Mental Health Sentiment Analysis Using Multimodal Deep Learning Models

Caleb Markus*, Etemi Joshua Garba, Yusuf Musa Malgwi

Department of Computer Science, Modibbo Adama University, Yola

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.101000269

Received: 05 October 2026; Accepted: 10 October 2026; Published: 01 November 2026

ABSTRACT

Mental health conditions affect hundreds of millions of people worldwide, and the way people express distress through text, speech and facial behaviour has become a growing source of signal for early detection. Deep learning models that combine more than one of these channels, commonly called multimodal models, are now one of the most active research directions in this space, yet the literature has grown too quickly and too unevenly for any one reader to take in by hand. This study is a systematic mapping review that combines a bibliometric analysis with a thematic synthesis. Six databases were searched, 9,044 records were identified, and 65 records were retained following the PRISMA 2020 guidelines. To keep the evidence base transparent, the corpus was separated into three strands, namely 42 primary empirical studies, which alone inform the performance synthesis, 18 secondary reviews and surveys, and 5 benchmark or methodological papers, the last two being used only for bibliometric mapping and technical context. Primary studies were appraised with a checklist extended to cover machine learning risks such as data leakage, class imbalance and validation design. Output rose sharply from 2022 onward, just over half of the corpus appeared in 2025 and 2026, and 66 percent of records are preprints. Multimodal fusion, mostly attention-based and transformer-based architectures evaluated on the DAIC-WOZ and E-DAIC corpora, is the largest theme. Where studies compared fused and single-modality models on the same data, fusion usually matched or exceeded the best single channel, but the gains were often small, rested on few peer-reviewed studies and small benchmark corpora, and cannot be generalised across tasks or populations. Explainability, external validation and cross-cultural testing remain uncommon. The review closes with a research agenda centred on explainable fusion, participant-level and cross-corpus validation, and clinically and ethically grounded evaluation.

Keywords: mental health, sentiment analysis, multimodal deep learning, depression detection, bibliometric analysis, thematic analysis, systematic literature review

INTRODUCTION

Mental health disorders are among the leading causes of disability worldwide, and the gap between the number of people who need care and the number who receive it remains wide in almost every country (World Health Organization, 2022). Recent reviews of digital and artificial intelligence-based detection open from the same premise: the burden is large, many cases go unrecognised for long periods, and scalable screening tools are needed (Cao et al., 2024; Guo et al., 2024; Khoo et al., 2024; Mocnik et al., 2025; Schlicher et al., 2025; VarastehNezhad et al., 2025). Conventional screening still depends on self-report questionnaires and clinical interviews, which are intermittent, subjective and costly in clinician time, and which can delay treatment (Chen et al., 2025; Khoo et al., 2024; Mandal et al., 2025; Mocnik et al., 2025; Sahu et al., 2025). For this reason, researchers have turned to detecting signs of depression, anxiety and related conditions from the way people write, speak and present themselves on camera, often without requiring a formal clinical instrument (Guo et al., 2024; Khan et al., 2025; Li, Yang, et al., 2025; Luo et al., 2026; Teng, Sun, et al., 2026).

A person's mental state rarely shows up in only one channel. Someone who is depressed may write in flatter, more self-focused language, speak with a slower and more monotonous voice, and show reduced facial expressiveness, often all at once, and models that rely on a single modality tend to miss whichever cue is

absent from that channel (Cenacchi et al., 2025; Khoo et al., 2024; Li, Yang, et al., 2025; Mocnik et al., 2025; Teng, Sun, et al., 2026). This observation has driven a shift toward multimodal deep learning, formally defined as learning from more than one input channel and combining them in a shared representation (Baltrusaitis et al., 2019), a direction that now dominates affective computing more broadly (Poria et al., 2017). Recent mental health studies combine text, speech and facial video in tri-modal systems, or add physiological and eye-tracking signals (Cenacchi et al., 2025; Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Li, Yang, et al., 2025; Rahman et al., 2025). Early work combined modalities through simple concatenation, whereas recent work relies on attention, transformer and large language model guided fusion (Chao et al., 2026; Chen et al., 2026; Haque et al., 2025; Mandal et al., 2025; Teng, Liu, et al., 2026; Teng, Sun, et al., 2026), following the wider shift toward attention mechanisms and deep architectures (Devlin et al., 2019; Hochreiter & Schmidhuber, 1997; LeCun et al., 2015; Vaswani et al., 2017).

Much of this architectural progress was first demonstrated on general-purpose affective computing benchmarks before being carried into mental health. The CMU-MOSEI corpus established one of the largest benchmarks for multimodal sentiment and emotion recognition from video, audio and text (Zadeh et al., 2018), and transformer-based models and surveys built on it (Gajjar & Ranaware, 2025; Lai et al., 2023; Shou et al., 2023; Sun et al., 2022; Xue & Marculescu, 2022) are the direct ancestors of the clinical fusion systems reviewed here. Those clinical systems are now evaluated on depression-specific corpora such as DAIC-WOZ, E-DAIC, EATD and D-Vlog, often several at once (Cenacchi et al., 2025; Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Haque et al., 2025; Mandal et al., 2025; Teng, Liu, et al., 2026), and the datasets themselves have become the subject of review (Al-Azani & El-Alfy, 2025).

The volume of published work has grown quickly enough that no single narrative review can keep pace with it, and a narrative review on its own cannot show a reader where the field's effort is concentrated. Recent reviews each cover one slice of the field: passive sensing (Khoo et al., 2024), large language models (Guo et al., 2024), bias in social media detection (Cao et al., 2024), affective computing for digital interventions (Schlicher et al., 2025), observable cues across disorders (Mocnik et al., 2025), and speech emotion recognition (Alhussein et al., 2025; Lombardo et al., 2025). A combined bibliometric and thematic review fills the gap between these slices. The bibliometric component maps the shape of the literature itself, including growth, venues, citation impact and keyword structure, while the thematic component goes inside that map and synthesises what the studies actually found (Mocnik et al., 2025; Schlicher et al., 2025; Tranfield et al., 2003).

Because several overlapping terms are used in this literature, the scope of the review needs stating at the outset. Sentiment analysis classifies the polarity or affective tone of an expression, emotion recognition assigns discrete or dimensional emotion labels, affective computing is the wider field concerned with sensing and modelling affect, and mental health condition detection infers the presence or severity of a clinically defined state such as depression, anxiety or post-traumatic stress disorder (Poria et al., 2017; Schlicher et al., 2025). These concepts overlap but are not interchangeable. The target construct of this review is the last of them, the detection or severity estimation of mental health conditions from affective signals in text, speech and facial or visual data. The phrase mental health sentiment analysis in the title is used in this sense, because affective signal is the shared mechanism through which all of the reviewed systems operate. Unimodal studies are retained because they are the reference point against which multimodal gains are judged, while general emotion recognition and sentiment benchmarks are retained only as contextual evidence and are never treated as evidence of clinical detection performance (Section 2.2).

This review addresses five research questions. RQ1 asks what the publication and citation trends are for mental health detection using multimodal deep learning. RQ2 asks which modality categories, text-based, speech-based, facial and visual, or fused multimodal, receive the most research attention. RQ3 asks what primary empirical studies in each category report about their own performance and limitations, and under what methodological conditions those results were obtained. RQ4 asks which datasets, architectures and fusion strategies recur most often. RQ5 asks what gaps remain, particularly in explainability, generalisation and clinical readiness, and what a defensible research agenda would look like. Section 2 sets out the methodology, Section 3 reports the bibliometric and thematic results, Section 4 discusses them against the research questions and states the limitations of the review, Section 5 proposes a research agenda, and Section 6 concludes.

METHODOLOGY

This review combines a quantitative bibliometric analysis with a qualitative thematic analysis and is reported in line with the PRISMA 2020 guidelines (Page et al., 2021). These were chosen over the earlier 2009 statement because they reflect more recent advances in how reviews identify, screen and report studies, and because they match the protocol-driven approach of recent reviews in this field (Alhussein et al., 2025; Cao et al., 2024; Guo et al., 2024; Khoo et al., 2024). The design is best described as a systematic mapping review with a focused empirical synthesis. The bibliometric map covers all 65 included records, whereas every statement about detection performance draws only on primary empirical studies. Separating these two purposes, rather than treating every record as equivalent evidence, is what allows the review to map a fast-moving field and still interpret its performance claims with appropriate caution (Kitchenham & Charters, 2007; Tranfield et al., 2003).

2.1 Search Strategy

Six databases were searched: Scopus, Web of Science, PubMed/MEDLINE, IEEE Xplore, the ACM Digital Library, and Google Scholar. No single source adequately covers both the clinical and the technical sides of this topic. Scopus and Web of Science were included for their structured multidisciplinary coverage and established use in systematic and bibliometric reviews (Tranfield et al., 2003); PubMed/MEDLINE for its depth in clinical and psychiatric literature; IEEE Xplore and the ACM Digital Library for deep learning and signal processing venues; and Google Scholar for its reach across preprint servers, where a large share of this literature first appears. Recent reviews in this area likewise combine biomedical and computing sources, including PubMed, Scopus, IEEE Xplore, the ACM Digital Library and Google Scholar (Alhussein et al., 2025; Cao et al., 2024; Guo et al., 2024; Lombardo et al., 2025; Schlicher et al., 2025).

Search terms combined three concept groups using Boolean operators, a mental health group (for example, depression, anxiety, mental health, sentiment), a modality and method group (for example, multimodal, deep learning, text, speech, audio, facial, video), and a detection group (for example, detection, classification, recognition, prediction). This three-group structure follows the recommendation for reviews in applied computing that search strings be built around the phenomenon, the method and the outcome so that no single broad term dominates the retrieved set (Kitchenham & Charters, 2007). Table 1 gives the string for each database in its native syntax, together with the fields searched and the restrictions applied. The search covered each database from its inception to the final search date, was limited to English-language records, and was not restricted by country, institution or publication type. All searches were last run on 28 September 2026. Google Scholar ranks results by relevance and does not support full Boolean export, so only the first 300 results were screened, in line with guidance on using Google Scholar in evidence reviews (Haddaway et al., 2015), and these were de-duplicated against the other five databases before screening.

Table 1. *Search string, fields and restrictions applied to each database*

Database	Fields searched	Search string and restrictions
Scopus	Title, abstract, keywords	TITLE-ABS-KEY (("mental health" OR depress *OR anxiety OR sentiment) AND (multimodal OR "deep learning" OR text OR speech OR audio OR facial OR video) AND (detect* OR classif *OR recogni* OR predict*)) AND LANGUAGE (english)
Web of Science	Topic (Core Collection)	TS=(("mental health" OR depress *OR anxiety OR sentiment) AND (multimodal OR "deep learning" OR text OR speech OR audio OR facial OR video) AND (detect* OR classif *OR recogni* OR predict*)), language filter English
PubMed/ MEDLINE	Title, abstract, MeSH terms	("mental health"[tiab] OR depress*[tiab] OR anxiety[tiab] OR sentiment[tiab] OR "Mental Disorders"[MeSH]) AND (multimodal[tiab] OR "deep learning"[tiab] OR "Deep Learning"[MeSH] OR text[tiab] OR speech[tiab] OR audio[tiab] OR facial[tiab] OR video[tiab]) AND (detect*[tiab] OR classif*[tiab] OR recogni*[tiab] OR predict*[tiab]), English filter

IEEE Xplore	Full text and metadata	Same Boolean string entered in the command search, restricted to Computer Science and Engineering subject areas
ACM Digital Library	Title, abstract, keywords	Same Boolean string applied to [Title], [Abstract] and [Keywords] fields
Google Scholar	All fields	Same Boolean string; first 300 results by relevance screened and de-duplicated against the other five databases

Note. All searches were limited to English-language records and run from database inception to the final search date, 28 September 2026.

2.2 Eligibility Criteria

Eligibility was defined at two levels, reflecting the two purposes of the review. For the empirical synthesis, studies were eligible if they were peer-reviewed journal articles, peer-reviewed conference papers or chapters, or openly available preprints reporting original empirical results; were written in English; applied at least one deep learning method, defined as any neural network-based architecture rather than a rule-based or classical statistical model (LeCun et al., 2015); and detected, classified or estimated the severity of a mental health condition, or an affective state explicitly linked to mental health, from text, speech, or facial and visual data. Preprints were included because the newest and most architecturally advanced work in this field frequently appears first on arXiv (Cenacchi et al., 2025; Chen et al., 2025; Haque et al., 2025; Mandal et al., 2025), but preprint status is flagged throughout and a peer-reviewed-only sensitivity check is reported in Section 3.3.

For the bibliometric map, two further types of record were retained when they met the same topical criterion, namely secondary reviews or surveys that synthesise deep learning-based detection of mental health or affective states from these modalities, and benchmark, dataset or methodological papers on which the empirical studies depend, such as challenge descriptions and general multimodal sentiment benchmarks. At title and abstract screening, reviews, editorials and abstracts were excluded unless they met this topical criterion, which is why 756 such records were excluded while a smaller set of directly relevant reviews was carried forward to full-text assessment. Records in these two contextual strands inform the publication, citation and keyword analyses and provide background to the synthesis, but they are never used as evidence of detection performance. Table 2 defines the three strands and their roles. Studies were excluded if they were outside this topical scope, used no deep learning or machine learning method, were editorials or abstracts

without original data, were not in English, could not be retrieved in full text, lacked enough methodological detail to be coded, or had been superseded by a later version of the same work. These exclusions are comparable to those applied in recent reviews of the field (Alhussein et al., 2025; Cao et al., 2024; Guo et al., 2024; Khoo et al., 2024; Lombardo et al., 2025).

Table 2. *Evidence strands in the included corpus and the role of each in the review*

Evidence strand	**Definition**	**Records, n**	**Role in this review**
Primary empirical	Original experiments that train or evaluate a deep learning model on text, speech, facial or visual, or multimodal data	42	Bibliometric mapping, thematic synthesis and all statements about detection performance
Secondary review or survey	Systematic, scoping, narrative or survey-type syntheses of related literature	18	Bibliometric mapping and research background only
Benchmark or methodological	Challenge and dataset descriptions, and general multimodal sentiment or fusion methods evaluated outside mental health	5	Bibliometric mapping and technical context only

Note. Two of the 42 primary studies examine general speech emotion recognition or the affective tone of language model responses rather than a mental health condition (Barhoumi & BenAyed, 2024;

VarastehNezhad et al., 2025) and are treated as contextual within the primary strand. Study-level assignments are listed in Appendix A.

2.3 Study Selection

Figure 1 shows the PRISMA 2020 flow of records (Page et al., 2021). The search returned 9,044 records. After 612 duplicates were removed, 8,432 records were screened on title and abstract and 8,268 were excluded, 3,450 as not relevant to the topic, 2,180 as outside the study criteria (for example, not concerned with mental health, not affective analysis, not multimodal or not deep learning), 1,620 as not using deep learning or machine learning methods, 756 as reviews, editorials or conference abstracts outside the topical criterion in Section 2.2, and 262 as not in English. Of the 164 full-text articles sought, 9 could not be obtained, 6 because access was restricted and 3 because no full text was available. The remaining 155 were assessed for eligibility and 90 were excluded, 48 because they did not address mental health detection from affective signals, 35 because they used neither deep learning nor a relevant multimodal method, 3 because they were not empirical, 2 for insufficient methodological detail and 2 as outdated or conceptually superseded. This left 65 records, comprising 42 primary empirical studies, 18 secondary reviews and 5 benchmark or methodological papers. Screening at both stages was carried out independently by two reviewers, with disagreements resolved through discussion with a third reviewer. The full list of included records, with the strand, theme and source type of each, is given in Appendix A.

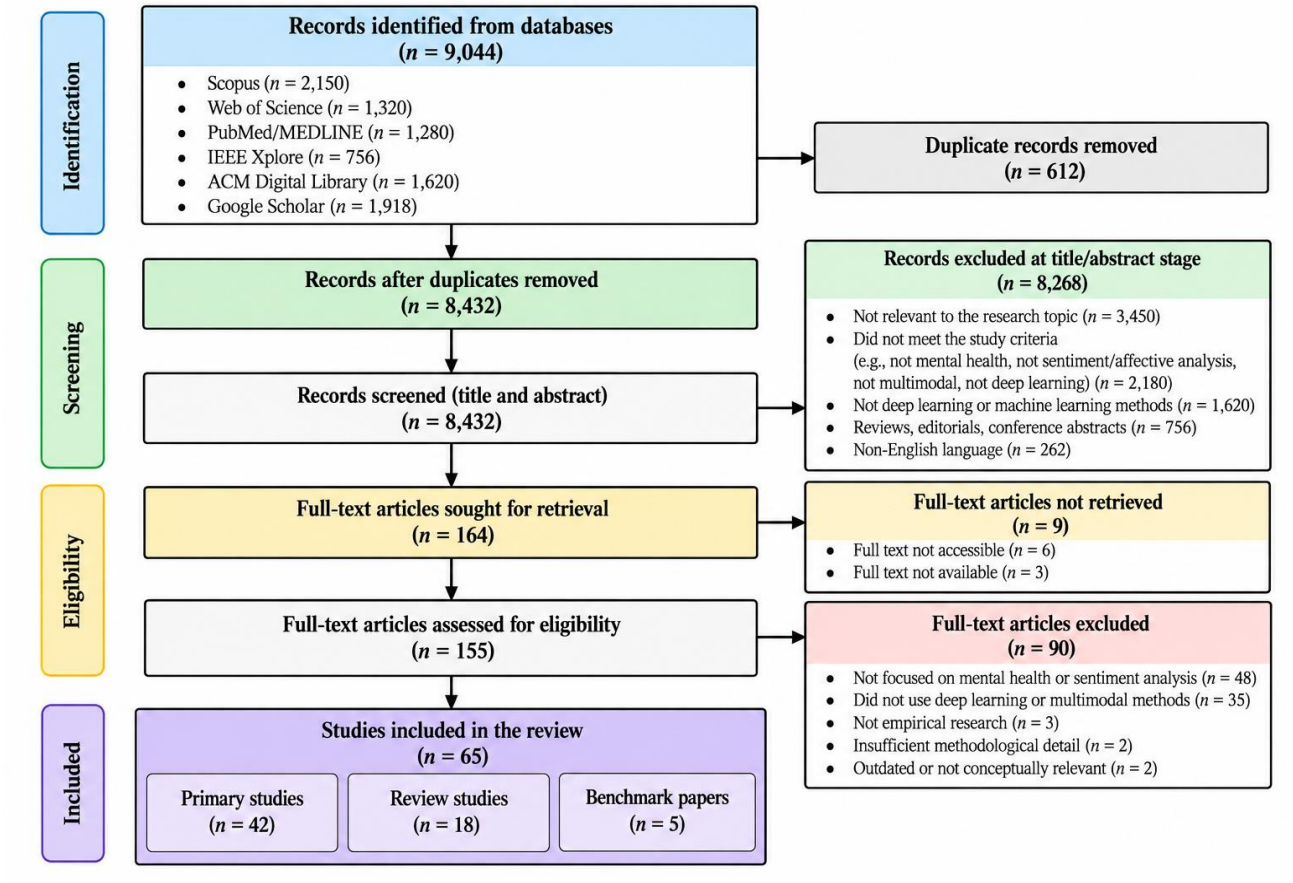

Figure 1. *PRISMA 2020 flow diagram of the study identification, screening, eligibility and inclusion process.*

2.4 Positioning Against Prior Reviews

Eighteen of the 65 records are reviews or surveys (Table 2). Table 3 compares this review with the six closest in scope, namely those addressing mental health detection across more than one modality or from a cross-cutting methodological angle (Cao et al., 2024; Guo et al., 2024; He et al., 2021; Khoo et al., 2024; Mocnik et al., 2025; Schlicher et al., 2025). Including such reviews in the bibliometric corpus, rather than treating them only as background, follows practice in bibliometric reviews, since how a field has already reviewed itself is part of its structure (Tranfield et al., 2003). They are, however, kept out of the performance synthesis.

Table 3. *Closest prior reviews within the included corpus and how this review differs from each*

Prior review	Scope	Method	How this review differs
Machine learning for multimodal mental health detection (Khoo et al., 2024)	Passive sensing approaches across modalities	Systematic review	Covers actively produced text, speech and video as well as passive data, and adds citation impact and keyword mapping
Large language model for mental health (Guo et al., 2024)	LLM-specific applications	Systematic review	Broader modality scope; not limited to LLM-based methods
Machine learning for mental illness detection on social media (Cao et al., 2024)	Bias and methodological challenges in text-only detection	Systematic review	Extends beyond text to speech, facial and fused modalities
Emotionally adaptive support (Schlicher et al., 2025)	Affective computing for mental health broadly	Narrative review	Restricts scope to detection and adds quantitative study counts and appraisal
Deep learning for depression recognition with audiovisual cues (He et al., 2021)	Audio and visual depression recognition	Review	Adds text and tri-modal fusion, covers work published after 2021, and adds bibliometric mapping
Multimodal observable cues in mood and anxiety disorders (Mocnik et al., 2025)	Review of reviews across observable cues	Review of reviews	Provides original thematic coding of primary studies rather than synthesising other reviews

2.5 Quality Appraisal

The 42 primary empirical studies were appraised using a checklist adapted from the Kitchenham and Charters guidelines (Kitchenham & Charters, 2007) and extended with items from the PROBAST risk-of-bias tool for prediction model studies (Wolff et al., 2019), shown in Table 4. Secondary reviews and benchmark papers were not scored, because the checklist is designed for primary studies. The Mixed Methods Appraisal Tool (Hong et al., 2018) was considered but not used, because the corpus consists almost entirely of quantitative model development and evaluation studies. Six of the eight criteria address risks specific to machine learning research that general checklists miss, namely how representative and well described the dataset is, whether data were split at participant level so that segments from one person cannot appear in both training and test sets, whether other sources of leakage were controlled, whether class imbalance was handled, whether the model was tested on an external corpus, and whether within-study baselines were reported (Kapoor & Narayanan, 2023; Wolff et al., 2019).

Table 4. *Quality appraisal criteria for primary empirical studies, adapted from Kitchenham and Charters (2007) and PROBAST (Wolff et al., 2019)*

Criterion	Scoring guide	Risk addressed
Clarity of aim and task definition	0 = not stated, 1 = implied, 2 = aim and task (classification, severity estimation or emotion recognition) explicitly stated	Reporting
Dataset description and representativeness	0 = absent, 1 = size only, 2 = size, population, language and recruitment described	Machine learning
Validation design	0 = single split with no detail, 1 = random or segment-level split, 2 = participant-level split, subject-wise cross-validation or official benchmark partition	Machine learning
Leakage control	0 = leakage likely or not addressed, 1 = partly addressed, 2 = preprocessing, feature selection and tuning confined to training data	Machine learning

Class imbalance	0 = ignored, 1 = reported only, 2 = reported and handled through resampling, weighting or class-sensitive metrics	Machine learning
External or cross-corpus testing	0 = none, 1 = held-out partition of the same corpus, 2 = independent corpus	Machine learning
Within-study baselines and ablation	0 = none, 1 = external baselines only, 2 = unimodal or ablation baselines on the same data	Machine learning
Appropriateness and completeness of metrics	0 = accuracy only on imbalanced data, 1 = partial, 2 = class-sensitive or error metrics suited to the task, with variance across runs or folds	Reporting

Note. Each criterion is scored 0 to 2, giving a total of 0 to 16. Totals of 12 to 16 are classed as higher quality, 8 to 11 as moderate and 0 to 7 as lower quality.

Each study was scored independently by two reviewers, and agreement before consensus was assessed with Cohen's kappa, shown in Equation (1), where p_o is the observed proportion of agreement and p_e is the proportion expected by chance. Values between 0.61 and 0.80 are conventionally read as substantial agreement (Landis & Koch, 1977). Disagreements were resolved by discussion, with a third reviewer consulted where consensus could not be reached.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

Quality informed the synthesis directly. Directional claims about modality or fusion performance in Sections 3 and 4 are based on higher and moderate quality studies, while results from lower quality studies are reported descriptively but are not used to support comparative conclusions.

2.6 Data Extraction

For each included record, the following were extracted into a structured spreadsheet: bibliographic details (authors, year, venue and peer-review status), evidence strand, modality or modalities used, the task addressed (binary classification, multiclass classification, severity estimation or emotion recognition), the condition studied, the architecture applied (for example, CNN, BiLSTM, transformer or graph neural network), the evaluation dataset with its size, language and population, the validation design and split level, class balance and how it was handled, whether an external test set was used, the fusion strategy where relevant, any explainability method, the reported metrics and results, and stated limitations. These fields extend those extracted by recent reviews, which record models, modalities, feature extraction, datasets and diagnostic performance (Alhussein et al., 2025; Cao et al., 2024; Guo et al., 2024; Khoo et al., 2024; Lombardo et al., 2025). The spreadsheet underlies the descriptive summary in Section 3, the quality appraisal and the coding summarised in Figure 2.

2.7 Bibliometric Analysis Approach

The bibliometric component summarises the included corpus descriptively and relationally, covering publication year, thematic category and source type, citation impact (citation counts and citations per year), journals represented, the co-occurrence of title terms, and the co-occurrence of keywords. Structured bibliometric analyses of this kind are well established for mapping a field (Aria & Cuccurullo, 2017; van Eck & Waltman, 2010) and have been applied to mental health and affective computing (Cao et al., 2024; Khoo et al., 2024; Mocnik et al., 2025; Schlicher et al., 2025). All statistics are computed from the 65 included records rather than the 9,044 retrieved records, so that trends reflect the literature that met the eligibility criteria and not the raw output of the search strings (Tranfield et al., 2003).

Citation counts, keywords and reference data were retrieved from OpenAlex (Priem et al., 2022) on 4 October 2026. OpenAlex was chosen because it is open and free and provides citation counts comparable to subscription indexes in many fields (Haunschild & Bornmann, 2024; Thelwall & Jiang, 2025), which suits a review that must be reproducible without a subscription. Records were matched by DOI and, where no DOI match existed, by exact title. Matches were found for 56 of the 65 records. The nine records without a match are excluded from citation figures rather than set to zero, so citation totals are lower bounds, and because

unmatched records tend to be recent, the citation share of the newest work is probably understated; all citation figures should be read with this in mind. Author names were standardised to surname and initials and checked manually for variant spellings, preprint and published versions of the same study were merged under the version of record, and countries were taken from the affiliations printed in each paper and standardised to current country names. Keywords are those assigned by OpenAlex rather than the authors' own, because author keywords were missing for many preprints, and they are labelled as such. Before counting, keyword synonyms and singular and plural forms were merged. For the title-term map, titles were lower-cased, stop words and generic words such as study, approach and using were removed, terms were lemmatised, and terms occurring in at least two titles were retained. Links were weighted by co-occurrence and normalised with the association strength measure used in VOSviewer (van Eck & Waltman, 2010). Keyword clusters were identified by modularity-based community detection on the normalised network, and all figures were produced in Python using the NetworkX and Matplotlib libraries. Country counts for Figure 4 come from the affiliations listed in each paper, and institution-level affiliation analysis and co-citation analysis were not carried out.

2.8 Thematic Analysis Approach

The thematic analysis followed the six-phase approach of Braun and Clarke (Braun & Clarke, 2006), chosen over alternatives such as grounded theory because it is designed for a defined, bounded set of texts that must be read closely and grouped into patterns, without the theoretical apparatus of methods built to generate new theory. The six phases were familiarisation with each full text, generating initial codes for modality, task, architecture, fusion strategy, validation and findings, searching for candidate themes, reviewing them against the coded extracts, defining and naming the final themes, and producing the synthesis reported in Section 3.

Four themes were defined a priori from the modality categories of the protocol and refined against the corpus, namely text-based sentiment analysis, speech and audio-based detection, facial and visual emotion-based detection, and multimodal fusion approaches. This split follows the standard taxonomy of input modalities in multimodal machine learning (Baltrusaitis et al., 2019) and recurs as the organising structure of several recent reviews (Khoo et al., 2024; Lombardo et al., 2025; Mocnik et al., 2025). These four themes organise the thematic synthesis. For the bibliometric figures only, two supplementary classes are added, general cross-modality records and foundational or benchmark papers, so that records not tied to a single modality can still be counted. This is why six categories appear in Figure 5 and Table 5. The supplementary classes are a counting device for the bibliometric map and are not themes of the synthesis. Within each theme, the synthesis also distinguishes the evidence strand of each record (Table 2) and the task it addresses, separating binary depression classification, severity estimation, detection of other conditions and general emotion recognition, because results from these tasks are not directly comparable. Figure 2 shows the coding schema.

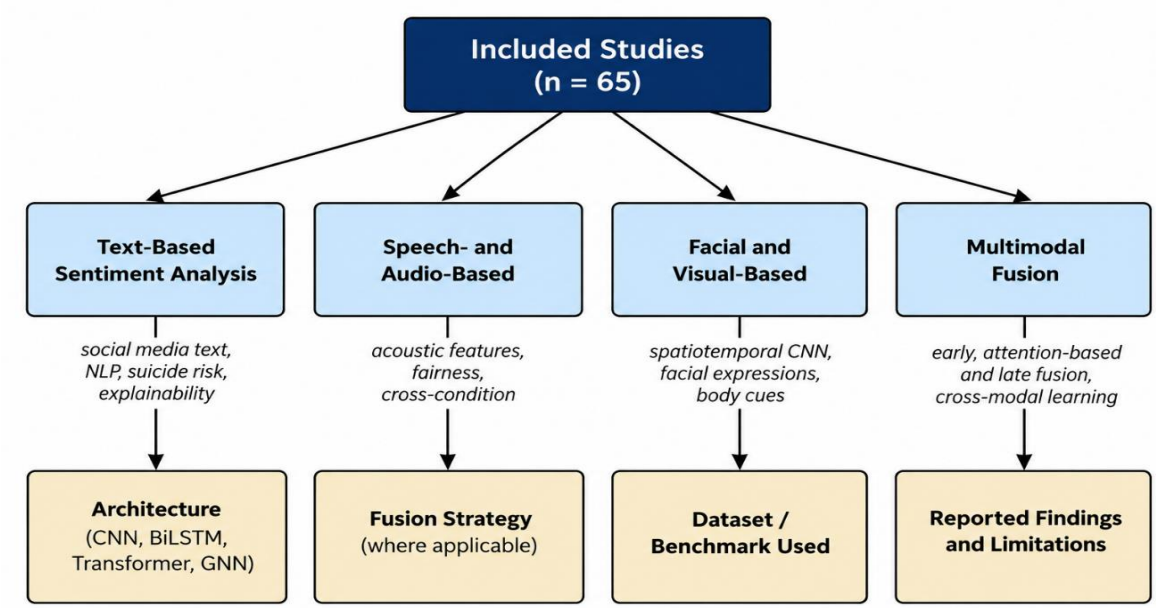

Figure 2. *Coding schema used for the thematic analysis, from the included studies to the four themes and their constituent codes.*

2.9 Evaluation Metrics Reported in the Literature

The included studies report results using a largely consistent set of metrics, so they are defined once here. For a binary detection task, precision, given in Equation (2), is the proportion of cases predicted positive that are truly positive, where *TP* is the number of true positives and *FP* the number of false positives. It is reported in fusion and text studies alike (Arji et al., 2023; Chao et al., 2026; Chen et al., 2026; Haque et al., 2025; Khan et al., 2025; Li, Yang, et al., 2025).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

Recall, given in Equation (3), is the proportion of truly positive cases that the model identifies, where *FN* is the number of false negatives, and it is reported alongside precision (Arji et al., 2023; Chao et al., 2026; Chen et al., 2026; Haque et al., 2025; Khan et al., 2025; Liu et al., 2022).

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

The F1-score, given in Equation (4), is the harmonic mean of precision and recall and is the most commonly reported summary metric across the included studies (Arji et al., 2023; Chao et al., 2026; Chen et al., 2026; Haque et al., 2025; Khan et al., 2025; Li, Yang, et al., 2025), since it balances the two in a way neither can alone. It is therefore used as the primary comparison metric for classification studies in Section 3.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Accuracy, given in Equation (5), is the proportion of all cases classified correctly, where *TN* is the number of true negatives (Arji et al., 2023; Chao et al., 2026; Chen et al., 2026; Haque et al., 2025; Khan et al., 2025; Liu et al., 2022). It is reported for completeness but not used for comparison, because on imbalanced corpora such as DAIC-WOZ, where non-depressed participants outnumber depressed participants (Gratch et al., 2014), a model can reach high accuracy while missing most positive cases.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

Studies that estimate depression severity on a continuous scale such as the PHQ-8 instead report the mean absolute error, shown in Equation (6), where $\hat{y}_i$ is the predicted score for participant *i*, y_i is the true score and *n* is the number of participants. Lower values indicate closer agreement with the clinical ground truth, in contrast to Equations (2) to (5), and the measure is used in severity estimation work (Cenacchi et al., 2025; Li, Yang, et al., 2025; Mandal et al., 2025). Because classification metrics and the mean absolute error answer different questions, results from classification and severity estimation studies are reported separately in Section 3 and are never compared directly.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{6}$$

2.10 Multimodal Fusion, as Modelled Across the Included Studies

Although architectures vary considerably, most fusion approaches can be placed on a continuum defined by where in the pipeline the modalities are combined (Baltrusaitis et al., 2019; Poria et al., 2017). In early or feature-level fusion, low-level features from each modality are concatenated before modelling. In intermediate or representation-level fusion, each modality is first encoded into its own learned representation, and these representations are combined before a shared classifier. In late or decision-level fusion, a separate model is trained for each modality and only their predictions are combined. The distinction matters because the three strategies differ in how they capture cross-modal interaction and how they cope with missing data. Equation (7) gives a generic weighted representation-level fusion, where *F* is the fused representation, F_t, F_s and F_v are the text, speech and visual representations, and α_t, α_s and α_v are learned or fixed weights that sum to one.

Because Equation (7) combines representations rather than predictions, it describes intermediate fusion and not late fusion.

$$F = \alpha_t F_t + \alpha_s F_s + \alpha_v F_v, \alpha_t + \alpha_s + \alpha_v = 1 \quad (7)$$

In attention-based fusion, the dominant approach among recent studies and one that builds on the attention mechanism for sequence modelling (Vaswani et al., 2017), the weights in Equation (7) are learned as a function of the input, typically through a softmax over a compatibility score for each modality, as shown in Equation (8), where e_k is the score for modality k and the denominator sums over all M modalities. Recent depression detection models use cross-modal and mutual attention in this way (Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Haque et al., 2025; Mandal et al., 2025; Teng, Sun, et al., 2026).

$$\alpha_k = \frac{\exp(e_k)}{\sum_{m=1}^{M} \exp(e_m)} \quad (8)$$

Decision-level late fusion, in contrast, combines the outputs of independently trained unimodal models, as in Equation (9), where $\hat{y}_k$ is the prediction, either a class probability or a severity score, produced by the model for modality k, w_k is its weight and $\hat{y}$ is the final prediction. Because each unimodal model can run on its own, late fusion degrades more gracefully when a modality is missing, but it cannot learn interactions that appear only when modalities are observed together.

$$\hat{y} = \sum_{k=1}^{M} w_k \hat{y}_k, \quad \sum_{k=1}^{M} w_k = 1 \quad (9)$$

The attention-weighted formulation lets a model down-weight a modality that is noisy, missing or less informative for a given individual, which several recent studies report as its main advantage over simple concatenation (Chao et al., 2026; Chen et al., 2025; Haque et al., 2025; Mandal et al., 2025; Teng, Sun, et al., 2026). Equations (7) to (9) are referred to in the synthesis of fusion studies in Section 3.9.

2.11 Citation Impact Measure

Raw citation counts favour older papers, so a normalised measure is also reported. The average citations per year of study i, given in Equation (10), divides its citation count C_i by the number of years since publication, counting the publication year as year one, where y_i is the publication year. Age-adjusted citation rates are a standard complement to raw counts in citation analysis (Aria & Cuccurullo, 2017; Haunschild & Bornmann, 2024; Thelwall & Jiang, 2025), and they matter here because 34 of the 65 records were published in 2025 or 2026 and have had little time to be cited.

$$\mathrm{CPY}_i = \frac{C_i}{2026 - y_i + 1} \quad (10)$$

RESULTS

3.1 Publication Trend and Geographical Distribution

Figure 3 shows the distribution of the 65 included studies by publication year. A single 2016 study anchors the corpus, the AVEC depression challenge description that established the DAIC-WOZ dataset many later fusion studies still build on (Valstar et al., 2016). Output stayed low and scattered through 2019 to 2021 (Bailey, 2020; Chlasta et al., 2019; Naderi et al., 2019; Woodward et al., 2019; Xie et al., 2021), then rose from 2022

onward. The final two years of the search window, 2025 and 2026, together account for 34 of the 65 studies, just over half the corpus, including the tri-modal, missing-modality, dynamic-fusion and adolescent screening work that defines the current frontier (Cenacchi et al., 2025; Chen et al., 2025; Haque et al., 2025; Luo et al., 2026; Mandal et al., 2025; Teng, Liu, et al., 2026; Teng, Sun, et al., 2026). A review completed even a year earlier would have missed roughly half the evidence base synthesised here.

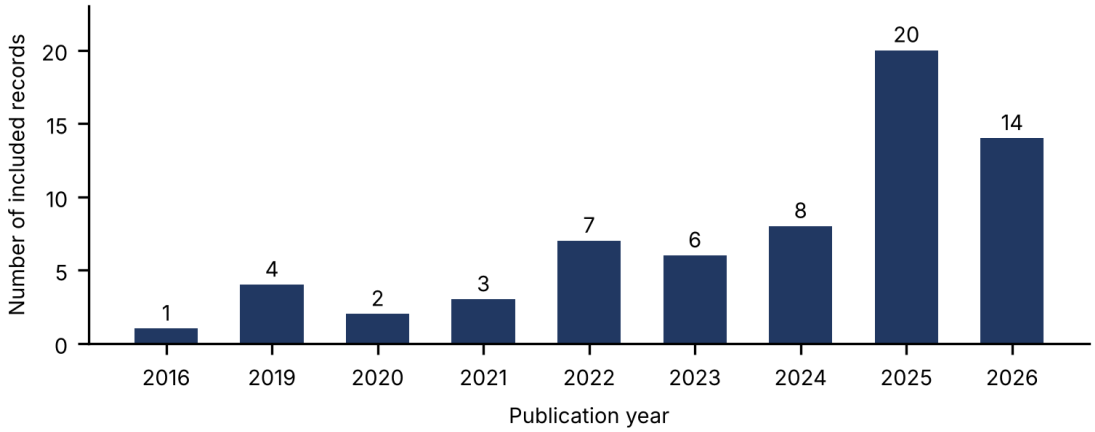

Figure 3. *Distribution of the 65 included studies by publication year.*

Figure 4 maps the countries represented in the corpus, using the affiliations listed in each paper. Contributions come from 24 countries on six continents. The United States contributes the most studies (14), followed by China (8), India and the United Kingdom (6 each), Canada and Germany (4 each), and Australia, Brazil, France, Japan and South Korea (3 each), while the remaining countries contribute one or two studies each. Research on multimodal mental health sentiment analysis is therefore produced across six continents, with the United States and China the largest single contributors. Country counts describe where the research was produced and not where the underlying data were collected, so they say little about the populations on which the models were trained. This matters because cross-cultural generalisation is one of the open problems identified in this review.

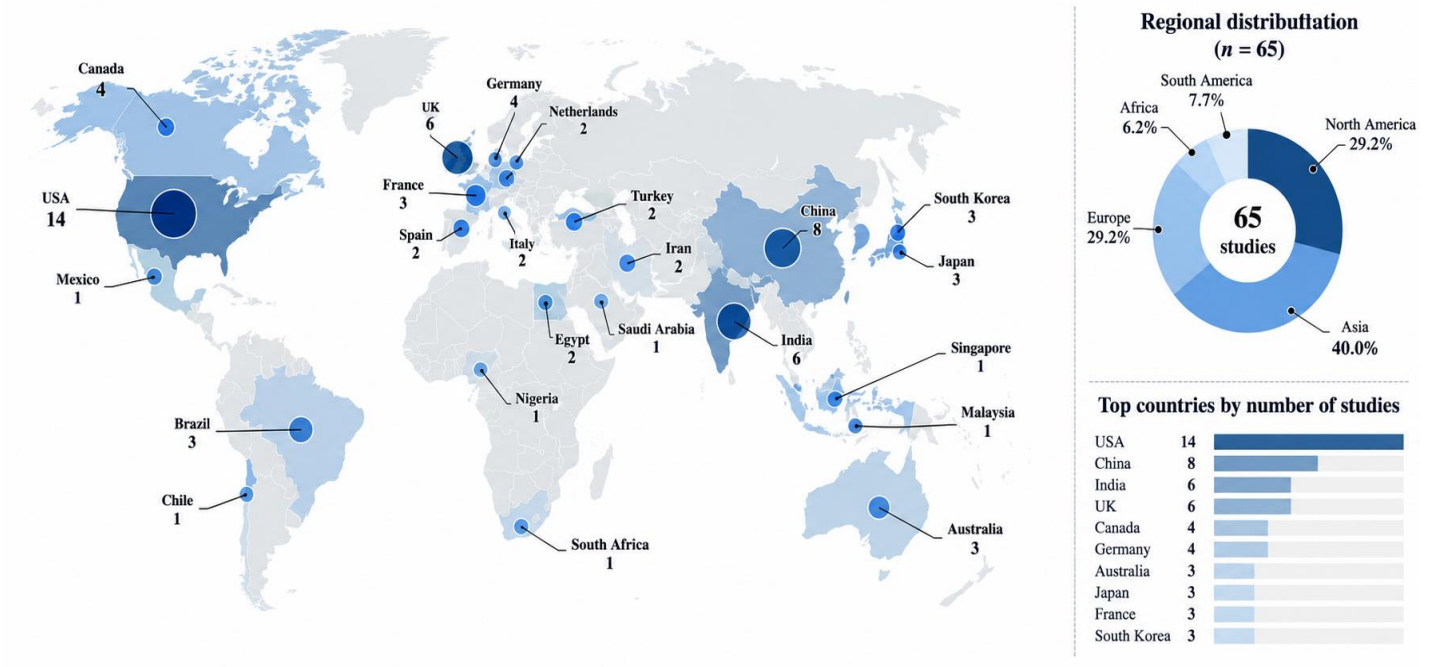

Figure 4. *Geographical distribution of the 65 included studies by country and region.*

3.2 Thematic Distribution and Citation Share

Figure 5 shows how the 65 records distribute across the four themes and the two supplementary bibliometric classes defined in Section 2.8 (panel a), and how many citations each has received (panel b, OpenAlex, 4 October 2026, 56 of 65 records with data). Multimodal fusion is the largest theme with 24 records, 37 percent of the corpus (Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Haque et al., 2025; Li, Yang, et al., 2025; Rahman et al., 2025), followed by text-based sentiment analysis with 15 (Belcastro et al., 2026; Ibrahimov et al., 2024; Kaur et al., 2026; Khan et al., 2025; Kim et al., 2025), speech and audio-based detection with 10 (Alhussein et al., 2025; Hassan et al., 2025; Lanzlinger et al., 2026; Li, Chng, & Guan, 2025; Lombardo et al.,

2025), and facial and visual detection with 6 (Alvarez Casado et al., 2022; Liu et al., 2022; Sahu et al., 2025; Shangguan et al., 2024; Zhang & Zhang, 2025). The supplementary classes hold 6 general cross-modality records, mostly reviews (Cao et al., 2024; Guo et al., 2024; Mocnik et al., 2025; Schlicher et al., 2025), and 4 foundational or benchmark papers (Al-Azani & El-Alfy, 2025; Chlasta et al., 2019; Gajjar & Ranaware, 2025; Valstar et al., 2016). These theme counts include the reviews and benchmark papers that sit within each theme. Counting primary empirical studies alone, the themes hold 17 fusion, 11 text-based, 6 speech-based and 5 facial and visual studies, with the remaining 3 primary studies in the supplementary classes (Appendix A). Citations are spread more evenly than record counts, with text-based and multimodal fusion records having each received 166 citations, general reviews 142, foundational papers 135, speech-based records 117, and facial and visual records 41. Per record, reviews and benchmark papers are therefore far more cited than the newest fusion architectures, which have had little time to accumulate citations.

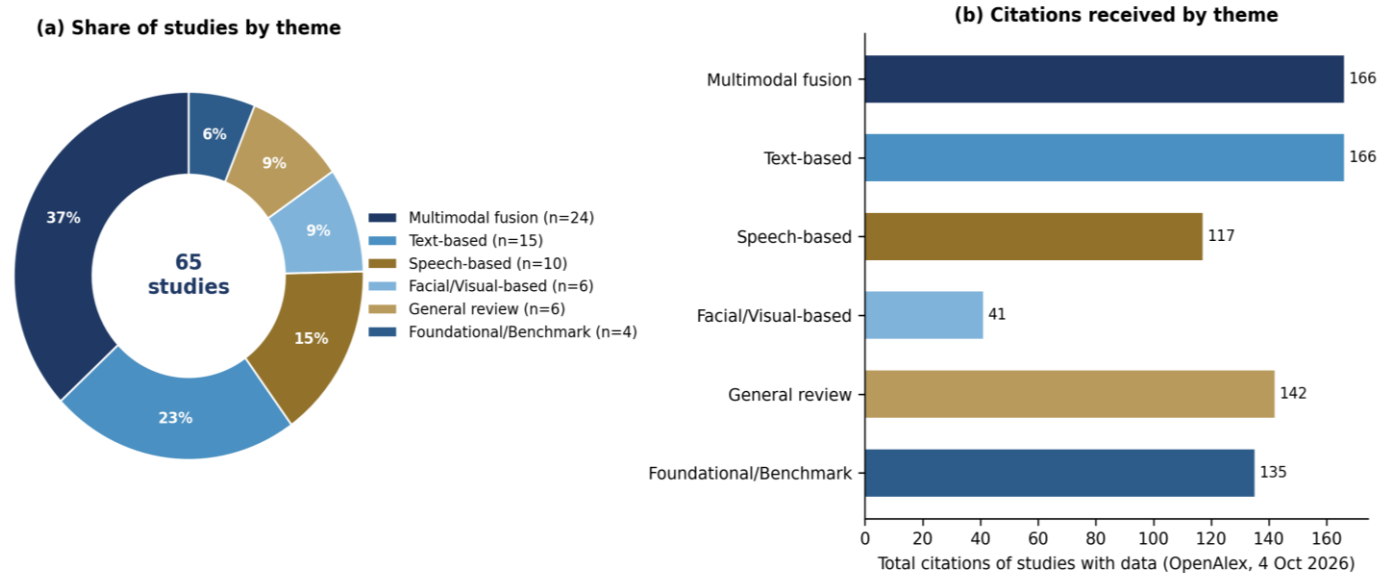

Figure 5. *Share of the 65 included records by theme (a) and total citations received by theme (b)*

3.3 Source Type

Table 5 summarises the corpus. Forty-three records, 66 percent, are arXiv preprints, 18 are peer-reviewed journal articles and 4 are peer-reviewed book or conference chapters. Among the 42 primary empirical studies, 28 are preprints and 14 are peer-reviewed. As Section 2.2 explains, preprints were included deliberately, and excluding them would have removed two-thirds of the evidence base, including most recent fusion, missing-modality and adolescent screening work (Cenacchi et al., 2025; Chen et al., 2025; Haque et al., 2025; Luo et al., 2026; Mandal et al., 2025; Teng, Liu, et al., 2026; Teng, Sun, et al., 2026). Peer-reviewed journal articles concentrate in the review and text-based themes (Arji et al., 2023; Kerz et al., 2023; Khan et al., 2025; Khoo et al., 2024; Lombardo et al., 2025; Schlicher et al., 2025).

Table 5. *Descriptive summary of the included corpus (N = 65)*

Characteristic	Category	Studies, n (%)
Publication period	2016 to 2021	10 (15%)
	2022 to 2023	13 (20%)
	2024 to 2026	42 (65%)
Thematic category	Multimodal fusion	24 (37%)
	Text-based	15 (23%)
	Speech-based	10 (15%)
	Facial/visual-based	6 (9%)
	General review	6 (9%)
	Foundational/benchmark	4 (6%)

Source type	arXiv preprint	43 (66%)
	Peer-reviewed journal article	18 (28%)
	Peer-reviewed book chapter	4 (6%)
Evidence strand	Primary empirical	42 (65%)
	Secondary review or survey	18 (28%)
	Benchmark or methodological	5 (8%)
Source type, primary studies only (n = 42)	arXiv preprint	28 (67%)
	Peer-reviewed	14 (33%)

Note. Percentages are rounded to the nearest whole number and may not sum exactly to 100 within each characteristic.

Table 5 makes one pattern explicit, namely that just under two-thirds of the records, and two-thirds of the primary empirical studies, are preprints. This reflects how quickly the subfield moves, but it also means that most reported results have not been through peer review and should be read as preliminary. To test whether the central conclusions depend on preprints, the synthesis was repeated using only the 14 peer-reviewed primary studies. The descriptive picture holds, since fusion accounts for 7 of the 14 and remains the most common design (Aishwarya et al., 2026; Chao et al., 2026; Chen et al., 2026; Li, Yang, et al., 2025; Pavithra & Trupthi, 2026; Rajkumar et al., 2026; Ramadevi et al., 2026). The comparative evidence, however, becomes thin. Among peer-reviewed studies, the clearest within-study comparison is the trimodal model that improved on the best bimodal combination by a small margin (Li, Yang, et al., 2025), alongside an interpretable ensemble that reported text as its strongest single channel (Rajkumar et al., 2026). The claim that fusion improves on single modalities therefore rests mainly on preprints and is treated as provisional (Section 3.9 and Table 6).

3.4 Term and Keyword Structure

To complement the thematic coding with an independent, quantitative view of the vocabulary, Figure 6 maps the co-occurrence of terms in the titles of the 65 studies, following the logic of a VOSviewer term map: node size reflects how often a term appears, link strength how often two terms appear in the same title, with a minimum occurrence of two. Depression, multimodal and detection are the three most frequent terms, appearing in 27, 24 and 24 titles respectively (preprocessing as described in Section 2.7), which mirrors the thematic distribution in Figure 5 (Cenacchi et al., 2025; Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Li, Yang, et al., 2025; Rahman et al., 2025).

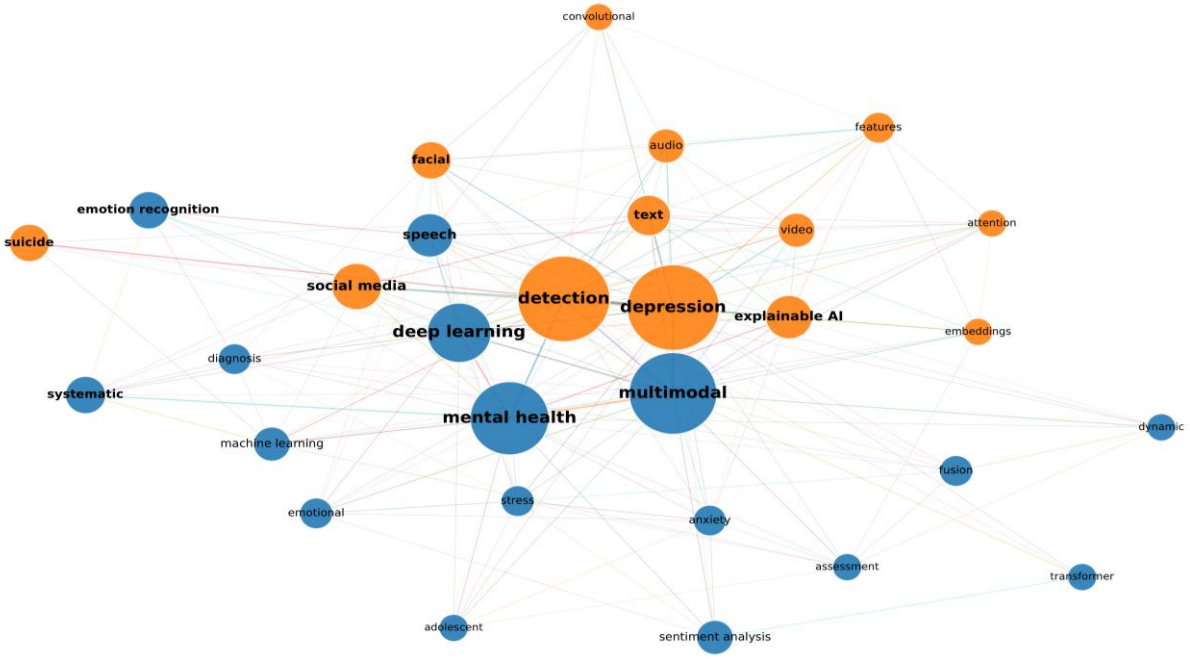

Figure 6. *Term co-occurrence network of the 65 included studies, based on title terms (minimum occurrence = 2).*

Figure 7 repeats the exercise with keywords rather than title terms. Using OpenAlex-assigned keywords for 56 studies and a minimum occurrence of three, it yields 31 keywords in three clusters. The largest cluster joins multimodal fusion, depression detection, social media, sentiment analysis, large language models and mental health, and corresponds to the text-based and fusion work (Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Guo et al., 2024; Kim et al., 2025). A second cluster groups explainable AI, depression, anxiety, depression severity, post-traumatic stress disorder and bipolar disorder, the clinical conditions studied (Belcastro et al., 2026; Cenacchi et al., 2025; Ibrahimov et al., 2024; Mandal et al., 2025; Mocnik et al., 2025). A third groups emotion recognition, speech analysis, facial expression analysis, physiological signals and affective computing, the sensing-oriented and affective computing work (Khoo et al., 2024; Li, Chng, & Guan, 2025; Lombardo et al., 2025; Sahu et al., 2025; Schlicher et al., 2025). Because these keywords are algorithm-assigned, they indicate topic structure rather than authors' own framing.

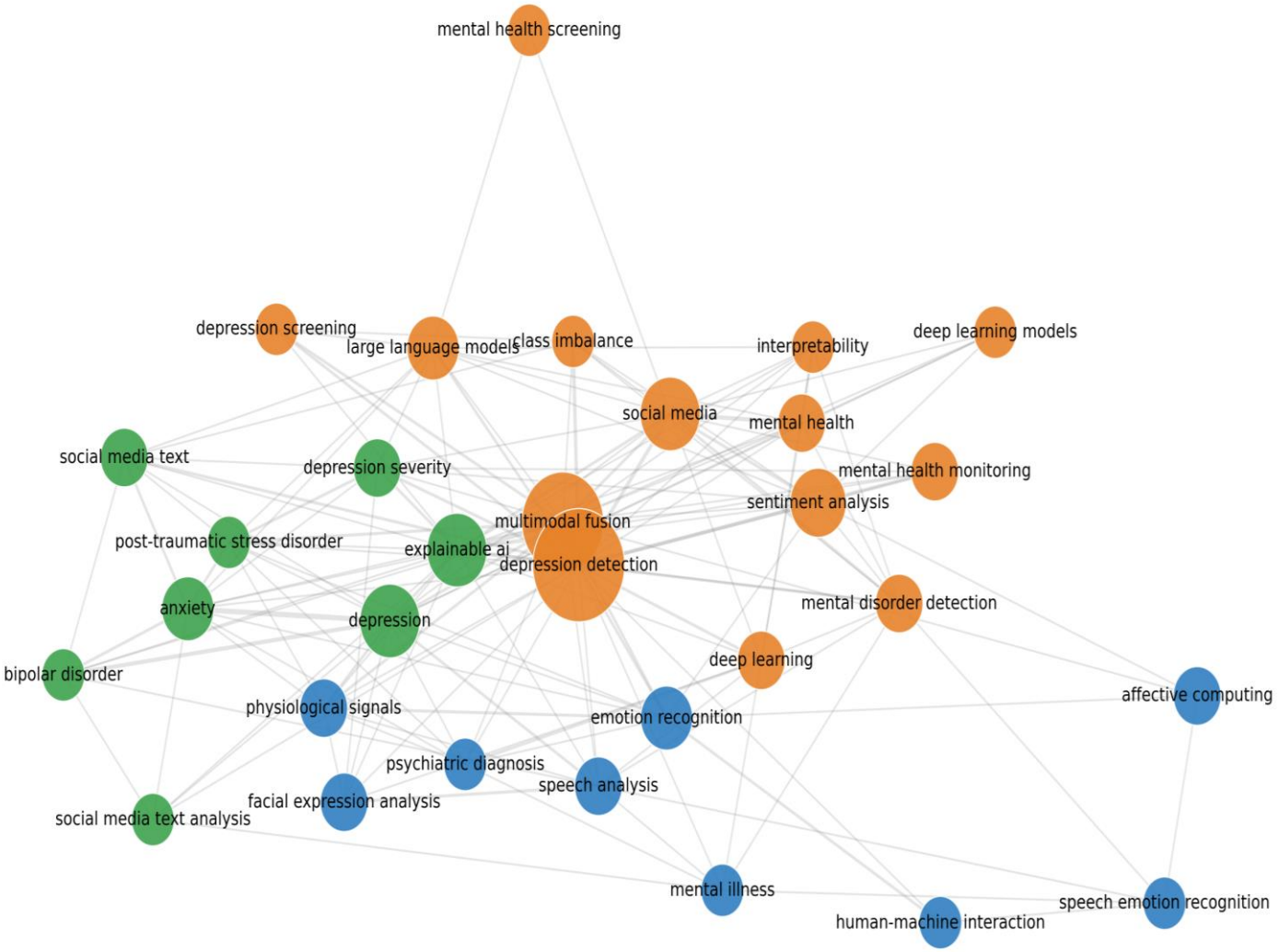

Figure 7. *Keyword co-occurrence network of the included studies*

Figure 8 maps the terms in Figure 6 onto the four themes used for the thematic synthesis, separating terms specific to one theme from cross-cutting terms such as depression and detection that recur across all four (Chao et al., 2026; Chen et al., 2026; Kim et al., 2025; Li, Yang, et al., 2025; Lombardo et al., 2025; Sahu et al., 2025).

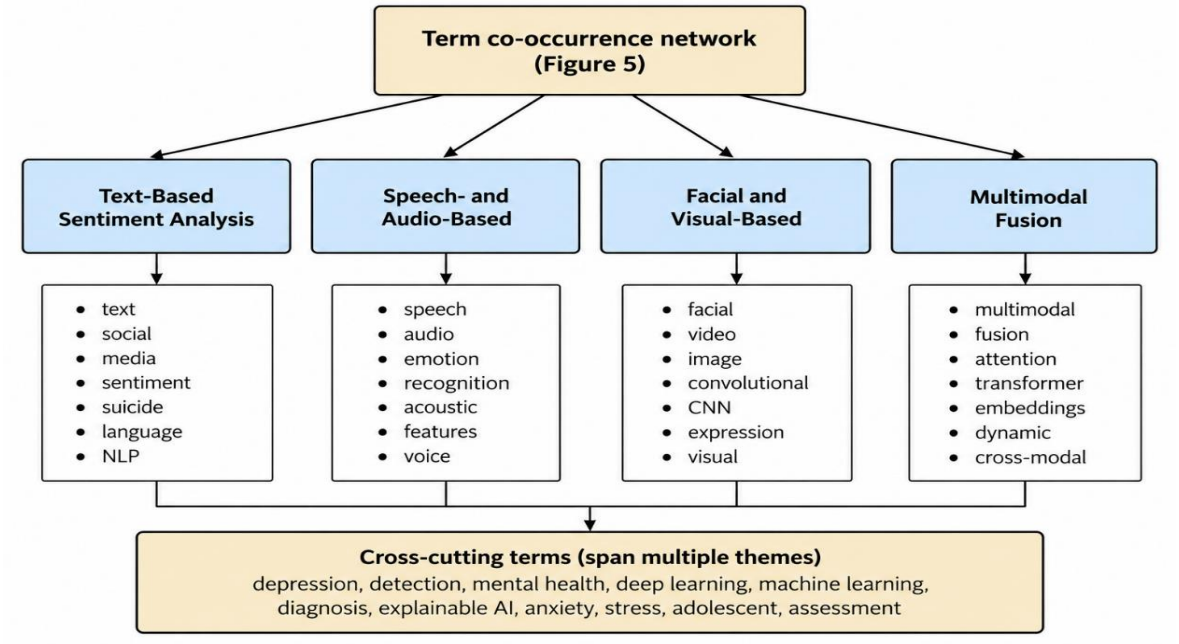

Figure 8. *Conceptual mapping between the title-term co-occurrence network (Figure 6) and the four thematic categories used in this review.*

Figure 9 links themes, keywords and source types in a three-field plot. The ten most frequent keywords are depression detection (18 studies), multimodal fusion (14), social media, depression and explainable AI (7 each), sentiment analysis (6), and emotion recognition, large language models and anxiety (5 each), plus physiological signals (4). Multimodal fusion studies supply 12 of the 14 papers keyed to multimodal fusion and 8 of the 18 keyed to depression detection (Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Haque et al., 2025; Mandal et al., 2025), while social media and large language models draw mainly on text-based studies (Belcastro et al., 2026; Cao et al., 2024; Guo et al., 2024; Kim et al., 2025; Nguyen, 2024). In nine of the ten keyword groups preprints outnumber journal articles, the exception being emotion recognition, which is consistent with the source profile in Table 5 (Cenacchi et al., 2025; Ibrahimov et al., 2024; Teng, Liu, et al., 2026; Teng, Sun, et al., 2026; VarastehNezhad et al., 2025).

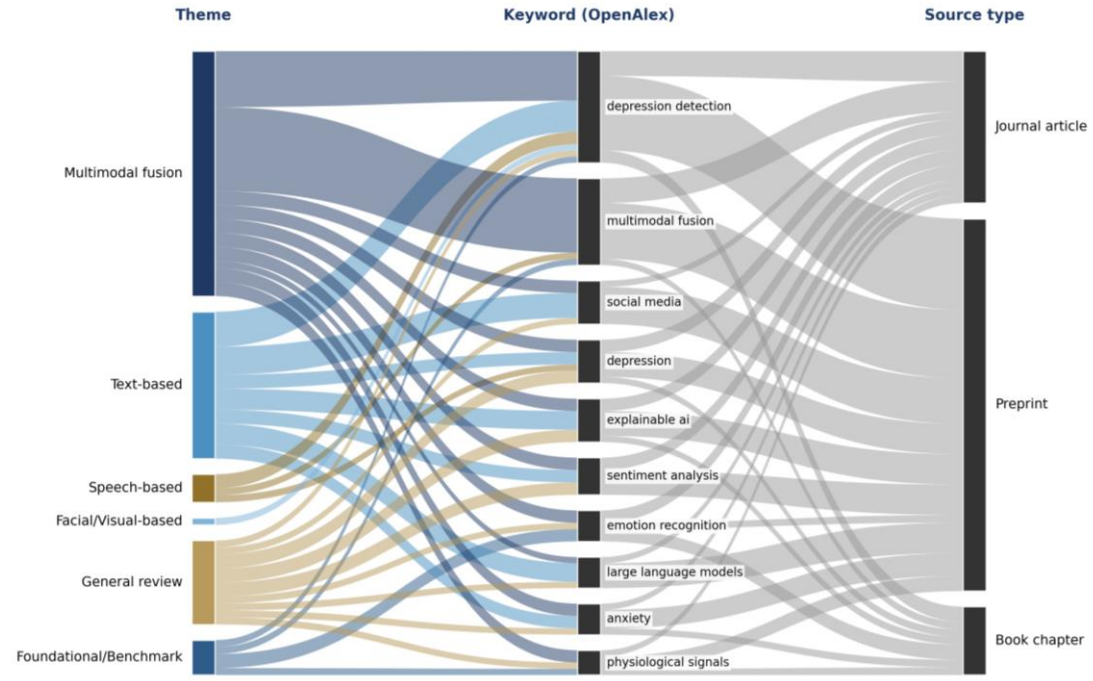

Figure 9. *Three-field plot linking theme, keyword (OpenAlex-assigned) and source type for the included studies.*

3.5 Journals and Citation Impact

Figure 10 shows the 18 peer-reviewed journal articles by journal. No journal accounts for more than 3 of the 65 studies, with Artificial Intelligence Review and Frontiers in Psychiatry each represented three times, which is consistent with a field still consolidating around its venues (Al-Azani & El-Alfy, 2025; Alhussein et al., 2025; Barhoumi & BenAyed, 2024; Kerz et al., 2023; Li, Yang, et al., 2025; Sahu et al., 2025).

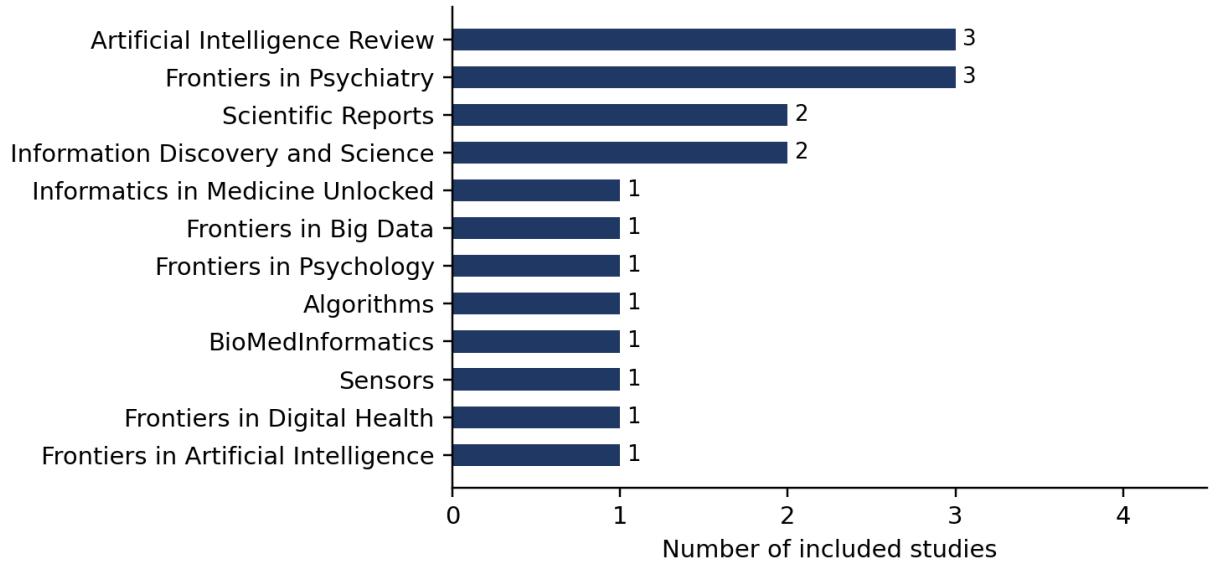

Figure 10. *Journals represented among the 18 peer-reviewed journal articles in the included corpus.*

Citation impact is concentrated. Of the 56 studies with data, the total is 767 citations, the median is 1, and 23 studies have no citations yet. Figure 11 ranks the ten most cited studies: the passive sensing review leads with 148 citations (Khoo et al., 2024), followed by the Facebook suicide risk study with 116 (Ophir et al., 2020), the AVEC 2016 challenge with 109 (Valstar et al., 2016), the explainable AI study with 102 (Kerz et al., 2023), and the real-time speech emotion recognition study with 63 (Barhoumi & BenAyed, 2024). Most of the top ten are reviews, benchmark papers, or speech and text studies, and the only fusion-theme paper among them is itself a review (Al-Azani & El-Alfy, 2025; Alhussein et al., 2025; Arji et al., 2023; Liu et al., 2022; Schlicher et al., 2025).

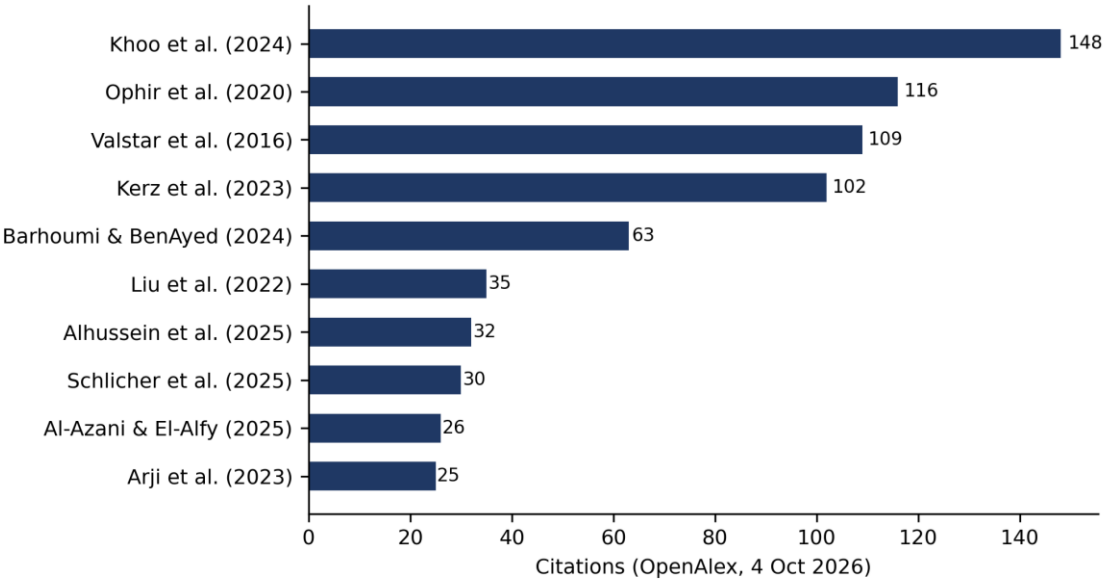

Figure 11. *Ten most cited included studies (OpenAlex, 4 October 2026; 56 of 65 studies with data).*

Figure 12 adjusts for age using Equation (10). The ranking changes at the margin but not at the top: the passive sensing review is again first at 49.3 citations per year, followed by the explainable AI study (25.5), the speech

emotion recognition study (21.0), the suicide risk study (16.6) and the conversational speech emotion review (16.0) (Alhussein et al., 2025; Barhoumi & BenAyed, 2024; Kerz et al., 2023; Khoo et al., 2024; Ophir et al., 2020). The newest architectures are largely absent from both figures, not because they matter less but because 2025 and 2026 papers have had little time to be cited (Cenacchi et al., 2025; Chao et al., 2026; Chen et al., 2026; Haque et al., 2025; Mandal et al., 2025; Teng, Sun, et al., 2026), which is a caution against reading citation rank as a measure of methodological importance.

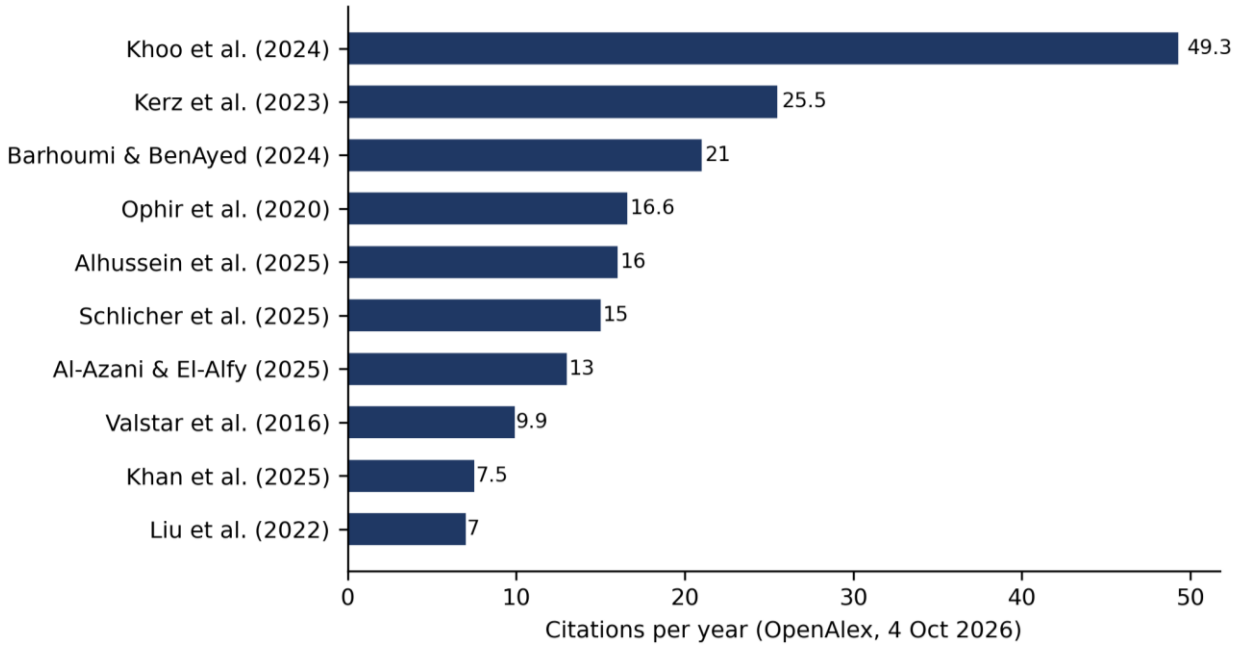

Figure 12. *Average citations per year of the top ten included studies (OpenAlex, 4 October 2026; 56 of 65 studies with data).*

3.6 Text-Based Sentiment Analysis for Mental Health

Text-based detection is the most mature of the four themes, both in the sense that it has the longest history in this corpus and in the sense that its methods have converged most clearly on a common approach. Of its 15 records, 11 are primary empirical studies (Belcastro et al., 2026; Farruque et al., 2022; Garcia et al., 2023; Kaur et al., 2026; Khan et al., 2025; Kim et al., 2025; Long et al., 2022; Nguyen, 2024; Ophir et al., 2020; Saeed & Ahmed, 2025; Zevallos et al., 2024) and 4 are reviews or surveys used for context (Arji et al., 2023; Garg, 2021; Ibrahimov et al., 2024; Ji et al., 2019). Across the primary studies, the dominant pattern is a transformer-based language model, usually BERT, RoBERTa or a domain-adapted variant, fine-tuned on social media text and frequently paired with a BiLSTM layer to capture sequential structure, reflecting the broader shift toward transformer-based architectures (Devlin et al., 2019). A study summarised in one included systematic review reports accuracies of 94, 96 and 98 percent for LSTM, BiLSTM and a BERT plus knowledge-distillation model on the same Reddit and Twitter depression and anxiety task (Arji et al., 2023), a pattern that should be read cautiously, because accuracy on such social media corpora is sensitive to class balance and to how labels were assigned. One primary study applies a BiLSTM-based framework that combines linguistic and temporal posting-behaviour features for early detection of mental health crises (Saeed & Ahmed, 2025), another applies comparable architectures in a shared-task setting focused on signs of depression in social media text (Garcia et al., 2023), and a third applies temporal modelling to clinical depression trajectories over time rather than a single snapshot (Farruque et al., 2022). More recent work has pushed toward explainability rather than raw accuracy alone, tracking shifts in a user's depression status over time through explainable digital trace analysis (Belcastro et al., 2026) and showing that large language model-derived embeddings can outperform raw text embeddings on depression severity classification while remaining more interpretable to a clinician reading the output (Kim et al., 2025).

Suicide risk detection forms a distinct sub-thread within the text-based theme, spanning six of the fifteen records, four primary studies (Long et al., 2022; Nguyen, 2024; Ophir et al., 2020; Zevallos et al., 2024) and

two reviews (Garg, 2021; Ji et al., 2019). These records converge on a similar conclusion from different angles, that deep learning models can detect quantifiable linguistic signals of suicide risk in ordinary social media posts but that reported performance depends heavily on dataset quality rather than model choice alone. The earliest deep neural network study in this sub-thread reported only moderate discrimination, with an area under the curve between 0.697 and 0.746 across model variants (Ophir et al., 2020). One study extends the sub-thread to a multilingual setting, addressing the largely English-language focus of earlier work (Zevallos et al., 2024), and a large language model-based approach targets settings with limited labelled training data (Nguyen, 2024).

Two further studies sit at the boundary between the text-based theme and general explainability research. A systematic review specifically of explainable AI approaches to text-based mental disorder detection finds that attention-based interpretability methods are now the norm, but that formal validation of whether those attention weights actually correspond to clinically meaningful signal remains rare (Ibrahimov et al., 2024), and a related study proposes a clinician-in-the-loop hybrid framework that combines transformer-based detection with human validation of each prediction rather than relying on the model's output alone (Kaur et al., 2026). A broader systematic review of deep learning algorithms across the full range of mental disorders, not limited to text, situates this text-based cluster within the wider field and reports that sentiment analysis specifically is among the most frequently applied techniques across the 85 studies it synthesises (Arji et al., 2023). Recent work extends explainability beyond text to dynamic summaries, acoustic indicators and clinician-readable reports (Belcastro et al., 2026; Kim et al., 2025; Lanzlinger et al., 2026; Mocnik et al., 2025; Teng, Liu, et al., 2026).

3.7 Speech- and Audio-Based Mental Health Detection

The ten records in the speech and audio theme comprise six primary studies (Bailey, 2020; Barhoumi & BenAyed, 2024; Hassan et al., 2025; Lanzlinger et al., 2026; Li, Chng, & Guan, 2025; Naderi et al., 2019) and four reviews or surveys (Alhussein et al., 2025; He et al., 2021; Lombardo et al., 2025; Woodward et al., 2019). Together with an early DAIC-based screening study classed with the foundational papers (Chlasta et al., 2019), the primary studies share a consistent methodological backbone of acoustic feature extraction, typically Mel-frequency cepstral coefficients, chroma features or raw spectrograms, followed by a convolutional or recurrent encoder, consistent with the long short-term memory and convolutional architectures formalised for sequence and signal processing (Hochreiter & Schmidhuber, 1997). The early study applies deep convolutional neural networks directly to spectrograms of participant speech from the DAIC corpus for automated depression screening (Chlasta et al., 2019), establishing a pattern that later speech-based studies still follow in outline, even as the architecture has shifted toward self-supervised and transformer-based audio encoders (Li, Chng, & Guan, 2025). An earlier multimodal study models audio speech samples for mental disorder prediction, combining acoustic features with basic text cues before the more elaborate fusion architectures discussed in Section 3.9 became common (Naderi et al., 2019). One primary study addresses general speech emotion recognition rather than a mental health condition (Barhoumi & BenAyed, 2024) and is read here as evidence about transferable methods, not about clinical detection. A recurring finding across this theme is that speech carries detectable signal even when recording conditions are far from clinical, and that this signal links to specific acoustic correlates, such as reduced pitch variation and slower articulation rate, rather than to opaque statistical regularities alone (Lanzlinger et al., 2026).

Two methodological concerns recur across this theme more than any other. First, fairness: one included study reports that depression detection accuracy from audio features differs measurably between male and female speakers trained on the same DAIC-WOZ corpus, a finding with direct implications for any clinical deployment of these systems (Bailey, 2020). Second, scope beyond depression: speech emotion recognition methods originally built for general affective computing are increasingly being repurposed for mental health screening specifically (Alhussein et al., 2025; Barhoumi & BenAyed, 2024), and a recent scoping review of this repurposing finds that speech-based screening has now been applied, with varying success, to post-traumatic stress disorder and bipolar disorder in addition to depression, though with substantially smaller evidence bases for each (Lombardo et al., 2025). A broader survey of technologies for mental well-being situates speech-based screening alongside other passive-sensing approaches as part of a wider move toward

continuous, low-burden monitoring (Woodward et al., 2019), a review focused specifically on audiovisual cues for depression recognition reaches a similar conclusion while noting that audio alone typically underperforms audio combined with visual cues (He et al., 2021), and a study specifically addressing embedding techniques for multimodal mental illness assessment extends this acoustic feature extraction approach into the fusion architectures discussed next (Hassan et al., 2025).

3.8 Facial and Visual Emotion-Based Mental Health Detection

The six included facial and visual studies are the smallest of the four themes, but they are methodologically distinct enough to warrant separate treatment (Alvarez Casado et al., 2022; Liu et al., 2022; Sahu et al., 2025; Shangguan et al., 2024; Xie et al., 2021; Zhang & Zhang, 2025). Of the six records, five are primary studies and one is a survey used for context (Shangguan et al., 2024). One of the primary studies uses EEG rather than visual input (Zhang & Zhang, 2025); it was coded with the non-verbal sensing studies and is interpreted as physiological rather than facial evidence. Several apply three-dimensional or spatiotemporal convolutional networks directly to facial video recorded during a structured task, such as a self-rating depression scale questionnaire, rather than to still images (Xie et al., 2021), on the reasoning that depression shows up in how facial expression changes over the course of an interaction, not only in any single frame. A related study applies remote photoplethysmography, extracting physiological heart-rate signal directly from facial video rather than from expression alone, as a complementary visual-channel biomarker for depression (Alvarez Casado et al., 2022). One study trains a convolutional network on the FER2013 facial expression dataset and reports 70.71 percent accuracy, which the authors note already exceeds the roughly 65 percent human accuracy typically reported for the same task, before combining this facial channel with body movement features to estimate depression severity specifically among patients with major depressive disorder (Liu et al., 2022). A further study extends facial video analysis specifically to social anxiety detection, moving the theme's focus beyond depression alone (Sahu et al., 2025), and a contemporary survey of facial expression analysis situates these clinical applications within the wider landscape of facial expression recognition technology, including its use in Internet of Things and ambient sensing contexts (Shangguan et al., 2024).

A pattern worth flagging on its own is how often this theme's studies extend beyond adults. One included study applies an EEG-based analysis to identify emotion dysregulation as an early marker of adolescent mental health difficulty, using multimodal physiological and behavioural signals rather than facial video alone (Zhang & Zhang, 2025), a direction echoed more directly within the multimodal fusion theme discussed next, where two further included studies address adolescent-specific screening in more detail (Du et al., 2023; Luo et al., 2026). This theme also shows a clear sign of methodological self-awareness: several of its studies note that most facial-based work analyses spatial and temporal information separately rather than jointly, and identify this separation, rather than any lack of data, as a main technical bottleneck still limiting facial-based detection relative to the other modalities covered in this review. Adolescent-focused work is now a visible strand, spanning privacy-preserving challenges, school-based screening and student sentiment analysis (Du et al., 2023; Khan et al., 2025; Luo et al., 2026; Teng, Sun, et al., 2026; Zhang & Zhang, 2025).

3.9 Multimodal Fusion Approaches

Multimodal fusion is both the largest theme in this review and the one where methodological diversity is greatest. Of its 24 records, 17 are primary empirical studies (Aishwarya et al., 2026; Cenacchi et al., 2025; Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Du et al., 2023; Haque et al., 2025; Li, Yang, et al., 2025; Mandal et al., 2025; Pavithra & Trupthi, 2026; Rahman et al., 2025; Rajkumar et al., 2026; Ramadevi et al., 2026; Seneviratne & Espy-Wilson, 2022; Teng, Liu, et al., 2026; Teng, Sun, et al., 2026; Zhou et al., 2026), 4 are reviews or surveys (Khoo et al., 2024; Lai et al., 2023; Patil & Dixit, 2026; Shou et al., 2023), and 3 are benchmark or methodological papers, namely an adolescent screening challenge (Luo et al., 2026) and two general multimodal sentiment methods evaluated on general benchmarks rather than mental health data (Sun et al., 2022; Xue & Marculescu, 2022). Only the 17 primary studies inform the performance statements below. Within this set, three fusion strategies recur, namely concatenation-based early fusion, attention-based intermediate fusion of the kind formalised in Equations (7) and (8), and decision-level late fusion as in Equation (9). The DAIC-WOZ and E-DAIC corpora, introduced in the foundational studies underlying this

theme (Gratch et al., 2014; Valstar et al., 2016), remain the dominant evaluation benchmark, and a dedicated review of multimodal datasets for emotional AI situates these corpora within the wider landscape of available benchmarks (Al-Azani & El-Alfy, 2025).

Reported gains from fusion should be read with three qualifications, namely that the studies use different datasets, tasks and metrics, that most gains are reported on small benchmark corpora, and that few studies report uncertainty around their estimates. Table 6 therefore restricts the comparison to studies that report fused and single-modality or reduced-modality results within the same study, separated by task. For severity estimation, the trimodal model of Li, Yang, et al. (2025) reached a mean absolute error of 5.75 against 5.84 for the best bimodal combination, with every unimodal baseline performing worse, an improvement that is consistent but small. Tri-modal severity work across depression and post-traumatic stress disorder found that the most informative modality differed by condition, with text contributing most for depression and audio and facial cues more for PTSD (Cenacchi et al., 2025). For classification, question-wise analysis found that a text and video combination, rather than all three modalities, performed best on most interview questions (Mandal et al., 2025), and an interpretable attention-guided ensemble reported text as the strongest single channel (Rajkumar et al., 2026). Headline results reported without a matched unimodal comparison, such as precision, recall, F1-score and accuracy of 0.92, 0.90, 0.91 and 0.94 on DAIC-WOZ and 0.95, 0.85, 0.90 and 0.97 on E-DAIC for a modality decomposition and contrastive learning model (Chen et al., 2026), show what fused systems can achieve on these benchmarks but cannot on their own show how much of the gain comes from fusion. Performance also depends on the corpus, since the same bidirectional cross-modal attention design reached an F1-score of 0.820 on the EATD-Corpus and performed differently on DAIC-WOZ (Chao et al., 2026). Other designs extend fusion to articulatory coordination features with hierarchical text attention (Seneviratne & Espy-Wilson, 2022) and to eye-tracking, facial and acoustic features in a graph convolutional model (Rahman et al., 2025). Taken together, the evidence supports a narrower conclusion than is often stated, namely that adding modalities tends to help when this is tested within a study, that the benefit is often modest, and that which modalities help depends on the condition, the task and the data.

Table 6. *Within-study comparisons of fused and single-modality or reduced-modality models among primary fusion studies*

Study and status	Task and data	Comparison reported	What the comparison shows
Li, Yang, et al. (2025), peer-reviewed	Depression severity estimation from video, audio and text	Trimodal against each bimodal and unimodal model (MAE)	Trimodal best (5.75 against 5.84 for best bimodal); all unimodal models worse
Cenacchi et al. (2025), preprint	Severity estimation for depression and PTSD from three modalities	Contribution of each modality by condition	Text most informative for depression; audio and facial cues more informative for PTSD
Mandal et al. (2025), preprint	Depression detection from clinical interviews, question by question	Modality combinations per interview question	Text and video best on most questions; all three modalities not always best
Rajkumar et al. (2026), peer-reviewed	Depression classification from text, speech and video	Individual modalities within an attention-guided ensemble	Text the strongest single modality
Chao et al. (2026), peer-reviewed	Binary depression classification on EATD-Corpus and DAIC-WOZ	Same fusion design across two corpora	F1-score of 0.820 on EATD-Corpus; performance differs by corpus
Chen et al. (2025), preprint	Cross-domain depression recognition	Complete against missing-modality conditions	Models trained for missing modalities degrade less when a modality is dropped

Attention-guided and transformer-based fusion architectures, in the general sense formalised in Equation (8), now dominate the more recent studies in this theme. One system combines text, speech and video through an attention-guided ensemble specifically designed to remain interpretable, reporting that text alone is consistently the strongest individual modality in its dataset, which the authors attribute to the richer semantic and emotional content available in language compared with acoustic or visual cues alone (Rajkumar et al., 2026). A closely related framework applies the same reliability-aware attention principle to real-time stress and emotional state prediction, extending fusion beyond depression specifically to broader affective state monitoring using speech, EEG and facial cues together (Ramadevi et al., 2026), and a further pair of studies apply comparable attention-guided neural architectures to early diagnosis of mental health disorders more broadly (Aishwarya et al., 2026; Pavithra & Trupthi, 2026). This finding, that text tends to be the strongest single modality even within fusion-focused studies, recurs often enough across this theme to be treated as a genuine pattern rather than a one-off result, and it has direct implications for how future fusion architectures might be weighted, a point returned to in Section 4. The same preference for learned, input-dependent weighting appears in recent mutual, bidirectional cross-modal and question-wise designs (Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Haque et al., 2025; Mandal et al., 2025).

Several included studies within this theme are themselves systematic or survey-level treatments of fusion methodology rather than single new architectures, and are drawn on here as supporting synthesis: a comprehensive survey of multimodal conversational emotion recognition (Shou et al., 2023), a dedicated survey of multimodal sentiment analysis techniques more broadly (Lai et al., 2023), a systematic review of machine learning and deep learning approaches to mental health diagnosis specifically (Patil & Dixit, 2026), and a systematic review of passive sensing approaches to multimodal mental health detection, already discussed in Table 3 (Khoo et al., 2024). Two further studies address dynamic rather than static fusion specifically, proposing mechanisms that adjust how modalities are weighted at inference time rather than fixing the fusion strategy during training (Sun et al., 2022; Xue & Marculescu, 2022), which is conceptually the clearest demonstration of the attention-weighted formulation in Equation (8) in this corpus, although both are evaluated on general multimodal benchmarks rather than mental health data, so they illustrate the mechanism rather than its clinical value.

A smaller but conceptually important group of studies within this theme addresses graceful degradation: what happens to a fusion model's performance when one modality is missing or corrupted, which is a realistic concern for any system intended for real-world deployment rather than a clean laboratory benchmark. One study specifically targets cross-domain depression recognition under missing modalities and reports that models explicitly trained to handle this scenario degrade far less sharply than standard fusion architectures when a modality is dropped at test time (Chen et al., 2025), and explainability within fusion architectures is addressed directly by at least two included studies, which generate natural-language summaries or structured attribution reports alongside each prediction rather than a bare label and confidence score (Haque et al., 2025; Teng, Liu, et al., 2026). Related work extends fusion to co-occurring conditions, jointly modelling depression and post-traumatic stress disorder through a tri-modal severity-fused diagnostic approach (Cenacchi et al., 2025), applying psychology-inspired multimodal stimuli to differential diagnosis across several mental disorders at once (Zhou et al., 2026), and combining summary-guided cross-task fusion specifically for structured, scale-based mental health assessment (Teng, Sun, et al., 2026). A further study improves fusion through question-wise modality weighting, recognising that different interview questions elicit different amounts of signal from each modality (Mandal et al., 2025), and a prediction of depression using video, audio and text data jointly, discussed earlier in this section, rounds out the theme's coverage of the field's most heavily benchmarked corpora (Li, Yang, et al., 2025).

3.10 Methodological Conditions Behind Reported Performance

Across themes, the conditions under which results were obtained explain more of the variation in reported performance than architecture alone. Dataset size is the first condition. DAIC-WOZ, the most reused clinical corpus, contains interviews with 189 participants (Gratch et al., 2014), and its official test partition holds fewer than fifty (Valstar et al., 2016), so a change of one or two predictions moves the F1-score by several points. High scores on test sets this small carry wide uncertainty, yet confidence intervals and repeated runs are rarely

reported. Class imbalance is the second condition. Depressed participants are a minority in DAIC-WOZ and E-DAIC, and accuracy on these corpora overstates performance when the minority class is missed, which is why the F1-score is used for comparison in this review.

Validation design is the third condition, and the one most likely to inflate results. Many speech and video studies divide each interview into short segments, and if segments from the same participant fall into both training and test sets, a model can learn to recognise the speaker rather than the condition, a form of data leakage that produces optimistic results (Kapoor & Narayanan, 2023). Participant-level splitting or use of the official benchmark partitions avoids this, but the split level was not always stated in the primary studies. External testing is the fourth condition. Only a few primary studies evaluated on an independent corpus, for example across the EATD-Corpus and DAIC-WOZ (Chao et al., 2026) or across domains under missing modalities (Chen et al., 2025), so most reported performance reflects within-corpus fit. Missing modalities are the fifth. Real recordings often lack a usable channel, yet most fusion systems were trained and tested with every modality present, and the study that tested degradation directly found standard fusion architectures to be fragile (Chen et al., 2025). The comparative statements in this review rest on the higher and moderate quality studies.

The extent to which primary studies address the two gaps most often named in this field can also be counted. At least eight of the 42 primary studies treat explainability as an explicit design objective (Belcastro et al., 2026; Haque et al., 2025; Kaur et al., 2026; Kerz et al., 2023; Kim et al., 2025; Lanzlinger et al., 2026; Rajkumar et al., 2026; Teng, Liu, et al., 2026), and only one validates its explanations with clinicians (Kaur et al., 2026). For cross-cultural generalisation the picture is weaker still. Only a few primary studies evaluate on non-English or multilingual data, among them multilingual suicide text detection (Zevallos et al., 2024) and Mandarin interview data in the EATD-Corpus (Chao et al., 2026), and transfer between cultural or linguistic populations was rarely tested directly.

3.11 Cross-Cutting Observations

Three patterns cut across all four themes rather than belonging to any one of them. First, explainability has moved from an afterthought to an explicit design goal, but unevenly: the text-based theme has the most mature explainability literature, the facial and visual theme the least. Second, dataset reuse is extensive; DAIC-WOZ and its extension E-DAIC, FER2013, and various Reddit or Twitter derived corpora recur across dozens of the included studies, which strengthens comparability between studies but also means that much of this literature has been validated on a narrower range of populations and cultural contexts than the volume of published work might suggest. Third, and most directly relevant to this review's own topic, genuinely trimodal or higher fusion, combining text, speech and visual data together rather than any pair of the three, remains a minority practice even within the multimodal fusion theme itself, with most fusion studies still combining only two of the three core modalities. One further study analyses the sentiment and emotional tone of large language model responses to depression, anxiety and stress queries (VarastehNezhad et al., 2025), pointing to an emerging direction not captured by the four themes defined in Section 2.8, in which general-purpose language models become part of the mental health interaction itself and their affective behaviour, not only their detection accuracy, needs evaluation. These patterns are visible in recent explainability work (Belcastro et al., 2026; Ibrahimov et al., 2024; Lanzlinger et al., 2026; Mocnik et al., 2025; Teng, Liu, et al., 2026), in the repeated use of DAIC-WOZ, E-DAIC and EATD (Cenacchi et al., 2025; Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Haque et al., 2025), and in the scarcity of genuinely tri-modal systems (Cenacchi et al., 2025; Chen et al., 2025; Chen et al., 2026; Li, Yang, et al., 2025; Rahman et al., 2025). Section 4 discusses what each of these patterns implies for the field's near-term research priorities.

DISCUSSION

The bibliometric and thematic results tell the same story from two angles. Bibliometrically, the field is young, fast-growing and dominated by preprints: 34 of the 65 studies appeared in 2025 or 2026 and 43 are preprints (Cenacchi et al., 2025; Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Haque et al., 2025; Mandal et al., 2025; Teng, Sun, et al., 2026). Thematically, the effort has shifted from asking whether a single modality carries useful signal, which the text-based and speech-based themes largely answered years ago (Arji et al.,

2023; Khan et al., 2025; Li, Chng, & Guan, 2025; Lombardo et al., 2025; Ophir et al., 2020), toward asking how best to combine modalities, which is where the 24 fusion studies concentrate (Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Haque et al., 2025; Li, Yang, et al., 2025; Rahman et al., 2025). Citation impact tells a complementary story: it sits with reviews, benchmarks and a few text and speech studies (Barhoumi & BenAyed, 2024; Kerz et al., 2023; Khoo et al., 2024; Ophir et al., 2020; Valstar et al., 2016), while the fusion architectures that define the frontier have barely begun to be cited. Multimodal fusion is therefore not one approach among several in this field; it is where most of the new work is being placed.

A second pattern is the recurring finding that text is often the strongest single modality even within fusion architectures. One interpretable fusion system reports text as its strongest individual channel (Rajkumar et al., 2026), tri-modal severity work finds that text contributes most to depression severity while audio and facial cues matter more for PTSD (Cenacchi et al., 2025), and question-wise analysis finds that a text and video combination performs best on most questions (Mandal et al., 2025). Where adding a modality was tested within a study it tended to help, but the gains were modest and the best combination was not always the full set (Li, Yang, et al., 2025; Mandal et al., 2025). The common claim that multimodal fusion consistently outperforms unimodal approaches is therefore stronger than the evidence allows. A more defensible reading is that fusion is a promising design whose benefit depends on the condition, the task, the corpus and the validation design. Fusion work would benefit from reporting modality-wise ablations on the same data as standard practice, so that readers can judge what each added modality contributes in a given population and context.

Explainability is the clearest cross-cutting gap. Text-based detection has the most developed explainability literature (Belcastro et al., 2026; Ibrahimov et al., 2024; Kerz et al., 2023; Kim et al., 2025), yet even there formal validation that attention weights correspond to clinically meaningful signal is rare (Ibrahimov et al., 2024; Kaur et al., 2026). The facial and visual theme has almost none (Sahu et al., 2025; Shangguan et al., 2024; Zhang & Zhang, 2025), and only a handful of fusion studies produce natural-language summaries or structured attributions alongside predictions (Haque et al., 2025; Teng, Liu, et al., 2026). With at least eight of 42 primary studies treating explainability as a design goal and one validating explanations with clinicians, there is a mismatch between the field's stated purpose of supporting rather than replacing clinical judgement and its research output, and this is the area where the review most strongly recommends redirected effort.

Reliance on a few benchmark corpora also deserves a cautious reading. DAIC-WOZ, E-DAIC, EATD, D-Vlog and FER2013 recur across dozens of studies (Cenacchi et al., 2025; Chao et al., 2026; Chen et al., 2025; Chen et al., 2026; Haque et al., 2025; Mandal et al., 2025), which makes results comparable but means much of the literature's claimed generalisability rests on a narrow slice of populations, languages and cultures (Al-Azani & El-Alfy, 2025; Cao et al., 2024; Chen et al., 2025; Teng, Liu, et al., 2026; Zevallos et al., 2024). The keyword structure in Figures 7 and 9 reinforces the point: depression detection is the most frequent keyword by a wide margin, whereas anxiety appears in five studies and bipolar disorder in three (Cao et al., 2024; Cenacchi et al., 2025; Lombardo et al., 2025; Mocnik et al., 2025; Schlicher et al., 2025), so conclusions about multimodal detection are, in practice, conclusions about depression.

The results also need to be read against what clinical use would require. Nearly all the primary studies are research-stage classifiers evaluated retrospectively on benchmark or convenience data, and none of the coded studies reports prospective evaluation within a clinical pathway. They therefore show that a signal can be detected, not that a tool is fit for screening or diagnosis. Clinically validated screening requires prospective testing in the intended population, calibration of risk scores, agreement with structured clinical assessment, and evidence that use improves outcomes. The costs of error are also asymmetric. A false negative may leave a person at risk without support, which is most serious in suicide risk detection, while false positives at population scale can overload services, cause distress and label people who are well. Reporting sensitivity and specificity at clinically meaningful thresholds, rather than a single F1-score, would let readers judge these trade-offs (Chancellor & De Choudhury, 2020).

Multimodal systems also raise sharper ethical questions than text-only ones. Voice, facial video and physiological signals are biometric data that can identify a person and are treated as specially protected data under data protection law in many jurisdictions, so consent, storage and secondary use need stronger

safeguards than are usually described in the included studies. Inferring mental health status from social media without the knowledge of the people posting raises further questions of consent and autonomy (Chancellor et al., 2019). Emotional expression also differs across cultures, languages, genders and age groups, and a model trained on English-speaking adults in one interview setting may misread people outside that population, as the gender gap already observed in audio-based depression detection shows (Bailey, 2020). Applying such models outside their training populations, or in settings such as schools and workplaces where the person assessed has little power to refuse, risks harms that strong benchmark scores do not reveal. Future studies should report the demographic and linguistic composition of their training data, test for subgroup differences, and state the intended use of the model and its limits.

Returning to the research questions, the field has grown sharply since 2022, with citation impact concentrated in reviews and benchmark papers (RQ1). Multimodal fusion receives the most attention, followed by text-based work (RQ2). Primary studies in every category report strong results, but mostly on small, reused corpora, with limited external validation and inconsistent control of leakage and class imbalance, so established findings are fewer than the volume of work suggests (RQ3). DAIC-WOZ and E-DAIC, transformer encoders and attention-based intermediate fusion recur most often (RQ4). The main gaps are explainability, cross-population validation, robustness to missing modalities, coverage of conditions beyond depression, and clinical evaluation (RQ5). Of these findings, the growth of the field, the dominance of fusion and the reliance on a few corpora are well established, whereas the size of the fusion advantage, the value of large language model-based detection and the clinical usefulness of current systems remain emerging questions.

This review has limitations of its own. Two-thirds of the records are preprints whose results may change in peer review, and although a peer-reviewed sensitivity check was run, the peer-reviewed subset is small. Quality appraisal relied on what studies reported, so poor reporting cannot be distinguished from poor practice. OpenAlex lacked records for nine studies, which makes the citation analysis incomplete and probably understates the impact of recent work. The search was limited to English and to six databases, which may under-represent work from non-English-speaking regions and so reinforce the cultural narrowness the review itself identifies. Finally, the four-theme structure was defined a priori, and directions that cut across it may be under-described.

FUTURE RESEARCH DIRECTIONS

The patterns in Section 4 translate into eight directions for future work, summarised in Table 7 alongside the gap each responds to. They draw on open problems that recent reviews and studies identify, namely explainable fusion (Belcastro et al., 2026; Ibrahimov et al., 2024; Lanzlinger et al., 2026; Mocnik et al., 2025; Teng, Liu, et al., 2026), cross-corpus and cross-lingual evaluation (Al-Azani & El-Alfy, 2025; Cao et al., 2024; Chen et al., 2025; Zevallos et al., 2024), robustness to missing modalities (Chao et al., 2026; Chen et al., 2025; Haque et al., 2025; Mandal et al., 2025), rigorous validation (Kapoor & Narayanan, 2023), conditions beyond depression (Cenacchi et al., 2025; Guo et al., 2024; Khoo et al., 2024; Lombardo et al., 2025), and clinically and ethically grounded evaluation (Chancellor et al., 2019; Chancellor & De Choudhury, 2020).

Table 7. *Proposed future research agenda*

Identified gap	Suggested research direction	Suggested approach
Explainability concentrated in text-based studies	Extend validated explainability methods to speech-based and facial/visual detection	Attention visualisation, feature attribution, clinician validation studies
Few true trimodal (text, speech, visual) systems	Report systematic modality-wise ablations in all new fusion work	Trimodal architectures with unimodal and bimodal baselines reported as standard
Heavy reliance on DAIC-WOZ, E-DAIC and FER2013	Evaluate on culturally and linguistically distinct corpora	Cross-corpus and cross-lingual validation studies
Depression dominates; anxiety, PTSD and bipolar disorder are rare	Build and share multimodal corpora for other conditions	Differential and comorbid detection studies

Most evidence is preprint, and recent work is uncited	Encourage peer-reviewed replication of high-performing preprint architectures	Journal-targeted replication and extension studies; re-run citation analysis annually
Limited handling of missing or corrupted modalities	Report robustness under missing-modality conditions as standard	Graceful degradation testing following the missing-modality studies in this corpus
Validation practices vary and leakage risk is often unreported	Make participant-level splits, external test sets and uncertainty reporting standard	Subject-independent cross-validation, held-out corpora, confidence intervals or repeated runs
Research-stage classifiers are rarely evaluated clinically, and ethical safeguards are under-reported	Move from retrospective benchmarks to prospective, consented evaluation	Prospective validation against structured clinical assessment, subgroup fairness audits and data governance reporting

CONCLUSION

This review combined a bibliometric analysis with a thematic synthesis to map research on mental health detection using multimodal deep learning. From 9,044 records identified across six databases, 65 records met the eligibility criteria under PRISMA 2020 (Page et al., 2021), comprising 42 primary empirical studies, 18 reviews and 5 benchmark or methodological papers. The field is young and growing quickly, with just over half of the records published in 2025 and 2026, multimodal fusion is its largest theme, and citation impact is concentrated in a few reviews and benchmark papers while the newest architectures have yet to be cited. Where fused and single-modality models were compared within the same study, fusion usually matched or exceeded the best single channel, but the gains were often modest, the best combination depended on the condition and task, and the evidence rests largely on preprints evaluated on small, reused corpora with little external validation. These limitations of the evidence base, together with the review's own reliance on reported information and on partial citation coverage, mean that current results show that mental health signals can be detected across modalities, not that multimodal systems are ready for clinical screening. Explainability, participant-level and cross-corpus validation, robustness to missing modalities, coverage of conditions beyond depression, and clinically and ethically grounded evaluation are the most pressing open problems, and the agenda in Section 5 is offered as a concrete starting point.

REFERENCES

1. Aishwarya, V., Venkataraman, D., & Shunmuga Priya, M. C. (2026). Multimodal system for monitoring mental and emotional state with deep learning model. In I. Bhiradi & J. Machado (Eds.), *Intelligent control, robotics, and industrial automation* (pp. 585-597). Springer. https://doi.org/10.1007/978-981-95-2901-8_44
2. Al-Azani, S., & El-Alfy, E. S. M. (2025). A review and critical analysis of multimodal datasets for emotional AI. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-025-11271-1
3. Alhussein, G., Ziogas, I., Saleem, S., & Hadjileontiadis, L. J. (2025). Speech emotion recognition in conversations using artificial intelligence: A systematic review and meta-analysis. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-025-11197-8
4. Alvarez Casado, C., Lage Canellas, M., & Bordallo Lopez, M. (2022). *Depression recognition using remote photoplethysmography from facial videos* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2206.04399
5. Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007
6. Arji, G., Erfannia, L., Alirezaei, S., & Hemmat, M. (2023). A systematic literature review and analysis of deep learning algorithms in mental disorders. *Informatics in Medicine Unlocked, 40*, Article 101284. https://doi.org/10.1016/j.imu.2023.101284
7. Bailey, A. (2020). *Gender bias in depression detection using audio features* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2010.15120

8. Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607
9. Barhoumi, C., & BenAyed, Y. (2024). Real-time speech emotion recognition using deep learning and data augmentation. *Artificial Intelligence Review, 58*, Article 49. https://doi.org/10.1007/s10462-024-11065-x
10. Belcastro, L., Gervino, F., Marozzo, F., Talia, D., & Trunfio, P. (2026). *Explainable detection of depression status shifts from user digital traces* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.14995
11. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa
12. Cao, Y., Dai, J., Wang, Z., Zhang, Y., Shen, X., Liu, Y., & Tian, Y. (2024). *Machine learning approaches for mental illness detection on social media: A systematic review of biases and methodological challenges* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2410.16204
13. Cenacchi, F., Richards, D., & Cao, L. (2025). *Tri-modal severity fused diagnosis across depression and post-traumatic stress disorders* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2510.20239
14. Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M. B., & De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 79-88). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287587
15. Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine, 3*, Article 43. https://doi.org/10.1038/s41746-020-0233-7
16. Chao, C. M., Yuen, K. T., & Du, W. (2026). Bidirectional cross-modal attention with adaptive thresholding for multimodal depression detection across the EATD-Corpus and DAIC-WOZ datasets. *BioMedInformatics, 6*(5), Article 70. https://doi.org/10.3390/biomedinformatics6050070
17. Chen, J., Ning, Y., Zhang, P., Chen, H., & Zheng, H. (2026). Multimodal depression detection based on modality feature decomposition and contrastive learning. *Algorithms, 19*(6), Article 452. https://doi.org/10.3390/a19060452
18. Chen, J., Tan, M., Lu, H., Xu, Q., Wang, Z., Zeng, R., & Hu, X. (2025). *Towards stable cross-domain depression recognition under missing modalities* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2512.06447
19. Chlasta, K., Wolk, K., & Krejtz, I. (2019). *Automated speech-based screening of depression using deep convolutional neural networks* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1912.01115
20. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
21. Du, Z., Jiang, C., Yuan, X., Zhai, S., Lei, Z., Ma, S., Liu, Y., Ye, Q., Xiao, C., Huang, Q., Xu, M., Yu, D., & Qin, P. (2023). *GAME: Generalized deep learning model towards multimodal data integration for early screening of adolescent mental disorders* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.10077
22. van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3
23. Farruque, N., Goebel, R., Sivapalan, S., & Zaiane, O. (2022). *Deep temporal modelling of clinical depression through social media text* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2211.07717
24. Gajjar, J., & Ranaware, K. (2025). *Multimodal sentiment analysis on CMU-MOSEI dataset using transformer-based models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.06110
25. Garcia, E., Gomes, J., Barbosa Junior, A., Borges, C., & da Silva, N. (2023). *DeepLearningBrasil@LT-EDI-2023: Exploring deep learning techniques for detecting depression in social media text* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2311.05047
26. Garg, M. (2021). *Quantifying the suicidal tendency on social media: A survey* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2110.03663

27. Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., & Morency, L. P. (2014). The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 3123-3128). European Language Resources Association.
28. Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). *Large language model for mental health: A systematic review* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2403.15401
29. Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PLOS ONE, 10*(9), Article e0138237. https://doi.org/10.1371/journal.pone.0138237
30. Haque, M. R., Islam, M. M., Raju, S. M. T. U., Altaheri, H., Nassar, L., & Karray, F. (2025). *MDD-Net: Multimodal depression detection through mutual transformer* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2508.08093
31. Hassan, A. A., Ali, A. A., Fouda, A. E., Hanafy, R. J., & Fouda, M. E. (2025). *Leveraging embedding techniques in multimodal machine learning for mental illness assessment* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2504.01767
32. Haunschild, R., & Bornmann, L. (2024). The use of OpenAlex to produce meaningful bibliometric global overlay maps of science on the individual, institutional, and national levels. *PLOS ONE, 19*(12), Article e0308041. https://doi.org/10.1371/journal.pone.0308041
33. He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., Guo, C., Wang, H., Ding, S., Wang, Z., Pan, X., & Dang, W. (2021). *Deep learning for depression recognition with audiovisual cues: A review* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2106.00610
34. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
35. Hong, Q. N., Pluye, P., Bujold, M., & Vedel, I. (2018). *Mixed Methods Appraisal Tool (MMAT), version 2018* (Registration of Copyright No. 1148552). Canadian Intellectual Property Office.
36. Ibrahimov, Y., Anwar, T., & Yuan, T. (2024). *Explainable AI for mental disorder detection on social media: A survey and outlook* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2406.05984
37. Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2019). *Suicidal ideation detection: A review of machine learning methods and applications* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1910.12611
38. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804
39. Kaur, A., Singh, P., Chatterjee, N., Shubneet, Yadav, A. R., & Talwandi, N. S. (2026). Explainable AI for early detection of mental health disorders using social media data: A hybrid deep learning approach with clinician-in-the-loop validation. In A. Swaroop, B. Virdee, S. D. Correia, & Z. Polkowski (Eds.), *Proceedings of Data Analytics and Management: ICDAM 2025* (pp. 504-514). Springer. https://doi.org/10.1007/978-3-032-03751-0_42
40. Kerz, E., Zanwar, S., Qiao, Y., & Wiechmann, D. (2023). Toward explainable AI (XAI) for mental health detection based on language behavior. *Frontiers in Psychiatry, 14*, Article 1219479. https://doi.org/10.3389/fpsyt.2023.1219479
41. Khan, H. U., Naz, A., Alarfaj, F. K., & Almusallam, N. (2025). Analyzing student mental health with RoBERTa-Large: A sentiment analysis and data analytics approach. *Frontiers in Big Data*. https://doi.org/10.3389/fdata.2025.1615788
42. Khoo, L. S., Lim, M. K., Chong, C. Y., & McNaney, R. (2024). Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches. *Sensors, 24*(2), Article 348. https://doi.org/10.3390/s24020348
43. Kim, S., Imieye, O., & Yin, Y. (2025). *Interpretable depression detection from social media text using LLM-derived embeddings* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2506.06616
44. Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (EBSE Technical Report EBSE-2007-01). Keele University and University of Durham.
45. Lai, S., Xu, H., Hu, X., Ren, Z., & Liu, Z. (2023). *Multimodal sentiment analysis: A survey* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2305.07611

46. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310
47. Lanzlinger, J., Muller, K. O. E., Stiller, B., & Rodrigues, B. (2026). *Towards interpretable depression detection: Linking acoustic features to DSM-5 indicators* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2608.26148
48. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. https://doi.org/10.1038/nature14539
49. Li, Y., Chng, E. S., & Guan, C. (2025). *Hierarchical self-supervised representation learning for depression detection from speech* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2510.08593
50. Li, Y., Yang, X., Zhao, M., Wang, J., Yao, Y., Qian, W., & Qi, S. (2025). Predicting depression by using a novel deep learning model and video-audio-text multimodal data (IMDD-Net). *Frontiers in Psychiatry*. https://doi.org/10.3389/fpsyt.2025.1602650
51. Liu, D., Liu, B., Lin, T., Liu, G., Yang, G., Qi, D., Qiu, Y., Lu, Y., Yuan, Q., Shuai, S. C., Li, X., Liu, O., Tang, X., Shuai, J., Cao, Y., & Lin, H. (2022). Measuring depression severity based on facial expression and body movement using deep convolutional neural network. *Frontiers in Psychiatry*. https://doi.org/10.3389/fpsyt.2022.1017064
52. Lombardo, C., Esposito, G., Carbone, S., Serrano, S., & Mento, C. (2025). Speech analysis and speech emotion recognition in mental disease: A scoping review. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2025.1645860
53. Long, S., Cabral, R., Poon, J., & Han, S. C. (2022). *A quantitative and qualitative analysis of suicide ideation detection using deep learning* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2206.08673
54. Luo, Z., Wang, J., Qi, T., Wu, Y., Matsui, T., Wu, S., Zhao, X., Qian, K., Yue, Z., Takiguchi, T., Wang, F., & Ishiguro, H. (2026). *AdoDAS: A privacy-preserving multimodal challenge for adolescent depression, anxiety, and stress assessment* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2609.07038
55. Mandal, A., Atzil-Slonim, D., Solorio, T., & Gurevych, I. (2025). *Enhancing depression detection via question-wise modality fusion* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.20496
56. Mocnik, G., Rehberger, A., Smogavc, Z., Mlakar, I., Smrke, U., & Mocnik, S. (2025). Multimodal observable cues in mood, anxiety, and borderline personality disorders: A review of reviews to inform explainable AI in mental health. *Frontiers in Artificial Intelligence, 8*, Article 1696448. https://doi.org/10.3389/frai.2025.1696448
57. Naderi, H., Soleimani, B. H., Matwin, S., Rempel, S., & Uher, R. (2019). *Multimodal deep learning for mental disorders prediction from audio speech samples* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1909.01067
58. Nguyen, V. (2024). *Leveraging large language models for suicide detection on social media with limited labels* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2410.04501
59. Ophir, Y., Tikochinski, R., Asterhan, C., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual Facebook posts. *Scientific Reports, 10*, Article 16685. https://doi.org/10.1038/s41598-020-73917-0
60. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71
61. Patil, S. C., & Dixit, M. (2026). A systematic review of machine learning and deep learning for mental health diagnosis. *Smart Innovation, Systems and Technologies, 454*, 339-348. https://doi.org/10.1007/978-3-032-07837-7_25
62. Pavithra, R. S., & Trupthi, M. (2026). Multimodal emotion-aware neural networks for early diagnosis of mental health disorders. In A. Kumar, G. Ghinea, & S. Merugu (Eds.), *Proceedings of the 4th International Conference on Cognitive and Intelligent Computing, Volume 2 (ICCIC 2024)*. Springer. https://doi.org/10.1007/978-981-95-0144-1_19
63. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98-125. https://doi.org/10.1016/j.inffus.2017.02.003

64. Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2205.01833
65. Rahman, S., Deb, S., Chowdhury, M. S. I., Sourov, M. J. A., & Shamsuddin, M. (2025). *MF-GCN: A multi-frequency graph convolutional network for tri-modal depression detection using eye-tracking, facial, and acoustic features* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2511.15675
66. Rajkumar, K., Sunkavalli, J., Vadlamani, S., Uddagiri, C., Varma, P., & Anitha, R. (2026). MoodSenseAI: An attention-guided multi-source multimodal deep learning system for interpretable depression detection from text, speech and video. *Discover Computing, 29*(1). https://doi.org/10.1007/s10791-026-10177-y
67. Ramadevi, Y., Shailaja, T., Gireesh, C., Valaboju, S., Raja Sekar, M., . . . Aakuthota, N. K. (2026). Reliability-aware multimodal deep learning framework for real-time stress and emotional state prediction using speech, EEG and facial cues. *Discover Computing, 29*, Article 442. https://doi.org/10.1007/s10791-026-10338-z
68. Saeed, Q. B., & Ahmed, I. (2025). *Early detection of mental health issues using social media posts* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.07653
69. Sahu, N. K., Harshit, N. S., Uikey, R., & Lone, H. R. (2025). *Beyond questionnaires: Video analysis for social anxiety detection* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2501.05461
70. Schlicher, M., Li, Y., Murthy, S. M. K., Sun, Q., & Schuller, B. W. (2025). Emotionally adaptive support: A narrative review of affective computing for mental health. *Frontiers in Digital Health*. https://doi.org/10.3389/fdgth.2025.1657031
71. Seneviratne, N., & Espy-Wilson, C. (2022). *Multimodal depression classification using articulatory coordination features and hierarchical attention based text embeddings* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2202.06238
72. Shangguan, Z., Dong, Y., Guo, S., & Leung, V. C. (2024). *Facial expression analysis and its potentials in IoT systems: A contemporary survey* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2412.17616
73. Shou, Y., Meng, T., Ai, W., Fu, F., Yin, N., & Li, K. (2023). *A comprehensive survey on multi-modal conversational emotion recognition with deep learning* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2312.05735
74. Sun, L., Lian, Z., Liu, B., & Tao, J. (2022). *Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2208.07589
75. Teng, S., Liu, J., Sun, H., Li, Y., Chai, S., Hou, R., Tateyama, T., Lin, L., & Chen, Y. W. (2026). *Dynamic summary generation for interpretable multimodal depression detection* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2604.11334
76. Teng, S., Sun, H., Hou, R., Yu, H., Song, Y., Lin, L., Liu, J., Chai, S., & Chen, Y. W. (2026). *DynaBridge: Dynamic summary-guided cross-task multimodal fusion for DASS-structured mental health assessment* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.25679
77. Thelwall, M., & Jiang, X. (2025). Is OpenAlex suitable for research quality evaluation and which citation indicator is best? *Journal of the Association for Information Science and Technology, 76*(12), 1660-1681. https://doi.org/10.1002/asi.70020
78. Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*(3), 207-222. https://doi.org/10.1111/1467-8551.00375
79. Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres, M., Scherer, S., Stratou, G., Cowie, R., & Pantic, M. (2016). *AVEC 2016: Depression, mood, and emotion recognition workshop and challenge* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1605.01600
80. VarastehNezhad, A., Tavasoli, R., Elyasi, S., LotfiNia, M., & Farbeh, H. (2025). *AI in mental health: Emotional and sentiment analysis of large language models' responses to depression, anxiety, and stress queries* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2508.11285
81. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates. https://doi.org/10.48550/arXiv.1706.03762
82. Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of

prediction model studies. *Annals of Internal Medicine, 170*(1), 51-58. https://doi.org/10.7326/M18-1376

83. Woodward, K., Kanjo, E., Brown, D., McGinnity, T. M., Inkster, B., Macintyre, D., & Tsanas, A. (2019). *Beyond mobile apps: A survey of technologies for mental well-being* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1905.00288
84. World Health Organization. (2022). *Mental disorders* [Fact sheet]. https://www.who.int/news-room/fact-sheets/detail/mental-disorders
85. Xie, W., Liang, L., Lu, Y., Luo, H., & Liu, X. (2021). *Deep 3D-CNN for depression diagnosis with facial video recording of self-rating depression scale questionnaire* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2107.10712
86. Xue, Z., & Marculescu, R. (2022). *Dynamic multimodal fusion* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2204.00102
87. Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 2236-2246). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1208
88. Zevallos, R., Schoene, A., & Ortega, J. (2024). *The first multilingual model for the detection of suicide texts* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2412.15498
89. Zhang, Z., & Zhang, L. (2025). Emotion dysregulation as a marker in adolescent mental health with EEG-based prediction model. *Scientific Reports, 15*, Article 38178. https://doi.org/10.1038/s41598-025-22067-2
90. Zhou, Z., Wu, J., Lei, Z., Guo, J., Yu, Z., Chu, Y., Zhang, X., Zhao, Q., Wang, Q., Hao, S., Guo, Y., & Hong, R. (2026). *Differential mental disorder detection with psychology-inspired multimodal stimuli* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2604.02798

Appendix A. Included Records by Theme, Evidence Strand and Source Type

Table A1 lists all 65 included records. Theme gives the thematic or supplementary bibliometric class used in Figure 5 and Table 5, and evidence strand follows the definitions in Table 2.

Table A1. *Included records by theme, evidence strand and source type*

Record	Theme	Evidence strand	Source type
Aishwarya et al. (2026)	Multimodal fusion	Primary empirical	Book or conference chapter
Al-Azani and El-Alfy (2025)	Foundational or benchmark	Secondary review	Journal article
Alhussein et al. (2025)	Speech and audio	Secondary review	Journal article
Alvarez Casado et al. (2022)	Facial and visual	Primary empirical	arXiv preprint
Arji et al. (2023)	Text-based	Secondary review	Journal article
Bailey (2020)	Speech and audio	Primary empirical	arXiv preprint
Barhoumi and BenAyed (2024)	Speech and audio	Primary empirical	Journal article
Belcastro et al. (2026)	Text-based	Primary empirical	arXiv preprint
Cao et al. (2024)	General cross-modality	Secondary review	arXiv preprint
Cenacchi et al. (2025)	Multimodal fusion	Primary empirical	arXiv preprint
Chao et al. (2026)	Multimodal fusion	Primary empirical	Journal article
Chen et al. (2025)	Multimodal fusion	Primary empirical	arXiv preprint
Chen et al. (2026)	Multimodal fusion	Primary empirical	Journal article
Chlasta et al. (2019)	Foundational or benchmark	Primary empirical	arXiv preprint
Du et al. (2023)	Multimodal fusion	Primary empirical	arXiv preprint
Farruque et al. (2022)	Text-based	Primary empirical	arXiv preprint
Gajjar and Ranaware (2025)	Foundational or benchmark	Benchmark or methodological	arXiv preprint
Garcia et al. (2023)	Text-based	Primary empirical	arXiv preprint
Garg (2021)	Text-based	Secondary review	arXiv preprint
Guo et al. (2024)	General cross-modality	Secondary review	arXiv preprint
Haque et al. (2025)	Multimodal fusion	Primary empirical	arXiv preprint
Hassan et al. (2025)	Speech and audio	Primary empirical	arXiv preprint
He et al. (2021)	Speech and audio	Secondary review	arXiv preprint
Ibrahimov et al. (2024)	Text-based	Secondary review	arXiv preprint
Ji et al. (2019)	Text-based	Secondary review	arXiv preprint
Kaur et al. (2026)	Text-based	Primary empirical	Book or conference chapter
Kerz et al. (2023)	General cross-modality	Primary empirical	Journal article
Khan et al. (2025)	Text-based	Primary empirical	Journal article
Khoo et al. (2024)	Multimodal fusion	Secondary review	Journal article
Kim et al. (2025)	Text-based	Primary empirical	arXiv preprint
Lai et al. (2023)	Multimodal fusion	Secondary review	arXiv preprint
Lanzlinger et al. (2026)	Speech and audio	Primary empirical	arXiv preprint
Li, Chng, and Guan (2025)	Speech and audio	Primary empirical	arXiv preprint
Li, Yang, et al. (2025)	Multimodal fusion	Primary empirical	Journal article
Liu et al. (2022)	Facial and visual	Primary empirical	Journal article

Lombardo et al. (2025)	Speech and audio	Secondary review	Journal article
Long et al. (2022)	Text-based	Primary empirical	arXiv preprint
Luo et al. (2026)	Multimodal fusion	Benchmark or methodological	arXiv preprint
Mandal et al. (2025)	Multimodal fusion	Primary empirical	arXiv preprint
Mocnik et al. (2025)	General cross-modality	Secondary review	Journal article
Naderi et al. (2019)	Speech and audio	Primary empirical	arXiv preprint
Nguyen (2024)	Text-based	Primary empirical	arXiv preprint
Ophir et al. (2020)	Text-based	Primary empirical	Journal article
Patil and Dixit (2026)	Multimodal fusion	Secondary review	Book or conference chapter
Pavithra and Trupthi (2026)	Multimodal fusion	Primary empirical	Book or conference chapter
Rahman et al. (2025)	Multimodal fusion	Primary empirical	arXiv preprint
Rajkumar et al. (2026)	Multimodal fusion	Primary empirical	Journal article
Ramadevi et al. (2026)	Multimodal fusion	Primary empirical	Journal article
Saeed and Ahmed (2025)	Text-based	Primary empirical	arXiv preprint
Sahu et al. (2025)	Facial and visual	Primary empirical	arXiv preprint
Schlicher et al. (2025)	General cross-modality	Secondary review	Journal article
Seneviratne and Espy-Wilson (2022)	Multimodal fusion	Primary empirical	arXiv preprint
Shangguan et al. (2024)	Facial and visual	Secondary review	arXiv preprint
Shou et al. (2023)	Multimodal fusion	Secondary review	arXiv preprint
Sun et al. (2022)	Multimodal fusion	Benchmark or methodological	arXiv preprint
Teng, Liu, et al. (2026)	Multimodal fusion	Primary empirical	arXiv preprint
Teng, Sun, et al. (2026)	Multimodal fusion	Primary empirical	arXiv preprint
Valstar et al. (2016)	Foundational or benchmark	Benchmark or methodological	arXiv preprint
VarastehNezhad et al. (2025)	General cross-modality	Primary empirical	arXiv preprint
Woodward et al. (2019)	Speech and audio	Secondary review	arXiv preprint
Xie et al. (2021)	Facial and visual	Primary empirical	arXiv preprint
Xue and Marculescu (2022)	Multimodal fusion	Benchmark or methodological	arXiv preprint
Zevallos et al. (2024)	Text-based	Primary empirical	arXiv preprint
Zhang and Zhang (2025)	Facial and visual	Primary empirical	Journal article
Zhou et al. (2026)	Multimodal fusion	Primary empirical	arXiv preprint