



Contents lists available at [Journal ELORA](#)

JRTI (Jurnal Riset Tindakan Indonesia)

ISSN: 2502-079X (Print) ISSN: 2503-1619 (Electronic)

Journal homepage: <https://jrti.eloracenter.org/jrti>



Reading the unspoken: a systematic review of multimodal artificial intelligence for detecting psychological and affective states from nonverbal cues

Ifdil Ifdil^{1*}, Nur Adila Wafiqoh Zulvi¹, M. Fahli Zatrachadi²

¹ Universitas Negeri Padang, Padang, Indonesia

² Universitas Islam Negeri Sultan Syarif Kasim Riau, Riau, Indonesia

Article Info

Article history:

Received Jul 12th, 2025

Revised Aug 20th, 2025

Accepted Sep 28th, 2025

Keyword:

Affective computing

Emotion recognition

Multimodal artificial intelligence

Nonverbal cues

Psychological assessment

ABSTRACT

The psychological assessment is limited to only relying on the patients' self-reports, as they may not even understand or have the desire or capacity to express the distress they show with their body language such as facial movements, voice, gaze, posture, and physiology. This systematic literature review brings together the findings of multimodal artificial intelligence (AI), mainly the use of such a system in detecting psychological and affective states from nonverbal cues and examines whether it is ready for mental-health and counseling applications. Using PRISMA 2020, we identified the records from a Scopus search, then screened them by the predefined eligibility criteria and evaluated them with the tool for Mixed-Method Appraisal. The original list of 669 articles was reduced first by removing the duplicates to 665 screened articles, and among them, we selected 20 articles released in 2024, 2025 that fulfilled the criteria. The themes identified in the five modality Channels facial & micro expression, speech & acoustic, text & linguistic, physiologic, and body movement together with their amalgamation were compared against the target states which were mainly depression, general emotion, and concealed affect. Multimodal approach generally yielded better results than unimodal methods and recognized concealed distress and suicidality, but the research is limited in scope, data sets being very small, external validation is missing and ethical precautions are not even considered yet.



© 2025 The Authors.

This is an open access article under the CC BY-NC-SA license
(<https://creativecommons.org/licenses/by-nc-sa/4.0>)

Corresponding Author:

Ifdil Ifdil,

Universitas Negeri Padang, Padang, Indonesia

Email: ifdil@konselor.org

Introduction

Mental health disorders rank amongst the biggest and fastest-growing health issues internationally, yet a big chunk of the cases are never discovered or the diagnosis comes too late. Traditional psychological evaluation largely depends on two reliable but subjective methods, self rating scales and clinical interviews, both of which are vulnerable to recall bias, social desirability, limited insight and the constraints of clinician time. As a consequence, an increasing number of computer-based efforts are being directed at finding nonverbal behavioral characteristics that would reflect the psychological state of the person. They might, for example, consist of different verbal and non-verbal cues (see e.g., Muzammel et al., 2021; Rejaibi et al., 2022; Cascella et al., 2024).

These studies are driven by the desire to develop a scalable and objective screening that would work as the professional judgment's aid, not a replacement. In fact, the rationale for a look beyond words in a clinical setting

is very much supported: much of what people go through is shown through their body rather than the spoken word and it seems that one may hide affect or conceal it and still it might "leak" through very short unintentional cues. Face micro-expressions is one such example that is very often associated with revealing real emotions people try to hide which is why psychotherapy and diagnosis can benefit from it (See Singh & Malik, 2025; Shuai et al., 2025). In some psychotherapy situations, the therapists have analyzed the non-verbal cues, i.e., expressions of the face, voice, and body at the same time, to determine the emotional nature of the interactions, in one case even psychodynamic play therapy (Halfon et al., 2021). As a result, nonverbal behavior, according to these remarks, can be seen as a source of information that is not only deep but also has been left almost untouched for purposes of assessment.

Affective computing has progressed from making rules for each feature separately and analyzing just one modality at a time to getting emotion patterns by deep learning models based on human face and speech (Thakur et al., 2025). More precisely, in mental health area, machine learning approaches have been aimed at identifying depressed individuals via features such as speech patterns and attention levels (Rejaibi et al., 2022; Park & Moon, 2022; Du et al., 2023), and meanwhile, larger surveys have been presenting various data collections, approaches, and experimental settings for analysis of cognitive behavioral aspects (E.g., Bhatt et al., 2023). However, these studies, although promising, are still pretty much separated by their modalities, with each study having its own datasets, evaluation methods, and reporting standards.

In more recent research there is a stronger emphasis placed on combining signals from various sources and expanding the use of the sensor network. One system that can continuously recognize the emotions of the patient fuses visual and physiological signals obtained through their faces (Mutawa & Hassouneh, 2024). When both facial and speech modalities are used together they are more successful than when they are used separately in identifying multidimensions of emotions (Tomar et al., 2024). The possibility of wearing devices, or self-monitoring, is opening up areas outside of clinics with examples such as wrist-mounted devices picking up the change in skin resistance for detecting stress and analyzing of attention (Zhu et al., 2023) and attention-fusion networks for the detection of mental-emotions (Xu et al., 2025) as well as multimodal-facial networks (Nemati et al., 2025). Therefore, the merging of different nonverbal channels is probably the main direction now.

Although there is a lot of research activity in this direction, the level of integration of evidence between different modality-based studies is quite low, and that's hampering the application of these modalities in the real clinics. Very few studies are able to map nonverbal features (and modalities) to their fusion techniques and target psychological states. Clinical populations are not equally distributed in studies, for example, impairment of one's ability to recognize emotions in face and voice has been one of the most important diagnostic characteristics of schizophrenia, see Zhao et al., 2025. Other markers proposed such as projective drawing also require further validation (Liu & Ye, 2025). It is true, that altogether there are indications that, besides having advantages in multimodal nature of the tests over the unimodal, the evidence is limited by the degree of variability in results, the lack of cross-study and external validation of the findings (Wang et al., 2025).

The second type of limitations is methodological and translational aspects. The availability of depression and emotional state data is often limited. As it motivates the use of privacy-preserving methods such as federated learning (Simć et al., 2024) and the generation of syntheses data across modalities to augment the medical data (Ibrahim et al., 2025). Longitudinal, real-world deployment for purposes of monitoring and relapse prevention has been shown only in limited settings (Zlatintsi et al., 2022). Standards for determining the validity, transparency, and how outputs should be framed to be communicative for affective outputs have not been fully developed (Resende Faria et al., 2024). It would be hard to assess whether these systems represent a sufficiently mature point to be employed responsibly in counseling and assessment of mental health through absence of synthesis. On the one hand, this article combines the latest available evidence regarding the use of multimodal AI techniques for identifying psychological states and affective states from silent cues

Apart from modalities, this article will focus also on target states, performance, and readiness for implementation while it is a good idea for us to keep the fact that there has been a dramatic increase in publications about this topic since 2023 on one side, and also that clinicians nowadays rely rather on the ability to piece different pieces of knowledge together than on proving that something is possible on the other. Guided by the PRISMA 2020 framework, the review addresses three research questions:

RQ1. Which nonverbal modalities and multimodal fusion strategies have been used by AI systems to detect psychological and affective states?

RQ2. Which psychological and affective target states are most frequently assessed, and what does the evidence indicate about detection performance?

RQ3. What methodological and translational limitations constrain deployment in mental-health and counseling assessment, and what future directions emerge?

By answering these questions, the review contributes an integrative modality-by-target map of the field and a prioritized research agenda for moving multimodal nonverbal assessment from proof-of-concept toward validated, ethical clinical support.

Method

Research Design and Framework

A systematic literature review (SLR) design was adopted to identify, appraise, and synthesize evidence in a transparent and reproducible manner, following established guidance for evidence-informed synthesis (Tranfield et al., 2003; Liberati et al., 2009). Reporting adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement and its explanation and elaboration (Page et al., 2021; Page, Moher, et al., 2021). The review was structured around the three research questions stated above and prioritized the mapping of methods and outcomes rather than a single pooled effect estimate.

Search Strategy and PICOS Framework

Search concepts were structured using the PICOS framework (Population, Intervention/Interest, Comparison, Outcome, Study design) to ensure systematic identification of key terms (Methley et al., 2014). Table 1 summarizes the framework. Each element was expanded into synonym clusters and combined with Boolean operators and truncation to capture lexical variants (e.g., depress*, emotion*, multimodal*). Field codes restricted matching to titles, abstracts, and keywords.

Table 1. <PICOS framework used to structure the search strategy.>

Element	Specification
Population (P)	Individuals assessed for psychological or affective states, including clinical (e.g., depression, bipolar disorder) and non-clinical populations
Interest (I)	Artificial-intelligence / machine-learning systems that analyze nonverbal cues (facial, vocal, linguistic, physiological, gaze, posture) individually or in multimodal fusion
Comparison (C)	Unimodal versus multimodal approaches, or AI-derived indices versus conventional scale/interview assessment (where reported; not required)
Outcome (O)	Detection, recognition, screening, or estimation of a psychological or affective state (e.g., depression, emotion, concealed affect, engagement)
Study design (S)	Peer-reviewed empirical studies and reviews (2020–2025), English language

The Boolean search string combined the three core concept blocks-artificial intelligence, nonverbal/multimodal cues, and psychological/affective states-as follows:

("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR transformer OR "large language model") AND (multimodal OR "multi-modal" OR nonverbal OR "non-verbal" OR "facial expression" OR "micro-expression" OR speech OR acoustic OR prosod OR gaze OR "eye tracking" OR physiological OR "body posture") AND (depress* OR anxiet* OR emotion* OR affect* OR "mental health" OR psycholog* OR suicid* OR distress)*

The string was applied to the TITLE-ABS-KEY field. Limiters restricted results to the 2020–2025 window and to English-language records. The string was piloted and iteratively refined; overly broad terms that returned large volumes of unrelated engineering records were tightened by anchoring the third block to psychological and affective outcomes.

Database and Information Sources

The primary bibliographic database searched was Scopus, selected for its broad multidisciplinary coverage of computer science, biomedical engineering, and psychology. The Scopus metadata export used for this review contained 669 records; this export served as the authoritative evidence base for identification, screening, and inclusion. The search was executed and exported for records published up to 2025.

Eligibility Criteria

Inclusion and exclusion criteria were defined a priori (Table 2).

Table 2. <Inclusion and exclusion criteria.>

Criterion	Inclusion	Exclusion
Topic	AI/ML analysis of nonverbal cues to detect a psychological or affective state	No psychological/affective target, or no AI analysis of nonverbal cues
Method	Multimodal or nonverbal-cue-based computational method	Unimodal non-behavioral or non-computational assessment only
Document type	Peer-reviewed article or review	Conference paper, book chapter, editorial, note, retracted item
Period	2020–2025	Before 2020
Language	English	Non-English
Availability	Full text retrievable	Full text not retrievable

Study Selection Process

Selection proceeded in stages consistent with PRISMA 2020. Duplicate records were removed first. Remaining records were screened on title and abstract against the eligibility criteria; records that were clearly out of scope were excluded. Potentially eligible records were assessed at full text, and studies meeting all criteria were retained for synthesis. Reasons for full-text exclusion were recorded and are reported in the flow diagram (Figure 1).

Quality Assessment

Methodological quality was appraised using the Mixed Methods Appraisal Tool (MMAT), version 2018 (Hong et al., 2018), suitable for the heterogeneous quantitative designs anticipated in this literature. Each empirical study was first classified by design category and then rated on five criteria (Yes = 1, Can't tell / partial = 0.5, No = 0). Consistent with MMAT guidance, total scores were used descriptively rather than as an exclusion threshold; no study was excluded on quality grounds alone. Secondary studies (reviews and meta-analyses) were appraised separately for review conduct and are flagged accordingly. Table 3 summarizes the appraisal.

Table 3. <MMAT quality-appraisal summary of included studies (Q1–Q5: five design-appropriate criteria; Y = Yes, CT = Can't tell/partial, N = No). Reviews were appraised separately.>

No	Study	Design	Q1	Q2	Q3	Q4	Q5	Score	Rating
1	Wang et al. (2025)	Review / meta-analysis	-	-	-	-	-	n/a	Appraised separately
2	Thakur et al. (2025)	Review / meta-analysis	-	-	-	-	-	n/a	Appraised separately
3	Singh & Malik (2025)	Review / meta-analysis	-	-	-	-	-	n/a	Appraised separately
4	Shuai et al. (2025)	Review / meta-analysis	-	-	-	-	-	n/a	Appraised separately
5	Sălăgean et al. (2025)	Quantitative descriptive	Y	Y	CT	Y	CT	4/5 (80%)	High
6	Flores et al. (2025)	Quantitative descriptive	Y	Y	CT	Y	CT	4/5 (80%)	High
7	Ye et al. (2025)	Quantitative descriptive	Y	Y	CT	Y	CT	4/5 (80%)	High
8	Zhang et al. (2025)	Quantitative descriptive	Y	Y	CT	Y	CT	4/5 (80%)	High
9	Jin et al. (2025)	Quantitative descriptive	Y	Y	CT	Y	Y	4.5/5 (90%)	High
10	Liu et al. (2025)	Quantitative non-randomized	Y	Y	Y	Y	Y	5/5 (100%)	High
11	Yang & Zheng (2025)	Quantitative descriptive	Y	Y	CT	Y	CT	4/5 (80%)	High
12	Zhu et al. (2025)	Quantitative descriptive	Y	Y	CT	Y	CT	4/5 (80%)	High
13	Eberhardt et al. (2025)	Quantitative descriptive	Y	Y	CT	Y	Y	4.5/5 (90%)	High
14	Hau et al. (2025)	Quantitative descriptive	Y	Y	CT	Y	Y	4.5/5 (90%)	High
15	Min et al. (2025)	Quantitative descriptive	Y	Y	CT	Y	Y	4.5/5 (90%)	High

No	Study	Design	Q1	Q2	Q3	Q4	Q5	Score	Rating
16	Zhu et al. (2025)	Quantitative non-randomized	Y	Y	Y	Y	Y	5/5 (100%)	High
17	Kuyucu et al. (2025)	Quantitative non-randomized	Y	Y	Y	Y	Y	5/5 (100%)	High
18	Guhan et al. (2025)	Quantitative descriptive	Y	Y	CT	Y	Y	4.5/5 (90%)	High
19	Chen et al. (2025)	Quantitative non-randomized	Y	Y	Y	Y	Y	5/5 (100%)	High
20	Jiang et al. (2024)	Quantitative descriptive	Y	Y	Y	Y	Y	5/5 (100%)	High

Data Extraction

For each included study the following were extracted: authors and year; source; study design; nonverbal modalities analyzed; target psychological/affective state; and principal findings. Extraction was performed against the source export, and bibliographic details (authors, titles, years, sources, DOIs) were used exactly as recorded to preserve data integrity.

Bibliometric and Descriptive Analysis

Descriptive analysis of the corpus characterized the annual distribution of records (Figure 2). For the included studies, modality coverage (Figure 3) and target-state distribution (Figure 4) were tabulated to support the thematic synthesis and to answer RQ1 and RQ2.

Data Synthesis

Because included studies were heterogeneous in design, datasets, and metrics, a narrative thematic synthesis was conducted rather than meta-analysis (Thomas & Harden, 2008). Studies were coded inductively by modality and target state, then grouped into themes aligned with the research questions; agreements, contradictions, and gaps across studies were examined.

Reporting and Documentation

The review followed the PRISMA 2020 checklist for reporting (Page et al., 2021). The flow of records from identification to inclusion is documented in Figure 1, and all synthesized claims are traceable to the included studies listed in Table 4.

Results and Discussions

Study Selection Results

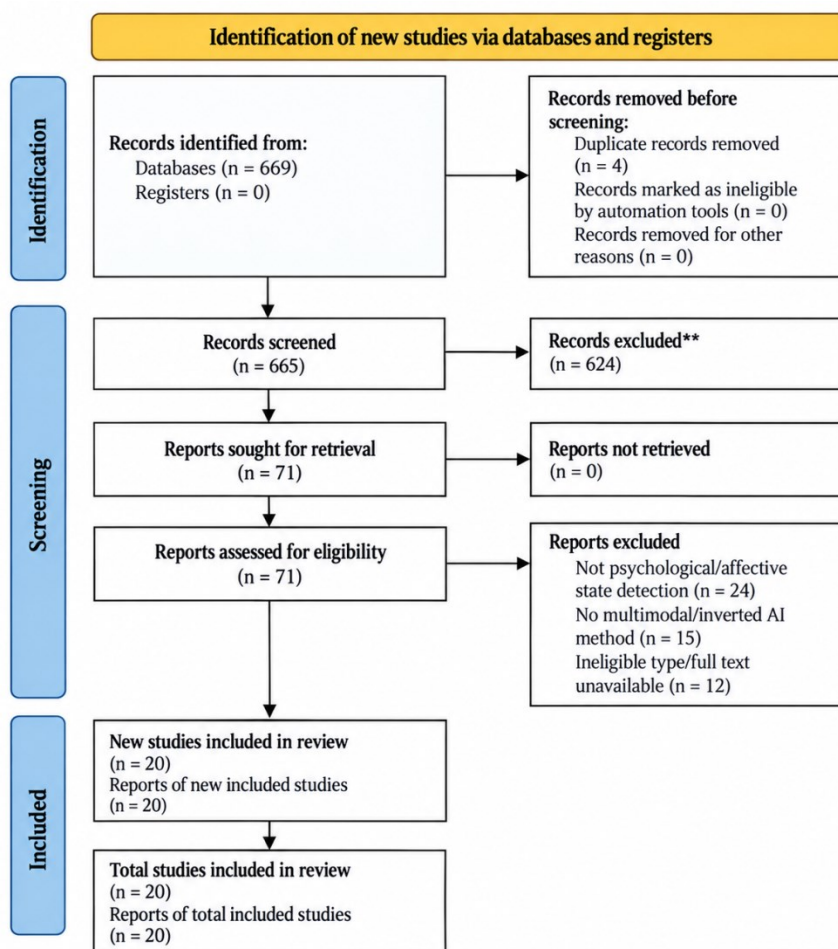


Figure 1. <PRISMA 2020 flow diagram of study identification, screening, eligibility, and inclusion (adapted from Page et al., 2021).>

The Scopus search identified 669 records. After removal of 4 duplicate records, 665 records were screened on title and abstract, of which 594 were excluded as clearly outside the review scope. The remaining 71 records were assessed at full text, and 51 were excluded: 24 did not target detection of a psychological or affective state, 15 did not employ a multimodal or nonverbal-cue-based AI method, and 12 were of an ineligible document type or had no retrievable full text. Twenty studies met all criteria and were included in the synthesis. These numbers are consistent across the abstract, this section, and Figure 1.

Descriptive Characteristics of Included Studies

Table 4 summarizes the 20 included studies with full titles and principal findings; Table 5 classifies each study by design, nonverbal modalities, target state, and source. The corpus is recent (19 of 20 studies published in 2025) and dominated by peer-reviewed journal articles, complemented by four synthesis papers.

Table 4. <Summary of included studies with full titles and key findings.>

No	Study (Authors, Year)	Full title	Key findings
1	Wang et al. (2025)	AI-assisted multi-modal information for the screening of depression: a systematic review and meta-analysis	Pooled analysis found AI-assisted multimodal measures (EEG, eye movement, audio-video, gait) outperformed unimodal approaches for depression screening, though heterogeneity and external validation remained limited.
2	Thakur et al. (2025)	A Comprehensive Review of Unimodal and Multimodal Emotion Detection: Datasets, Approaches, and Limitations	Deep learning extracts intrinsic emotional features more accurately across face and speech than handcrafted unimodal methods, but dataset variability and fusion complexity constrain generalization.

No	Study (Authors, Year)	Full title	Key findings
3	Singh & Malik (2025)	Unveiling hidden emotions: a review of microexpression recognition, classification, and datasets	Micro-expressions expose feelings people try to mask; enduring barriers are scarce high-quality annotated data, class imbalance, and weak real-world generalization.
4	Shuai et al. (2025)	Advances in Facial Micro-Expression Detection and Recognition: A Comprehensive Review	Automatic micro-expression detection and recognition are maturing with clear value for psychotherapy and diagnosis, yet the rapid, low-amplitude dynamics still challenge reliable recognition.
5	Sălăgean et al. (2025)	Seeing the Unseen: Real-Time Micro-Expression Recognition with Action Units and GPT-Based Reasoning	Coupling action-unit detection with GPT-based reasoning achieved real-time micro-expression recognition on CASME II, positioned for clinical assessment of emotional states.
6	Flores et al. (2025)	WavFace: A Multimodal Transformer-Based Model for Depression Screening	A transformer layer over pretrained audio and temporal-facial features improved depression screening from virtual-agent clinical interviews under small-sample constraints.
7	Ye et al. (2025)	DEP-Former: Multimodal Depression Recognition Based on Facial Expressions and Audio Features via Emotional Changes	An emotion-word-reading paradigm with a modality-adaptor transformer (DEP-Former) improved depression recognition by resolving cross-modal inconsistencies between face and voice.
8	Zhang et al. (2025)	A Multimodal Artificial Intelligence Model for Depression Severity Detection Based on Audio and Video Signals	A lightweight audio-video model estimated depression severity accurately while remaining deployable for real-time monitoring in resource-constrained settings.
9	Jin et al. (2025)	Diagnosis of depression based on facial multimodal data	Fusing spatiotemporal facial features with GCN-LSTM audio analysis yielded objective automatic depression diagnosis, reducing reliance on subjective rating scales.
10	Liu et al. (2025)	Multimodal depression recognition and analysis: Facial expression and body posture changes via emotional stimuli	An end-to-end model fusing facial expression and body posture under emotional stimuli enabled rapid preliminary depression screening versus healthy controls.
11	Yang & Zheng (2025)	Multi-level spatiotemporal graph attention fusion for multimodal depression detection	Graph-attention fusion modeling intra- and cross-modal relations outperformed concatenation-based fusion for depression detection under small-sample conditions.
12	Zhu et al. (2025)	RMER-DT: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers	A diffusion-transformer model recognized conversational emotion robustly even under random missing-modality data, relevant to mental-health interaction settings.
13	Eberhardt et al. (2025)	Decoding emotions: Exploring the validity of sentiment analysis in psychotherapy	Transformer-based sentiment analysis of psychotherapy transcripts showed partial construct and criterion validity, supporting NLP as an augmenting rather than standalone emotion measure.
14	Hau et al. (2025)	Let's talk about non-verbal communication: using AI and Machine learning for the investigation of interpersonal psychotherapeutic interactions	AI/ML can link multimodal nonverbal signals (posture, facial, gaze, prosody) to verbal content to map interpersonal synchronization in psychotherapy.
15	Min et al. (2025)	Automated Speech Analysis for Screening and Monitoring Bipolar Depression: Machine Learning Model Development and Interpretation Study	Between- and within-person speech classifiers screened bipolar depression and tracked longitudinal change to detect depressive recurrence.
16	Zhu et al. (2025)	Speech feature identification model for depressed individuals with suicidal ideation based on autobiographical memory	A model integrating vocal features and autobiographical memory distinguished depressed patients with suicidal ideation from those without, isolating markers of suicidality.

No	Study (Authors, Year)	Full title	Key findings
17	Kuyucu et al. (2025)	Emotion recognition in Virtual Reality using sensor fusion with eye tracking	Multi-sensor fusion of EEG and eye-tracking in immersive VR improved emotion recognition under controlled conditions, overcoming mobile-sensor limitations.
18	Guhan et al. (2025)	Developing a Machine Learning-Based Automated Patient Engagement Estimator for Telehealth: Algorithm Development and Validation Study	An ML estimator inferred patient engagement from limited telehealth nonverbal cues, supporting therapists in building the therapeutic relationship remotely.
19	Chen et al. (2025)	Comparative Efficacy of MultiModal AI Methods in Screening for Major Depressive Disorder: Machine Learning Model Development Predictive Pilot Study	AI models on facial, acoustic, and text features screened major depressive disorder, with multimodal audiovisual signals complementing subjective scales and interviews.
20	Jiang et al. (2024)	Multimodal Mental Health Digital Biomarker Analysis From Remote Interviews Using Facial, Vocal, Linguistic, and Cardiovascular Patterns	Temporal facial, vocal, linguistic, and cardiovascular biomarkers from tele-video interviews differentiated individuals with psychiatric disorders across multiple classification tasks.

Table 5. <Classification of included studies by design, nonverbal modalities, target state, and source.>

No	Study	Study design	Nonverbal modalities	Target state	Source
1	Wang et al. (2025)	Systematic review / meta-analysis	Facial/video, Speech/vocal, Physiological	Depression	npj Digital Medicine
2	Thakur et al. (2025)	Narrative review	Facial/video, Speech/vocal	Emotion	Expert Systems
3	Singh & Malik (2025)	Review	Facial micro-expression	Concealed emotion	Multimedia Tools and Applications
4	Shuai et al. (2025)	Review	Facial micro-expression	Concealed emotion	Information (Switzerland)
5	Sălăgean et al. (2025)	Quantitative (algorithm development)	Facial micro-expression	Concealed emotion	Applied Sciences (Switzerland)
6	Flores et al. (2025)	Quantitative (algorithm development)	Speech/vocal, Facial/video	Depression	IEEE Journal of Biomedical and Health Informatics
7	Ye et al. (2025)	Quantitative (algorithm development)	Facial/video, Speech/vocal	Depression	IEEE Transactions on Circuits and Systems for Video Technology
8	Zhang et al. (2025)	Quantitative (algorithm development)	Speech/vocal, Facial/video	Depression	Electronics (Switzerland)
9	Jin et al. (2025)	Quantitative (diagnostic)	Facial/video, Speech/vocal	Depression	Frontiers in Psychiatry
10	Liu et al. (2025)	Quantitative (case-control, n=146)	Facial/video, Body/posture	Depression	Journal of Affective Disorders
11	Yang & Zheng (2025)	Quantitative (algorithm development)	Facial/video, Speech/vocal, Text/linguistic	Depression	Biomedical Signal Processing and Control
12	Zhu et al. (2025)	Quantitative (algorithm development)	Speech/vocal, Facial/video, Text/linguistic	Emotion	Information Fusion
13	Eberhardt et al. (2025)	Quantitative validity study (85 transcripts)	Text/linguistic	Emotion	Psychotherapy Research
14	Hau et al. (2025)	Methodological / preliminary study	Facial/video, Speech/vocal, Body/posture	Emotion	Scandinavian Psychoanalytic Review

No	Study	Study design	Nonverbal modalities	Target state	Source
15	Min et al. (2025)	Quantitative (model development)	Speech/vocal	Depression	JMIR Medical Informatics
16	Zhu et al. (2025)	Quantitative (clinical, n=88)	Speech/vocal	Suicidal ideation	BMC Psychiatry
17	Kuyucu et al. (2025)	Quantitative (experimental, n=95)	Physiological, Eye-tracking	Emotion	Computers in Biology and Medicine
18	Guhan et al. (2025)	Quantitative (development & validation)	Facial/video, Body/posture	Patient engagement	JMIR Formative Research
19	Chen et al. (2025)	Quantitative (case-control, n=89)	Facial/video, Speech/vocal, Text/linguistic	Depression	JMIR Formative Research
20	Jiang et al. (2024)	Quantitative (observational, n=73)	Facial/video, Speech/vocal, Text/linguistic, Physiological	Psychiatric disorders (broad)	IEEE Journal of Biomedical and Health Informatics

Corpus Trends and Modality–Target Distribution

Annual output rose steadily across the corpus and peaked sharply in 2025 (Figure 2), reflecting intensifying interest in AI-based affective sensing. Across the 20 included studies, facial/video and speech/vocal channels were the most frequently analyzed, followed by text/linguistic, physiological, and body-movement channels, with most studies combining two or more channels in a fusion architecture (Figure 3). Target states were dominated by depression, followed by general emotion recognition and concealed affect, with individual studies addressing suicidal ideation, patient engagement, and broad psychiatric classification (Figure 4).

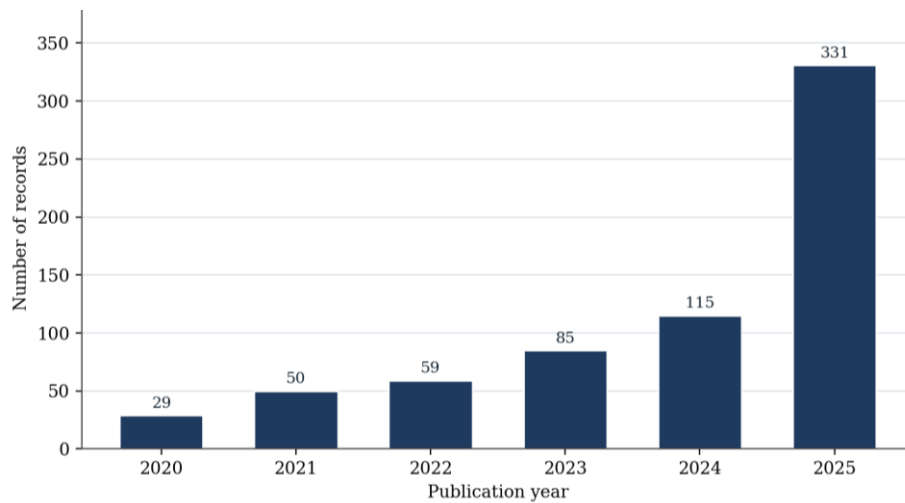


Figure 2. <Annual distribution of the Scopus corpus (669 records, 2020–2025).>

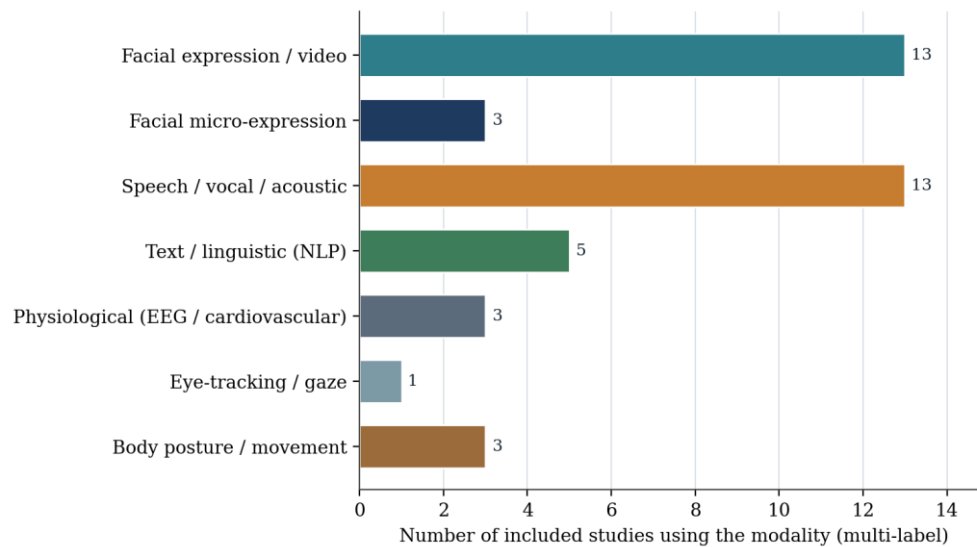


Figure 3. <Nonverbal modality channels analyzed across the 20 included studies (multi-label; a study may use several channels).>

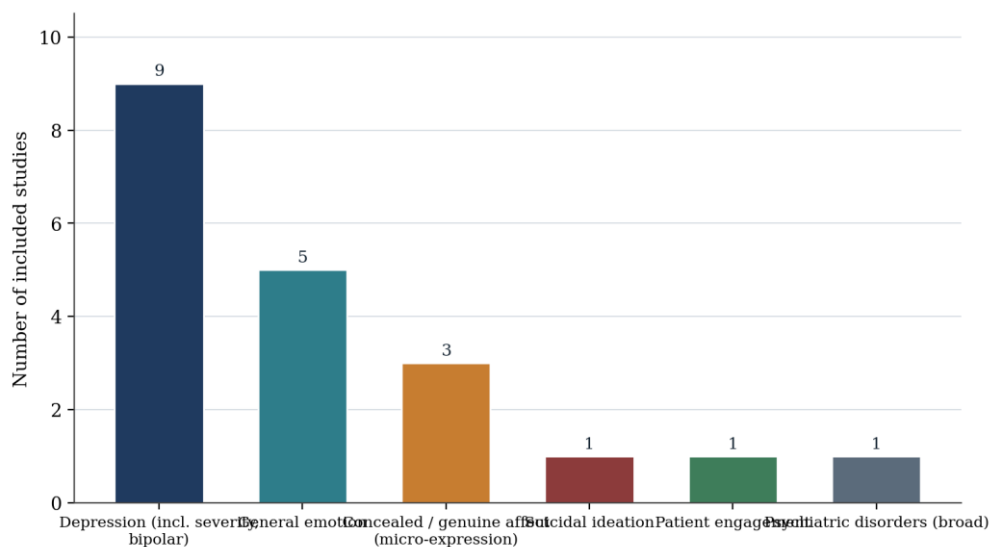


Figure 4. <Distribution of included studies by primary target psychological/affective state.>

Thematic Synthesis

RQ1 - Nonverbal modalities and fusion strategies

Included studies span the full range of nonverbal channels, and multimodal fusion was the prevailing design. Facial signals were central across studies, from micro-expression analysis that targets concealed affect (Singh & Malik, 2025; Shuai et al., 2025) to real-time micro-expression recognition combining action units with GPT-based reasoning (Sălăgean et al., 2025). Facial cues were most often fused with audio: WavFace layered a transformer over pretrained audio and temporal-facial features (Flores et al., 2025), DEP-Former aligned facial and vocal features through a modality adapter (Ye et al., 2025), and further audio–video models targeted depression and its severity (Zhang et al., 2025; Jin et al., 2025). Fusion strategy itself was a locus of innovation: graph-attention fusion modeled intra- and cross-modal relations beyond simple concatenation (Yang & Zheng, 2025), and a diffusion-transformer model preserved robustness under missing-modality conditions (X. Zhu et al., 2025).

Beyond face and voice, studies incorporated body posture under emotional stimuli (Liu et al., 2025), linguistic content via NLP sentiment analysis of psychotherapy transcripts (Eberhardt et al., 2025), and physiological sensing, including EEG with eye-tracking fusion in virtual reality (Kuyucu et al., 2025) and cardiovascular patterns from remote interviews (Jiang et al., 2024). Some work explicitly linked multimodal

nonverbal signals to verbal content to study interactional synchrony in psychotherapy (Hau et al., 2025), while others emphasized deployable or lightweight architectures for real-world monitoring (Zhang et al., 2025). In sum, the field has converged on multimodal fusion, with the facial–vocal pairing as the backbone and physiological and linguistic channels as increasingly common complements.

RQ2 - Target states and detection performance

Depression was by far the most frequent target, addressed through audiovisual, speech-only, and fusion pipelines (Flores et al., 2025; Ye et al., 2025; Zhang et al., 2025; Jin et al., 2025; Yang & Zheng, 2025), including bipolar depression tracked longitudinally from speech (Min et al., 2025) and comparative multimodal screening for major depressive disorder (Chen et al., 2025). A recurring, well-supported claim is that multimodal fusion outperforms unimodal baselines; the pooled evidence likewise favored AI-assisted multimodal measures for depression screening while cautioning about heterogeneity and limited external validation (Wang et al., 2025). General emotion recognition formed a second cluster, spanning conversational contexts (X. Zhu et al., 2025), immersive VR (Kuyucu et al., 2025), and the validity of NLP-based sentiment in therapy, which showed only partial construct and criterion validity (Eberhardt et al., 2025).

A third cluster targeted concealed or clinically high-stakes states. Micro-expression research explicitly frames its purpose as surfacing feelings people attempt to mask (Singh & Malik, 2025; Shuai et al., 2025; Sălăgean et al., 2025), directly addressing the review’s motivating problem of distress that clients do not verbalize. Individual studies extended detection to suicidal ideation, distinguishing depressed patients with and without suicidal ideation by integrating vocal features with autobiographical memory (Y. Zhu et al., 2025); to telehealth patient engagement inferred from limited nonverbal cues (Guhan et al., 2025); and to broad psychiatric classification from temporal multimodal biomarkers in remote interviews (Jiang et al., 2024). Reported performance was generally promising within study samples, but metrics were heterogeneous and rarely validated on independent clinical cohorts, so cross-study performance claims should be read as indicative rather than definitive.

RQ3 - Methodological and translational limitations

Three limitations recur across the corpus. First, datasets are small and often clinically narrow, which the studies themselves address through architectural regularization, graph-based fusion, and privacy-preserving or synthetic-data strategies rather than through larger, representative sampling (Yang & Zheng, 2025; Simć et al., 2024; Ibrahim et al., 2025). Second, external validation is scarce: most studies report within-sample performance, and even the pooled meta-analytic evidence flags heterogeneity and limited generalization (Wang et al., 2025). Third, validity and translational framing are underdeveloped—NLP sentiment showed only partial validity against clinical criteria (Eberhardt et al., 2025), and few studies evaluate deployment, longitudinal monitoring, or ethical safeguards, with notable exceptions in relapse-oriented monitoring systems (Zlatintsi et al., 2022). These gaps, rather than feasibility, now constitute the principal barrier to clinical use.

Comparative and Critical Analysis

Methodologically, the corpus is dominated by quantitative algorithm-development and descriptive designs, with case-control and experimental studies providing the strongest evidence (Table 5). This distribution mirrors the field’s engineering origins and helps explain both its rapid performance gains and its validation gap. Comparisons across studies are complicated by inconsistent datasets and metrics: depression pipelines variously report accuracy, AUC, or severity-estimation error, while emotion-recognition studies use categorical or dimensional targets. The clearest convergent signal is architectural-multimodal fusion, and increasingly attention- or graph-based fusion, reliably improves on unimodal and concatenation baselines (Ye et al., 2025; Yang & Zheng, 2025; Wang et al., 2025). Underused designs include prospective, ecologically valid studies in counseling settings and head-to-head comparisons against clinician assessment, both of which would strengthen translational claims.

Interpretation. The evidence supports the review’s premise that nonverbal channels carry assessment-relevant psychological information that AI can extract, and that fusing channels captures complementary signal that single modalities miss. The prominence of micro-expression work is theoretically important because it operationalizes the long-standing clinical intuition that suppressed affect leaks through involuntary cues (Singh & Malik, 2025; Shuai et al., 2025), aligning computational detection with counseling constructs of concealment and non-disclosure. Theoretical and practical implications. Practically, multimodal AI is best positioned as an augmenting layer over—not a replacement for—clinical judgment: as a screening aid, a monitoring tool between sessions, or a means of surfacing engagement and risk cues that are hard to read remotely (Guhan et al., 2025; Min et al., 2025). For counseling specifically, systems that link nonverbal signals to verbal content could support supervision and process research (Hau et al., 2025; Eberhardt et al., 2025). Theoretically, the modality-by-target

map assembled here suggests that affect and depression are comparatively well served, whereas concealed, latent, and high-risk states remain sparsely and unevenly addressed.

Comparison with prior reviews and contradictions. Consistent with the pooled findings of Wang et al. (2025) and prior emotion-detection syntheses (Thakur et al., 2025), multimodal approaches outperform unimodal ones; the present synthesis extends those conclusions across a wider set of target states and counseling-relevant applications. A genuine tension in the literature concerns validity: enthusiastic performance reports coexist with sobering validity evidence when models are tested against clinical criteria (Eberhardt et al., 2025), implying that high in-sample accuracy does not guarantee clinical meaning. Research gaps. At least three gaps stand out: (1) the scarcity of prospective, externally validated studies in real counseling or telehealth settings; (2) uneven coverage of concealed, latent, and high-risk states relative to the concentration on depression and general emotion; and (3) immature reporting of ethics, privacy, fairness, and the communicative framing of affective outputs (Simć et al., 2024; Ibrahim et al., 2025; Resende Faria et al., 2024).

Limitations of this review. First, identification relied on a single database export (Scopus); relevant records indexed only elsewhere may have been missed, and the duplicate count reflects that single-source export. Second, restriction to English-language, peer-reviewed articles and reviews may introduce language and publication bias. Third, the heterogeneity of metrics precluded meta-analysis, so synthesis is narrative and cross-study performance is characterized qualitatively. Future research agenda. Priorities include: prospective, multi-site validation against clinician assessment in counseling and telehealth; standardized datasets and reporting for multimodal affective models to enable comparison; deliberate targeting of concealed, latent, and high-risk states with appropriate safeguards; and privacy-preserving, transparent deployment with explicit attention to fairness and to how affective inferences are communicated to clinicians and clients. In summary, for RQ1, the field has converged on multimodal fusion built around facial and vocal cues; for RQ2, depression and general emotion are well covered while concealed and high-risk states are not; and for RQ3, small data, weak external validation, and immature ethics are the decisive barriers to clinical translation.

Conclusions

This review synthesized 20 recent studies on multimodal AI for detecting psychological and affective states from nonverbal cues. In answer to the research questions, the field has converged on multimodal fusion-anchored by facial and vocal channels and increasingly enriched by physiological and linguistic signals-and consistently finds that fusion outperforms unimodal detection; target states are dominated by depression and general emotion, with concealed affect, suicidal ideation, and engagement addressed more sparingly; and the principal constraints are small, clinically narrow datasets, limited external validation, and underdeveloped ethical safeguards. The review's core contribution is an integrative modality-by-target map that clarifies where the evidence is strong and where it is thin, together with a translational agenda. For practice, these systems are best deployed as objective aids that augment, rather than replace, professional judgment. Realizing their promise for counseling and mental-health assessment will require prospective clinical validation, standardized reporting, and privacy-preserving, transparent deployment.

References

- Bhatt, P., Sethi, A., Tasgaonkar, V., Shroff, J., Pendharkar, I., Desai, A., Sinha, P., Deshpande, A., Joshi, G., Rahate, A., Jain, P., Walambe, R., Kotecha, K., & Jain, N. K. (2023). Machine learning for cognitive behavioral analysis: datasets, methods, paradigms, and research directions. *Brain Informatics*. <https://doi.org/10.1186/s40708-023-00196-6>
- Cascella, M., Leoni, M. L. G., Shariff, M. N., & Varrassi, G. (2024). Artificial Intelligence-Driven Diagnostic Processes and Comprehensive Multimodal Models in Pain Medicine. *Journal of Personalized Medicine*. <https://doi.org/10.3390/jpm14090983>
- Chen, D., Wang, P., Zhang, X., Qiao, R., Li, N., Zhang, X., Zhang, H., & Wang, G. (2025). Comparative Efficacy of MultiModal AI Methods in Screening for Major Depressive Disorder: Machine Learning Model Development Predictive Pilot Study. *JMIR Formative Research*. <https://doi.org/10.2196/56057>
- Du, M., Liu, S., Wang, T., Zhang, W., Ke, Y., Chen, L., & Ming, D. (2023). Depression recognition using a proposed speech chain model fusing speech production and perception features. *Journal of Affective Disorders*. <https://doi.org/10.1016/j.jad.2022.11.060>

- Eberhardt, S. T., Schaffrath, J., Moggia, D., Schwartz, B., Jaehde, M., Rubel, J. A., Baur, T., André, E., & Lutz, W. (2025). Decoding emotions: Exploring the validity of sentiment analysis in psychotherapy. *Psychotherapy Research*. <https://doi.org/10.1080/10503307.2024.2322522>
- Flores, R., Tlachac, M. L., Shrestha, A., & Rundensteiner, E. A. (2025). WavFace: A Multimodal Transformer-Based Model for Depression Screening. *IEEE Journal of Biomedical and Health Informatics*. <https://doi.org/10.1109/JBHI.2025.3529348>
- Guhan, P., Awasthi, N., McDonald, K., Bussell, K., Reeves, G., Manocha, D., & Bera, A. (2025). Developing a Machine Learning-Based Automated Patient Engagement Estimator for Telehealth: Algorithm Development and Validation Study. *JMIR Formative Research*. <https://doi.org/10.2196/46390>
- Halfon, S., Doyran, M., Türkmen, B., Oktay, E. A., & Salah, A. A. (2021). Multimodal affect analysis of psychodynamic play therapy. *Psychotherapy Research*. <https://doi.org/10.1080/10503307.2020.1839141>
- Hau, S., Rugolon, F., Samuels, T. J., & Högman, L. (2025). Let's talk about non-verbal communication: using AI and Machine learning for the investigation of interpersonal psychotherapeutic interactions. *Scandinavian Psychoanalytic Review*. <https://doi.org/10.1080/01062301.2025.2539549>
- Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., ... Vedel, I. (2018). Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada.
- Ibrahim, M., Khalil, Y. A., Amirrajab, S., Sun, C., Breeuwer, M., Pluim, J., Elen, B., Ertaylan, G., & Dumontier, M. (2025). Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in Biology and Medicine*. <https://doi.org/10.1016/j.compbiomed.2025.109834>
- Jiang, Z., Seyedi, S., Griner, E., Abbasi, A., Rad, A. B., Kwon, H., Cotes, R. O., & Clifford, G. D. (2024). Multimodal Mental Health Digital Biomarker Analysis From Remote Interviews Using Facial, Vocal, Linguistic, and Cardiovascular Patterns. *IEEE Journal of Biomedical and Health Informatics*. <https://doi.org/10.1109/JBHI.2024.3352075>
- Jin, N., Ye, R., & Li, P. (2025). Diagnosis of depression based on facial multimodal data. *Frontiers in Psychiatry*. <https://doi.org/10.3389/fpsy.2025.1508772>
- Kuyucu, M., Sarikaya, M. A., Karakaş, T., Özkan, D. Y., Demir, Y., Bilen Ö., & Ince, G. (2025). Emotion recognition in Virtual Reality using sensor fusion with eye tracking. *Computers in Biology and Medicine*. <https://doi.org/10.1016/j.compbiomed.2025.111070>
- Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., ... Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine*, 6(7), e1000100. <https://doi.org/10.1371/journal.pmed.1000100>
- Liu, Y., Li, X., Wang, M., Bi, J., Lin, S., Wang, Q., Yu, Y., Ye, J., & Zheng, Y. (2025). Multimodal depression recognition and analysis: Facial expression and body posture changes via emotional stimuli. *Journal of Affective Disorders*. <https://doi.org/10.1016/j.jad.2025.03.155>
- Liu, Z., & Ye, C. C. (2025). Feature analysis of depression patients' house-tree-person drawings using convolutional neural networks. *Neuroscience Informatics*. <https://doi.org/10.1016/j.neuri.2025.100239>
- Methley, A. M., Campbell, S., Chew-Graham, C., McNally, R., & Cheraghi-Sohi, S. (2014). PICO, PICOS and SPIDER: A comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC Health Services Research*, 14, 579. <https://doi.org/10.1186/s12913-014-0579-0>
- Min, S., Yeum, T. -S., Shin, D., Rhee, S. J., Lee, H., Lee, H. -S., Park, S., Lee, J., & Ahn, Y. M. (2025). Automated Speech Analysis for Screening and Monitoring Bipolar Depression: Machine Learning Model Development and Interpretation Study. *JMIR Medical Informatics*. <https://doi.org/10.2196/79093>

- Mutawa, A. M., & Hassouneh, A. (2024). Multimodal Real-Time patient emotion recognition system using facial expressions and brain EEG signals based on Machine learning and Log-Sync methods. *Biomedical Signal Processing and Control*. <https://doi.org/10.1016/j.bspc.2023.105942>
- Muzammel, M., Salam, H., & Othmani, A. (2021). End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. *Computer Methods and Programs in Biomedicine*. <https://doi.org/10.1016/j.cmpb.2021.106433>
- Nemati, R., Shirini, K., & Gharehveran, S. S. (2025). FER-HA: a hybrid attention model for facial emotion recognition. *Journal of Supercomputing*. <https://doi.org/10.1007/s11227-025-07983-4>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372, n160. <https://doi.org/10.1136/bmj.n160>
- Park, J., & Moon, N. (2022). Design and Implementation of Attention Depression Detection Model Based on Multimodal Analysis. *Sustainability (Switzerland)*. <https://doi.org/10.3390/su14063569>
- Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2022). MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*. <https://doi.org/10.1016/j.bspc.2021.103107>
- Resende Faria, D., Weinberg, A. I., & Ayrosa, P. P. (2024). Multimodal Affective Communication Analysis: Fusing Speech Emotion and Text Sentiment Using Machine Learning. *Applied Sciences (Switzerland)*. <https://doi.org/10.3390/app14156631>
- Shuai, T., Beng, S., Khalid, F. B., & Rahmat, R. W. B. O. K. (2025). Advances in Facial Micro-Expression Detection and Recognition: A Comprehensive Review. *Information (Switzerland)*. <https://doi.org/10.3390/info16100876>
- Simić, N., Suzić, S., Milošević, N., Stanojević, V., Nosek, T., Popović, B., & Bajović, D. (2024). Enhancing Emotion Recognition through Federated Learning: A Multimodal Approach with Convolutional Neural Networks. *Applied Sciences (Switzerland)*. <https://doi.org/10.3390/app14041325>
- Singh, J., & Malik, P. (2025). Unveiling hidden emotions: a review of microexpression recognition, classification, and datasets. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-025-21100-w>
- Sălăgean, G. L., Leba, M., & Ionica, A. C. (2025). Seeing the Unseen: Real-Time Micro-Expression Recognition with Action Units and GPT-Based Reasoning. *Applied Sciences (Switzerland)*. <https://doi.org/10.3390/app15126417>
- Thakur, P., Kaur, N., Aggarwal, N., & Singh, S. (2025). A Comprehensive Review of Unimodal and Multimodal Emotion Detection: Datasets, Approaches, and Limitations. *Expert Systems*. <https://doi.org/10.1111/exsy.70103>
- Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8, 45. <https://doi.org/10.1186/1471-2288-8-45>
- Tomar, P. S., Mathur, K., & Suman, U. (2024). Fusing facial and speech cues for enhanced multimodal emotion recognition. *International Journal of Information Technology (Singapore)*. <https://doi.org/10.1007/s41870-023-01697-7>
- 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>
- Wang, L., Wang, C., Li, C., Murai, T., Bai, Y., Song, Z., Zhang, S., Zhang, Q., Huang, Y., Bi, X., & Jiang, J. (2025). AI-assisted multi-modal information for the screening of depression: a systematic review and meta-analysis. *npj Digital Medicine*. <https://doi.org/10.1038/s41746-025-01933-3>

- Xu, G., Jiang, H., Hou, H., & Liu, G. (2025). Self-Monitoring Deep Network Based on Multimodal Attention Fusion for Mental Emotion Recognition. *Journal of Cases on Information Technology*. <https://doi.org/10.4018/JCIT.393784>
- Yang, Y., & Zheng, W. (2025). Multi-level spatiotemporal graph attention fusion for multimodal depression detection. *Biomedical Signal Processing and Control*. <https://doi.org/10.1016/j.bspc.2025.108123>
- Ye, J., Yu, Y., Lu, L., Wang, H., Zheng, Y., Liu, Y., & Wang, Q. (2025). DEP-Former: Multimodal Depression Recognition Based on Facial Expressions and Audio Features via Emotional Changes. *IEEE Transactions on Circuits and Systems for Video Technology*. <https://doi.org/10.1109/TCSVT.2024.3491098>
- Zhang, L., Zhang, S., Zhang, X., & Zhao, Y. (2025). A Multimodal Artificial Intelligence Model for Depression Severity Detection Based on Audio and Video Signals. *Electronics (Switzerland)*. <https://doi.org/10.3390/electronics14071464>
- Zhao, W., Zhang, Q., Gao, L., Fan, N., Yun, Y., Song, J., Ji, Y., Wang, Y., Zhang, M., Yang, F., & Tan, S. (2025). A network analysis of facial and vocal emotion recognition deficits in schizophrenia. *Frontiers in Psychiatry*. <https://doi.org/10.3389/fpsy.2025.1598026>
- Zhu, L., Spachos, P., Ng, P. C., Yu, Y., Wang, Y., Plataniotis, K., & Hatzinakos, D. (2023). Stress Detection Through Wrist-Based Electrodermal Activity Monitoring and Machine Learning. *IEEE Journal of Biomedical and Health Informatics*. <https://doi.org/10.1109/JBHI.2023.3239305>
- Zhu, X., Wang, Y., Cambria, E., Rida, I., López, J. S., Cui, L., & Wang, R. (2025). RMER-DT: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers. *Information Fusion*. <https://doi.org/10.1016/j.inffus.2025.103268>
- Zhu, Y., Yin, Q., Xu, H., Xiao, F., Jiang, Q., Liang, M., Cheng, Q., & Liu, T. (2025). Speech feature identification model for depressed individuals with suicidal ideation based on autobiographical memory. *BMC Psychiatry*. <https://doi.org/10.1186/s12888-025-07635-0>
- Zlatintsi, A., Filntisis, P. P., Garoufis, C., Efthymiou, N., Maragos, P., Menychtas, A., Maglogiannis, I., Tsanakas, P., Sounapoglou, T., Kalisperakis, E., Karantinos, T., Lazaridi, M., Garyfalli, V., Mantas, A., Mantonakis, L., & Smyrnis, N. (2022). E-Prevention: Advanced Support System for Monitoring and Relapse Prevention in Patients with Psychotic Disorders Analyzing Long-Term Multimodal Data from Wearables and Video Captures. *Sensors*. <https://doi.org/10.3390/s22197544>