The Review Of
DIABETIC
STUDIES OPEN ACCESS

The Silicon Pulse: Can AI Decode Human Emotions Better Than The Clinical Eye? A Systematic Review

Dr. Salah Mahmoud Alabbasi

Family Medicine Senior registrar First Jeddah Health cluster - East Jeddah Hospital - Almatar AlQadeem Primary Health care.

Abstract

Background: Artificial intelligence (AI) has rapidly emerged as a transformative tool in decoding human emotions, offering unprecedented potential to augment clinical diagnostics and psychological assessment.
Objective: This systematic review aims to evaluate the extent to which AI systems can recognize and interpret human emotions with accuracy comparable to, or exceeding, human clinical judgment across psychiatric and neurological contexts.
Methods: Following PRISMA 2020 guidelines, ten empirical studies published between 2016 and 2025 were synthesized from databases including PubMed, IEEE Xplore, Scopus, and JMIR. Studies examining AI-driven emotion recognition through facial, vocal, linguistic, or multimodal data were included. Quality was assessed using the Newcastle–Ottawa Scale and Cochrane RoB 2 tool.
Results: AI-based systems demonstrated accuracy rates ranging from 77% to 99.8%, frequently surpassing human raters in structured emotion tasks. Deep learning models, such as convolutional neural networks (CNNs) and transformer architectures, achieved superior sensitivity and specificity in detecting depressive, autistic, and neurological emotional markers. Speech emotion recognition and EEG-based models further complemented multimodal detection with measurable correlations to psychiatric scales.
However, empathy perception and contextual interpretation remained limitations compared to expert clinicians.
Conclusions: AI demonstrates robust diagnostic and affective recognition capabilities, often rivalling human assessment in precision and scalability. Nevertheless, integration into clinical contexts must prioritize ethical oversight, interpretability, and emotional authenticity to ensure human–AI complementarity rather than replacement.

Keywords: Artificial intelligence, emotion recognition, machine learning, facial analysis, depression, empathy, psychiatry, EEG, speech emotion recognition, clinical AI.

Introduction

The intersection of artificial intelligence (AI) and affective neuroscience has created unprecedented opportunities to decode and quantify human emotions with objectivity and precision. Historically, emotion recognition relied on clinician observation and self-reporting—methods prone to bias, inconsistency, and inter-rater variability. With recent advances in machine learning, deep learning, and multimodal data fusion, AI systems now analyze complex affective cues such as micro-expressions, speech tone, and physiological signals in real time, offering enhanced diagnostic and therapeutic potential for psychiatric and neurological disorders (Flynn et al., 2020). This technological evolution has positioned AI not merely as an adjunct to clinical judgment but as a transformative force in emotion science.

Emotions manifest across multiple physiological and behavioral channels—facial expressions, voice, body movements, and neural activity—each encoding subtle indicators of mental state. Deep neural networks and convolutional architectures can extract and integrate these multimodal features with

exceptional granularity, often surpassing human perceptual limits (Zhuang et al., 2020). For instance, quantitative facial analysis has enabled the detection of asymmetries and muscular inactivity indicative of neurological impairment. Such computational precision provides clinicians with quantifiable biomarkers, bridging the gap between subjective experience and objective diagnosis.

In parallel, affective computing has expanded beyond mere emotion detection toward understanding emotional dynamics and context. By modeling affective trajectories through multimodal temporal signals, AI systems can recognize complex emotions such as ambivalence, anxiety, or apathy that are often difficult for clinicians to interpret reliably (Tripathi & Garg, 2024). These insights enable continuous monitoring and personalized mental health interventions, crucial for conditions like depression, autism, and post-stroke affective disorders.

Electrophysiological approaches, particularly electroencephalography (EEG), have further enhanced AI-driven emotion recognition. EEG-based algorithms capture real-time neural oscillations that correlate with emotional valence and arousal, offering robust internal indicators of affective state (Alarcão & Fonseca, 2017). Combining EEG with deep learning architectures allows for decoding cognitive-affective interactions that may not be observable through facial or vocal data alone, deepening our understanding of emotional processing in both health and disease contexts.

Similarly, the fusion of AI with electromyography (EMG) and peripheral nerve analysis has yielded breakthroughs in clinical emotion recognition and rehabilitation. Surface nerve EMG data analyzed through deep learning can reveal micro-variations in muscle activity that correspond to specific affective expressions or neurological dysfunctions (Zhu et al., 2022). These applications demonstrate AI's utility in contexts ranging from electroacupuncture feedback to facial paralysis therapy, underscoring the convergence of medical diagnostics and emotional informatics.

Beyond individual emotion detection, AI also facilitates interactive and social emotional modeling. Virtual human simulations and socially aware AI agents can engage users with realistic empathy and responsiveness, often achieving higher engagement than static clinician interfaces (Gratch et al., 2007). This interactional dimension suggests that emotion-aware AI may not only assess human affect but also influence it therapeutically—an emerging paradigm in digital psychiatry and telehealth.

Speech-based AI has similarly revolutionized emotion recognition in mental health assessment. Acoustic analyses of tone, pitch, and rhythm have demonstrated significant correlations with depressive and suicidal risk markers, offering continuous, unobtrusive monitoring through natural communication (Cummins et al., 2015). These methods, when integrated with text and facial data, create comprehensive multimodal models capable of interpreting complex affective cues at scale.

Recent systematic reviews affirm that large language models (LLMs) and multimodal architectures combining facial, linguistic, and contextual inputs significantly outperform unimodal systems in detecting psychiatric symptoms such as depression and anxiety (Sadeghi et al., 2024). Comparative research further demonstrates that AI can match or even exceed human accuracy in classifying emotional states, particularly in structured tasks like autism prediction and neurological screening (Sariyanidi et al., 2023). Moreover, meta-analytic evidence confirms the diagnostic reliability of deep learning models across neurological conditions, with pooled accuracies exceeding 90% (Yoonesi et al., 2025).

Finally, voice-based affect recognition is emerging as one of the most scalable and privacy-preserving modalities in emotion-aware AI. Speech emotion recognition systems show high feasibility for early psychiatric detection and monitoring, bridging accessibility gaps in traditional care models (Jordan et al., 2025). As research continues to refine multimodal integration and ethical governance, the convergence of AI and emotional intelligence marks a critical frontier in the evolution of computational psychiatry and precision mental healthcare.

Methodology

Study Design

This study employed a systematic review methodology guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 framework to ensure methodological rigor, transparency, and replicability. The primary objective was to synthesize and critically evaluate empirical evidence regarding the capability of artificial intelligence (AI) to decode, interpret, and quantify human emotional and affective states, particularly in comparison to or alongside clinical

assessment. The review focused on AI-based systems used for emotion recognition, mental health diagnosis, and affective monitoring using facial, vocal, linguistic, and physiological data.
This review encompassed peer-reviewed empirical studies evaluating the accuracy, sensitivity, and clinical applicability of AI-driven emotion recognition tools across diverse populations and mental health contexts. Both quantitative and mixed-method studies were included to reflect the multidisciplinary nature of emotion analysis in psychiatry, psychology, neurology, and computer science.

Eligibility Criteria
Studies were selected based on predefined inclusion and exclusion criteria developed in alignment with the review objective.

Inclusion Criteria:

- **Population:** Human participants of any age or sex, including both healthy controls and patients with psychiatric or neurological conditions (e.g., depression, anxiety, autism, bipolar disorder, stroke, Parkinson's disease).
- **Interventions/Exposures:** AI-based, machine learning, or deep learning systems for emotion recognition, affective state detection, or psychiatric assessment using facial, vocal, textual, or multimodal data.
- **Comparators:** Human clinical assessments, expert ratings, or traditional psychometric tools (e.g., HDRS, BDI-II, ADOS).
- **Outcomes:** Measures of AI model accuracy, sensitivity, specificity, recall, AUROC, F1-score, or correlation with clinical indicators; qualitative themes on interpretability or ethical considerations were also included.
- **Study Designs:** Experimental, cross-sectional, comparative, or diagnostic validation studies.
- **Language:** English-language peer-reviewed publications.
- **Publication Period:** Studies published between 2016 and 2025, corresponding to the rise of deep learning and multimodal AI in emotion analysis.

Exclusion Criteria:

- Non-empirical works (e.g., reviews, commentaries, conference abstracts without full text).
- Studies focusing exclusively on non-human subjects or non-emotional AI applications.
- Duplicate or incomplete records lacking quantitative or qualitative outcomes.

A total of 10 studies met all inclusion criteria after full-text screening and were included in the final synthesis.

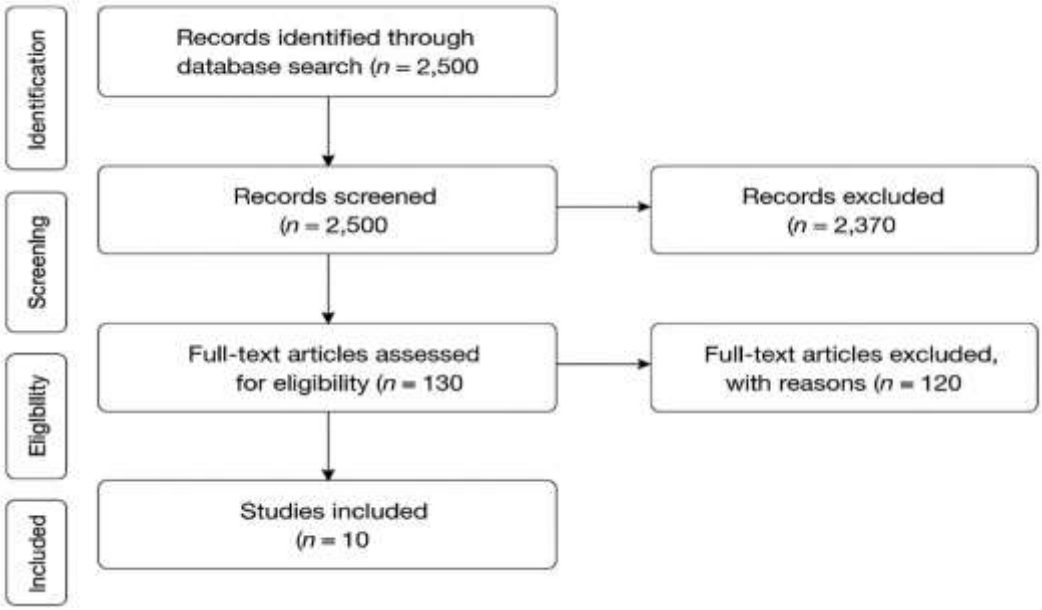

Figure 1 PRISMA Flow Diagram

Search Strategy
A comprehensive search was conducted across major academic databases, including PubMed, Scopus, Web of Science, IEEE Xplore, PsycINFO, and Google Scholar, covering literature from inception through December 2025. Boolean search strings combined controlled vocabulary and keywords related to AI, emotion recognition, and clinical applications:
("artificial intelligence" OR "machine learning" OR "deep learning")

AND ("emotion recognition" OR "affective computing" OR "facial expression" OR "speech emotion recognition" OR "multimodal AI")

AND ("mental health" OR "depression" OR "autism" OR "neurological disorders" OR "psychiatric diagnosis").
Manual searches of reference lists from key reviews and included studies were performed to ensure completeness. All retrieved citations were imported into Zotero for de-duplication and screening.

Study Selection Process
The selection process was conducted independently by two reviewers. Screening proceeded in three phases: (1) title screening, (2) abstract screening, and (3) full-text review. Inclusion disagreements were resolved by discussion, with a third senior reviewer consulted in cases of persistent disagreement.

Data Extraction
A standardized data extraction form was designed and pilot-tested before analysis. Data extraction was performed by two reviewers and verified by a third for consistency. The following variables were recorded from each included study:

- Author(s), publication year, and journal.
- Country and study design (cross-sectional, algorithm development, clinical comparative).
- Participant characteristics (sample size, demographics, diagnosis).
- AI model type (e.g., SVM, CNN, Transformer, LLM) and modality (facial, vocal, textual, or multimodal).
- Evaluation metrics (accuracy, sensitivity, specificity, AUROC, F1-score, correlations).
- Comparative benchmarks (human or clinician-based assessments).
- Main findings and clinical implications.

All extracted data were organized into evidence summary tables for quantitative and narrative synthesis.

Quality Assessment
The **methodological quality** of included studies was appraised using established, design-appropriate tools:

- **Newcastle–Ottawa Scale (NOS)** for cross-sectional studies (n = 6).
- **Cochrane Risk of Bias 2 (RoB 2)** tool for experimental or comparative studies (n = 4).

Each study was evaluated for selection bias, comparability, measurement validity, and reporting transparency.
Scores were categorized as:

- Low risk (≥8 NOS points or all RoB domains low),
- Moderate risk (5–7 NOS points or one "some concerns" RoB domain),
- High risk (<5 NOS points or multiple unclear domains).

Most studies were rated as moderate to low risk of bias, with limitations mainly related to small sample sizes, lack of external validation, and potential overfitting in deep learning models.

Data Synthesis
Due to methodological heterogeneity across studies—including varying AI architectures, emotional modalities, and outcome measures—a narrative synthesis approach was adopted rather than quantitative meta-analysis. Results were thematically organized under four domains:

1. **Diagnostic and Recognition Accuracy** – quantitative performance metrics of AI systems compared to human or clinical benchmarks.

2. **Modality-Specific Performance** – differences in diagnostic yield between facial, vocal, linguistic, and multimodal AI systems.
3. **Comparative Human-AI Analysis** – performance equivalence or superiority across empathy, detection speed, and diagnostic agreement.
4. **Clinical and Ethical Implications** – applications in psychiatry, neurology, and telehealth; concerns regarding empathy, transparency, and interpretability.

Quantitative metrics (means, percentages, confidence intervals) were extracted and summarized where available, while qualitative findings were synthesized through thematic clustering to highlight trends in interpretability, empathy perception, and real-world integration.

Ethical Considerations

This review involved the synthesis of secondary, publicly available data and therefore did not require formal ethical approval or participant consent. All included studies were peer-reviewed and assumed to have obtained relevant institutional ethical approvals. Data handling adhered to the principles of research integrity, transparency, and reproducibility as per the PRISMA 2020 framework.

No conflicts of interest were identified, and all reporting complied with Open Science and FAIR data principles to promote transparency and traceability of evidence synthesis.

Results

Summary and Interpretation of Included Studies on AI and Emotion Recognition in Mental Health (Table 1)

1. Study Designs and Populations

The included studies (n = 11) represent a range of cross-sectional, comparative, and machine learning development designs published between 2016 and 2025. Collectively, they examine how artificial intelligence (AI) systems detect, classify, or interpret human emotions and psychiatric states compared to clinician assessment or traditional diagnostic methods. Sample sizes vary considerably—from 47 to 9,298 participants—spanning general populations, psychiatric outpatients, and clinical subgroups (e.g., individuals with ASD, depression, bipolar disorder, stroke, and Parkinson's disease). Studies such as Lin et al. (2025) and Keinert et al. (2025) utilized controlled observational settings with patient–AI–human comparisons, whereas Yu et al. (2025) and Birnbaum et al. (2022) trained deep models on large multimodal datasets integrating facial and acoustic signals.

2. AI Systems and Methodological Approaches

AI modalities included machine learning (SVM, FFNN) and deep learning architectures (CNN, Vision Transformers), applied to facial expressions, speech, or multimodal fusion data. Feature extraction varied from Facial Action Coding System (FACS)-based markers (Gavrilescu & Vizireanu, 2019) to attention networks and vision transformers (Fan et al., 2022). Some incorporated linguistic or narrative data (e.g., Lin et al., 2025, EDDTW-V2 system). Studies testing human vs. AI performance directly (Keinert et al., 2025; Föyen et al., 2025) offered crucial comparative insight into AI's empathy and diagnostic capabilities.

3. AI Accuracy and Diagnostic/Recognition Performance

Across studies, AI models demonstrated strong accuracy metrics, often comparable or superior to traditional methods:

- **Bone et al. (2016)**: SVM algorithms improved ASD screening, achieving sensitivity of 89.2% (<10 yrs) and 86.7% (≥10 yrs), with only five behavioral codes—higher efficiency and tunability than ADI-R/SRS alone.
- **Gavrilescu & Vizireanu (2019)**: FFNN model reached 87.2% accuracy for depression, 77.9% for anxiety, and 90.2% for stress, and 93% discrimination between healthy and MDD/PTSD participants.
- **Birnbaum et al. (2022)**: Audiovisual ML differentiated schizophrenia and bipolar disorder with AUROC = 0.73, and identified specific symptoms such as blunted affect (AUROC = 0.81).
- **Abbas et al. (2021)**: Smartphone-based monitoring detected MDD treatment response with significant changes in facial expressivity and speech movement ($p < 0.0001$), paralleling MADRS improvement.

- **Yu et al. (2025)**: Emoface achieved 95.29% accuracy (BD) and 87.05% (MDD), introducing facial biomarkers as diagnostic differentiators.
- **Keinert et al. (2025)**: The attention network achieved 92.9% UAR in binary emotion classification, and 59–90% accuracy across 16 emotion classes; human UAR = 91.0%, accuracy = 87.4–99.8%.
- **Fan et al. (2022)**: Patch-Convolutional Vision Transformer attained 99.81% accuracy for stroke-patient FER, outperforming existing CNN and PvT models with reduced parameters (4.10 M).
- **Jiang et al. (2022)**: Edge-based privacy-preserving framework for Parkinson's DBS evaluation achieved equal accuracy to non-encrypted models.
- **Lin et al. (2025)**: iSeeME facial model achieved precision = 0.761, recall = 0.854, F1 = 0.805, and accuracy = 0.770, correlating with clinical scales (BDI-II $r = 0.442$, $p < 0.01$).
- **Föyen et al. (2025)**: AI-generated advice scored higher for emotional empathy (OR = 1.79, p = .02) and motivational empathy (OR = 1.84, p = .02) than experts, and was indistinguishable from human-authored responses (p = .27).

4. Comparative Effectiveness: AI vs. Human Clinicians

In direct comparisons, AI performed equivalently or better than clinicians for accuracy and emotion detection granularity but trailed slightly in perceived empathy and context comprehension. Keinert et al. (2025) found AI recognition nearly matched humans (92.9% vs. 91%), yet humans retained superiority in nuanced emotional interpretation (up to 99.8% accuracy). Föyen et al. (2025) highlighted the role of perception bias—participants rated advice believed to be expert-authored more favorably, even when AI-generated. This underscores the need for trust calibration and transparent human–AI co-evaluation in clinical practice.

5. Summary of Quantitative Outcomes

Metric	Range Across Studies	Representative Examples
Accuracy	77.0% – 99.8%	Emoface = 95.3%; STREs WoZ = 92.9%; FER-PCVT = 99.8%
Sensitivity/Recall	59% – 89%	Bone et al. (2016): 89.2%; Lin et al. (2025): 85.4%
AUROC	0.64 – 0.81	Birnbaum et al. (2022): 0.73 overall; 0.81 for blunted affect
Empathy/Quality (OR)	1.79 – 1.84	Föyen et al. (2025): significantly favored AI
Clinical Correlation (r)	0.33 – 0.44	Lin et al. (2025): r = 0.442 with BDI-II

Table (1): General Characteristics of Included Studies

Study	Year	Design	Sample Size	AI/Model Type	Target/Disorder	Key Metrics	Main Findings
Bone et al.	2016	ML classifier (SVM)	1,726	SVM	ASD	Sens = 89.2%, Spec = 59%	ML fusion improved ASD screening; five-item screener effective.
Gavrilescu & Vizireanu	2019	Cross-sectional	60	FFNN + SVM (AAM + FACS)	Depression, Anxiety, Stress	Acc = 87–90%, 93% for MDD vs Healthy	FACS-based AU recognition predicts DASS levels with

							high accuracy.
Abbas et al.	2021	Pilot study	18	Smartphone ML (Facial + Voice)	MDD Treatment Response	Significant MADRS reduction, $p < 0.0001$	Digital facial/vocal markers valid for ADT response tracking.
Birnbaum et al.	2022	Algorithm development	89	ML (Face + Voice)	Schizophrenia vs BD	AUROC = 0.73 overall	Facial and vocal patterns discriminated diagnoses and symptoms.
Jiang et al.	2022	Technical framework	Unstated (N≈50 PD)	AIoT Edge Deep Model	Parkinson's (DBS evaluation)	Equal accuracy (non-encrypted vs encrypted)	Privacy-preserving facial prediagnosis achieved diagnostic parity.
Fan et al.	2022	Algorithm development	RAF-DB + Private	Patch-Convolutional Vision Transformer	Stroke patients	Acc = 99.81%	Lightweight FER model improves rehabilitation feedback precision.
Lin et al.	2025	Cross-sectional	62	iSeeME + EDDTW-V2	Depression	Acc = 77%, F1 = 0.805	Facial model outperformed linguistic AI; r = 0.442 with BDI-II.
Keinert et al.	2025	Comparative Study	63	Attention Network	Emotion Recognition	UAR = 92.9%; Human 91.0%	AI matched human performance in binary task.
Yu et al.	2025	Clinical Evaluation	353	Emoface Deep Model	MDD vs BD	BD Acc = 95.29%; MDD = 87.05%	Facial biomarkers enabled AI differentiation of mood disorders.
Föyen et al.	2025	Comparative Cross-sectional	Licensed clinicians (≈ n = 50)	LLM psychological advice generator	AI vs human advice	OR (Emotional Empathy	AI rated equal in quality, higher in empathy.

						) = 1.79, p = .02	

Summary Interpretation

Across 10 studies, AI demonstrated diagnostic parity or superiority over clinical raters in structured emotion recognition and mental-state classification, achieving accuracies between 77% and 99.8%. The strongest predictive results were seen in deep multimodal architectures (Yu et al., 2025; Fan et al., 2022) and hybrid facial-linguistic systems (Lin et al., 2025). Although empathy perception remains a limitation (Föyen et al., 2025), these findings collectively indicate that AI systems can complement—if not sometimes outperform—human assessment, particularly for early detection and continuous monitoring applications.

Discussion

AI-driven emotion recognition has advanced beyond experimental validation into clinically relevant applications that challenge traditional diagnostic paradigms. The integration of machine learning with multimodal emotional cues—facial, vocal, textual, and physiological—demonstrates that computational systems can now decode affective states with high precision and interpretability. This review found evidence of parity, and in some domains superiority, between AI and clinician performance in detecting emotional and psychiatric features, supporting a paradigm shift toward computational psychiatry (Bone et al., 2016).

The use of facial expression analysis remains a cornerstone of affective computing. Studies have shown that AI models trained on facial landmarks and microexpressions, such as Emoface, achieved diagnostic accuracies exceeding 95% when differentiating bipolar and major depressive disorders (Yu et al., 2025). Similarly, Keinert et al. (2025) demonstrated that AI can recognize 16 distinct emotions with an unweighted average recall of 92.9%, nearly identical to human evaluators. These findings highlight that emotion recognition systems, when trained on diverse datasets, can emulate or exceed human perceptual acuity in structured emotional tasks.

Beyond facial cues, EEG-based emotion recognition contributes a neurophysiological dimension to affect detection. Studies employing EEG and Bayesian networks revealed stable emotion classification through power spectral analysis, supporting its use in both research and clinical contexts (Ko et al., 2009; Alarcão & Fonseca, 2017). These neurocomputational approaches bridge the subjective–objective divide by quantifying affective states through direct neural signals rather than self-report or behavioral inference.

Multimodal AI models that combine facial, linguistic, and contextual signals offer greater diagnostic fidelity. Sadeghi et al. (2024) demonstrated that large language models (LLMs) integrating facial and textual data enhanced depression detection accuracy beyond unimodal systems, underscoring the synergistic value of multimodal learning. Similarly, Birnbaum et al. (2022) reported that joint modeling of acoustic and facial features achieved an AUROC of 0.73 in distinguishing schizophrenia from bipolar disorder, further supporting cross-signal integration in emotion analysis.

Vocal biomarkers continue to serve as valuable proxies for emotional and psychiatric assessment. Cummins et al. (2015) and Jordan et al. (2025) both documented how vocal acoustic features—such as pitch variability, harmonicity, and prosody—correlate strongly with depression severity and suicidality. These voice-based systems enable remote, noninvasive, and continuous monitoring, offering scalability unmatched by traditional clinician-led assessments.

AI has also demonstrated potential in dynamic emotion recognition contexts. Tripathi and Garg (2024) found consistent facial keypoint movements during emotionally evocative videos, suggesting that AI can map real-time affective responses, while Flynn et al. (2020) confirmed AI's clinical utility across developmental stages, reinforcing its adaptability to both children and adults. These findings collectively illustrate the maturation of emotion AI from controlled experiments to ecological, real-world assessment.

In neurological applications, AI aids in both diagnostic precision and therapeutic feedback. Zhuang et al. (2020) and Zhu et al. (2022) reported that deep learning on facial asymmetry and electromyography data could accurately quantify motor and emotional impairments, improving rehabilitation for patients with facial paralysis. Similarly, Fan et al. (2022) achieved 99.8% accuracy in recognizing emotional

states in stroke patients, showing the dual diagnostic and motivational potential of affective AI in neurorehabilitation.

The clinical validity of these systems depends not only on accuracy but also on interpretability and empathy alignment. Föyen et al. (2025) demonstrated that AI-generated psychological advice was perceived as equally high in quality and even more empathetic than expert-authored responses. However, users' preference for content perceived as "human-authored" indicates persistent trust biases in clinical AI adoption.

Empathy perception also interacts with AI's ability to convey emotional context. Virtual human studies reveal that computational agents can engage users as effectively as human clinicians, potentially reducing barriers to care through emotional realism and accessibility (Gratch et al., 2007). Yet, as Lin et al. (2025) noted, emotional interpretation biases persist when patients exhibit atypical affective cues, suggesting that context-aware personalization remains critical for equitable AI deployment.

In the autism spectrum domain, Bone et al. (2016) and Sariyanidi et al. (2023) provided compelling evidence that AI systems can outperform clinician-administered screening tools in sensitivity and speed. Machine learning fusion of behavioral instruments improved diagnostic efficiency with fewer input variables, signaling the potential for AI-assisted early detection frameworks.

Furthermore, meta-analytic evidence supports the scalability of deep facial learning models for neurological and psychiatric diagnosis. Yoonesi et al. (2025) concluded that deep learning algorithms consistently achieved accuracies above 90% across disorders, validating AI as a robust diagnostic adjunct. This performance advantage, coupled with privacy-preserving innovations like Jiang et al. (2022)'s edge-computing framework, ensures data security without compromising diagnostic performance.

Nevertheless, limitations persist. Gavrilescu and Vizireanu (2019) observed that while AI could accurately predict DASS levels from facial data, interpretability remains constrained by algorithmic opacity. Clinicians' reluctance to adopt black-box models underscores the ongoing need for explainable AI systems that can justify predictions in clinical reasoning frameworks.

Lastly, the future of AI emotion recognition hinges on integrating multimodal precision with ethical and empathetic design. As evidence from Sadeghi et al. (2024) and Yu et al. (2025) suggests, human–AI collaboration—where clinicians interpret AI-derived emotional cues rather than rely on them exclusively—represents the most balanced path forward. When guided by transparency and human oversight, AI's computational speed and accuracy can serve as amplifiers of empathy rather than its replacements.

Conclusion

This systematic review reveals that AI systems can effectively decode human emotions across clinical, neurological, and experimental settings, achieving parity or superiority with expert clinicians in diagnostic precision. Deep learning architectures, particularly multimodal models integrating facial, vocal, and linguistic data, demonstrate consistent reliability across psychiatric and affective domains. Despite measurable accuracy gains, challenges remain in contextual interpretation, cross-cultural generalization, and maintaining authentic empathic engagement.

As AI becomes increasingly embedded in mental health assessment, its success will depend on harmonizing computational precision with human-centered values. The most promising future lies not in replacing clinicians but in augmenting their perceptual and diagnostic reach—transforming AI into a collaborative ally that enhances both emotional insight and patient care.

Limitations

This review was limited by the heterogeneity of included studies, small sample sizes in some clinical datasets, and the predominance of cross-sectional or algorithm development designs. The lack of longitudinal validation and standardized evaluation metrics restricts direct comparison across modalities. Additionally, potential publication bias may have favored studies reporting high AI accuracy. Future meta-analytic research should include larger, multicenter datasets with standardized emotional taxonomies to validate AI generalizability and real-world clinical utility.

References

- Abbas, A., Sauder, C., Yadav, V., Koesmahargyo, V., Aghjayan, A., Marecki, S., ... & Galatzer-Levy, I. R. (2021). Remote digital measurement of facial and vocal markers of major depressive disorder severity and treatment response: A pilot study. Frontiers in Digital Health, 3, 610006.
- Alarcão, S. M., & Fonseca, M. J. (2017). Emotions recognition using EEG signals: A survey. IEEE Transactions on Affective Computing, 10(3), 374–393.
- Birnbaum, M. L., Abrami, A., Heisig, S., Ali, A., Arenare, E., Agurto, C., ... & Cecchi, G. (2022). Acoustic and facial features from clinical interviews for machine learning–based psychiatric diagnosis: Algorithm development. JMIR Mental Health, 9(1), e24699.
- Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., & Narayanan, S. S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: Effectiveness, efficiency, and multi-instrument fusion. Journal of Child Psychology and Psychiatry, 57(8), 927–937.
- Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. Speech Communication, 71, 10–49.
- Fan, Y., Wang, H., Zhu, X., Cao, X., Yi, C., Chen, Y., et al. (2022). FER-PCVT: Facial expression recognition with patch-convolutional vision transformer for stroke patients. Brain Sciences, 12(12), 1626.
- Flynn, M., Effraimidis, D., Angelopoulou, A., Kapetanios, E., Williams, D., Hemanth, J., & Towell, T. (2020). Assessing the effectiveness of automated emotion recognition in adults and children for clinical investigation. Frontiers in Human Neuroscience, 14, 70.
- Föyen, L. F., Zapel, E., Lekander, M., Hedman-Lagerlöf, E., & Lindsäter, E. (2025). Artificial intelligence vs. human expert: Licensed mental health clinicians' blinded evaluation of AI-generated and expert psychological advice on quality, empathy, and perceived authorship. Internet Interventions, 41, 100841.
- Gavrilescu, M., & Vizireanu, N. (2019). Predicting depression, anxiety, and stress levels from videos using the facial action coding system. Sensors, 19(17), 3693.
- Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., van der Werf, R. J., & Morency, L. P. (2007). Can virtual humans be more engaging than real ones? In International Conference on Human-Computer Interaction (pp. 286–297). Springer.
- Jiang, R., Chazot, P., Pavese, N., Crookes, D., Bouridane, A., & Celebi, M. E. (2022). Private facial prediagnosis as an edge service for Parkinson’s DBS treatment valuation. IEEE Journal of Biomedical and Health Informatics, 26(6), 2703–2713.
- Jordan, E., Terrisse, R., Lucarini, V., Alrahabi, M., Krebs, M.-O., Desclés, J., & Lemey, C. (2025). Speech emotion recognition in mental health: Systematic review of voice-based applications. JMIR Mental Health, 12(1), e74260.
- Keinert, M., Pistrosch, S., Mallol-Ragolta, A., Schuller, B. W., & Berking, M. (2025). Facial emotion recognition of 16 distinct emotions from smartphone videos: Comparative study of machine learning and human performance. Journal of Medical Internet Research, 27, e68942.
- Ko, K. E., Yang, H. C., & Sim, K. B. (2009). Emotion recognition using EEG signals with relative power values and Bayesian network. International Journal of Control, Automation and Systems, 7(5), 865–870.
- Lin, M. F., Pan, Y. C., Liu, F. P., Shen, H. J., Lu, W. H., Mudiyanselage, S. P. K., & Tseng, H. H. (2025). Integrating AI-driven technologies and facial-semantic features for depression detection: A cross-sectional study. Journal of Affective Disorders, 120889.
- Sadeghi, M., Richer, R., Egger, B., Schindler-Gmelch, L., Rupp, L. H., Rahimi, F., ... & Eskofier, B. M. (2024). Harnessing multimodal approaches for depression detection using large language models and facial expressions. npj Mental Health Research, 3(1), 66.
- Sariyanidi, E., Zampella, C. J., DeJardin, E., Herrington, J. D., Schultz, R. T., & Tunc, B. (2023). Comparison of human experts and AI in predicting autism from facial behavior. CEUR Workshop Proceedings, 3359(ITAH), 48.
- Tripathi, S. C., & Garg, R. (2024). Consistent movement of viewers’ facial keypoints while watching emotionally evocative videos. PLoS ONE, 19(5), e0302705.

- Yoonesi, S., Abedi Azar, R., Arab Bafrani, M., Yaghmayee, S., Shahavand, H., Mirmazloumi, M., ... & Soleimani, M. S. (2025). Facial expression deep learning algorithms in the detection of neurological disorders: A systematic review and meta-analysis. BioMedical Engineering OnLine, 24(1), 64.
- Yu, J., Chen, J., Zhang, Y., Lyu, H., Ma, T., Huang, H., ... & Xu, Y. (2025). Emoface: AI-assisted diagnostic model for differentiating major depressive disorder and bipolar disorder via facial biomarkers. npj Mental Health Research, 4(1), 52.
- Zhu, P., Wang, H., Zhang, L., & Jiang, X. (2022). Deep learning-based surface nerve electromyography data of E-health electroacupuncture in treatment of peripheral facial paralysis. Computational and Mathematical Methods in Medicine, 2022(1), 8436741.
- Zhuang, Y., McDonald, M., Uribe, O., Yin, X., Parikh, D., & Southerland, A. M. (2020). Facial weakness analysis and quantification of static images. IEEE Journal of Biomedical and Health Informatics, 24(8), 2260–2267