



Multimodal Digital Signals for AI-Assisted Detection of Depression and Anxiety: A Comparative Analysis of Questionnaire, Language, Speech and Wearable Data

Mansi*¹, Dr. Grace M. BasanShrieh*²

*1 Phd Scholar, Department of Psychology, William Caery University, Meghalaya, India

*2 Assistant Professor and Guide, Department of Psychology, William Caery University, Meghalaya, India

Abstract- No single source of information captures the full experience of depression or anxiety. Artificial intelligence (AI) research has therefore turned to multiple digital signals, including standardised questionnaires, electronic health records, written language, speech, facial behaviour and passively sensed smartphone and wearable data. This paper compares these modalities for AI-assisted detection of depression and anxiety. Drawing on the literature synthesised in a doctoral study, it maps each modality to the symptom domains it can reflect (cognitive, emotional, behavioural, physiological and functional), summarises the features typically extracted, and appraises each against six practical criteria: clinical validity, scalability, participant burden, privacy protection, interpretability and cross-cultural robustness. It also compares early, late and hybrid fusion strategies. The analysis shows that the modalities are complementary rather than interchangeable. Questionnaires and language are strongest for cognitive and emotional symptoms, speech and facial signals for psychomotor and affective expression, and wearables for behavioural and physiological rhythms. Late fusion of modality-specific models offers the best balance of robustness to missing data and auditability for early clinical research. The paper concludes with design recommendations for building multimodal datasets that can support valid, fair and clinically interpretable AI screening.

Keywords- Depression; anxiety; multimodal machine learning; digital phenotyping; natural language processing; speech analysis; wearable sensing; data fusion

I. Introduction

Depressive and anxiety disorders affect an estimated 280 million and 301 million people worldwide, respectively [1], and both became markedly more common during the COVID-19 pandemic [2]. They often co-occur [3], and their expression varies widely between individuals: one person may present mainly with low mood and negative thinking, another with sleep disturbance, fatigue and withdrawal. Standard assessment relies on clinical interview supported by self-report instruments such as the PHQ-9 [4] and GAD-7 [5]. These instruments are accurate, but they capture only what a person is able and willing to report at one moment.

Artificial intelligence (AI) offers a way to supplement self-report with other signals that people produce in daily life, such as the words they write, the way they speak and their



patterns of movement and sleep [6], [7]. Reviews of AI for depression and anxiety document a growing number of studies using these modalities [8]–[10], but they seldom compare the modalities side by side on criteria that matter for clinical use. This paper aims to fill that gap. Its objectives are (i) to map each data modality to the symptom domains it can reflect, (ii) to summarise the features and methods used for each, (iii) to compare modalities on practical criteria for clinical research, and (iv) to evaluate strategies for fusing them.

II. Method of Analysis

The paper uses a structured comparative analysis of peer-reviewed literature identified during the doctoral study, searched in PubMed, Scopus, IEEE Xplore and Google Scholar. For each modality, the evidence was coded for the symptom domains captured, the representative features, the typical algorithms, and the reported strengths and limitations. Domain coverage (Fig. 1) and practical ratings (Fig. 2) were assigned by the authors on ordinal scales, based on consistency of evidence across the cited sources. They are qualitative syntheses intended to support design decisions, not pooled effect sizes.

III. Modality-by-Modality Analysis

1. Questionnaires and Clinical Records

Questionnaire items directly sample cognitive and emotional symptoms. A PHQ-9 score of 10 or more has shown sensitivity and specificity of about 0.88 against structured interviews [4], [11], and the GAD-7 performs similarly for anxiety [5]. ML can weight individual items and response patterns, but when a questionnaire cut-off is also used as the outcome label, the model learns the instrument rather than the disorder. Electronic health records add longitudinal information on diagnoses, prescriptions and service use, which makes them particularly informative about functional impact. They also carry inconsistent coding and site-specific practices [12].

2. Written Language

People with depression use more first-person singular pronouns and more negative emotion words than non-depressed peers [13]. Social media studies have extended these findings to large, naturalistic samples [14]. When consented Facebook posts were linked to medical records, language predicted a later clinical diagnosis of depression [15]. Transformer models such as BERT capture context and negation better than word-count dictionaries [16], and large language models add further capability along with risks of hallucination and privacy exposure [17]. Language is rich for cognitive and emotional content but reveals little about behavioural and physiological change.

3. Speech and Facial Behaviour

Depression is associated with monotone prosody, slower speech, longer pauses and altered voice quality, which reflect psychomotor retardation [18]. Benchmark challenges such as AVEC 2016 established shared audio-visual datasets and evaluation protocols for depression recognition from interviews [19]. A systematic review found promising accuracy across psychiatric disorders but noted small samples, inconsistent



feature sets and little cross-corpus testing [20]. Acoustic features are sensitive to recording conditions, language and accent, so models trained in one setting often transfer poorly to another.

4. Smartphone and Wearable Sensing

Digital phenotyping uses smartphones and wearables to measure behaviour continuously and passively [7], [21]. GPS-derived features such as location variance and circadian regularity have been associated with PHQ-9 severity [22]. Accelerometry, heart rate and sleep timing add physiological information. These signals carry little participant burden and capture behavioural and somatic change well, but they reflect emotional and cognitive states only indirectly. They are also affected by adherence, battery life and device differences [23].

	Cognitive	Emotional	Behavioural	Physiological / somatic	Functional
Questionnaire	High	High	Moderate	Moderate	Moderate
EHR / clinical notes	Moderate	Moderate	Moderate	Moderate	High
Language (text)	High	High	Low	Low	Low
Speech	Low	Moderate	Moderate	Moderate	Low
Facial / behavioural	Low	High	Moderate	Low	Low
Wearable / smartphone	Low	Low	High	High	Moderate

Fig. 1. Qualitative coverage of the five symptom domains by each data modality, synthesised by the authors from the reviewed literature [4], [6], [14], [18], [22]. Darker cells indicate stronger and more consistent evidence.

Table 1 summarises representative features and typical algorithms for each modality.

Table 1. Representative features and algorithms by modality

Modality	Representative features	Typical algorithms
Questionnaire	Item scores, totals, response latency	Logistic regression, random forest [24]
EHR	Diagnoses, prescriptions, utilisation, clinical notes	Gradient boosting [25], NLP on notes
Language	Pronoun and affect indices, topics, contextual embeddings	SVM [26], transformers [16]
Speech	Pitch, energy, pause length, speech rate, spectral features	SVM, CNN, LSTM [27]
Facial / behavioural	Action units, gaze, head movement	CNN, LSTM [28]
Wearable / smartphone	Location variance, steps, sleep, heart rate, phone use	Random forest, gradient boosting, LSTM

IV. Comparative Appraisal

Fig. 2 compares four principal modalities on six practical criteria. Questionnaires rate highest on clinical validity and interpretability but need active participation. Language scales most easily, since text is abundant, but is weaker on privacy and on cross-cultural robustness, because expressions of distress differ across languages and cultures. Speech occupies a middle position. Wearables minimise participant burden but have the weakest direct link to clinical diagnosis. No modality dominates on all criteria, which is the main argument for combining them.

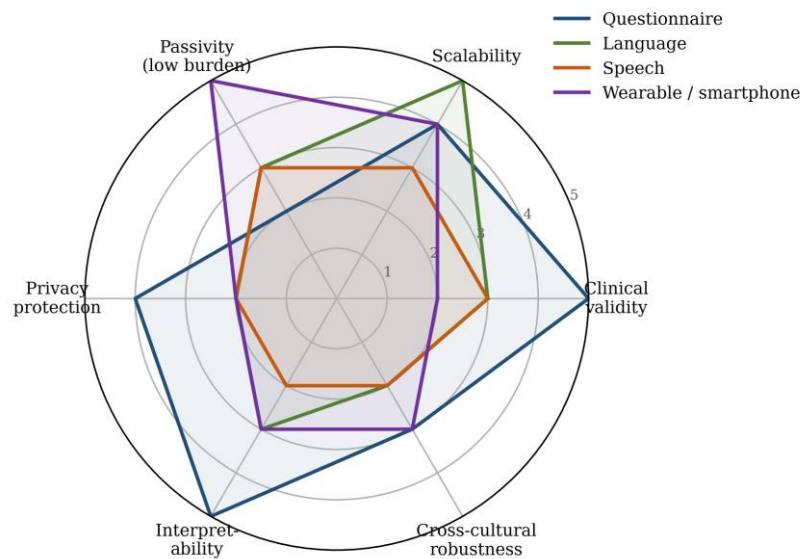


Fig. 2. Comparative appraisal of four modalities on six practical criteria (1 = weak, 5 = strong). Ratings are the authors' qualitative synthesis of the reviewed literature.

V. Fusion Strategies

Multimodal ML integrates heterogeneous inputs through three broad strategies [29], illustrated in Fig. 3 and compared in Table 2. Early fusion concatenates features into one vector for a single classifier. It can capture cross-modal interactions but fails when any modality is missing. Late fusion trains a separate model for each modality and combines their outputs by weighted averaging or stacking. It tolerates missing modalities and lets each channel be validated and explained separately. Hybrid fusion learns a joint representation through shared layers or attention [30]. It is the most expressive option but needs the most data and is the hardest to interpret.

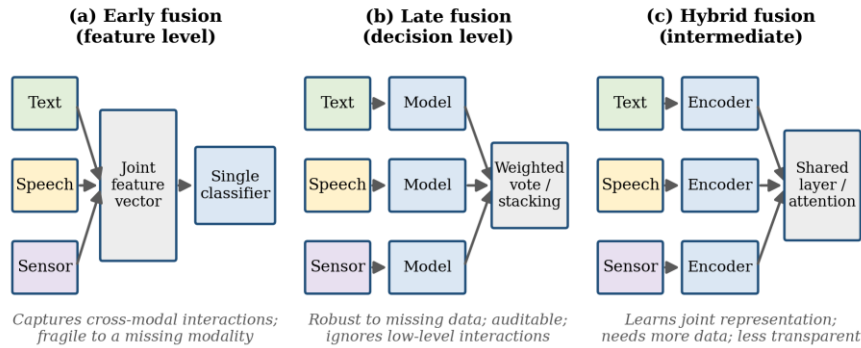


Fig. 3. Early, late and hybrid fusion strategies for combining text, speech and sensor data, following the taxonomy of Baltrušaitis *et al.* [29].

Table 2. Comparison of fusion strategies for clinical research

Criterion	Early fusion	Late fusion	Hybrid fusion
Cross-modal interaction	Captured	Limited	Captured
Missing modality	Fails or needs imputation	Handled naturally	Partly handled
Data requirement	Moderate	Low to moderate	High
Explainability	Moderate	High (per modality)	Low
Suitability for small clinical samples	Moderate	High	Low

VI. Discussion and Recommendations

The analysis supports four recommendations for multimodal AI research in depression and anxiety. First, **anchor the outcome to a clinical reference standard**, such as a structured interview, and treat questionnaires as predictors rather than labels [11]. Second, **start with late fusion** of modality-specific models and a transparent logistic-regression baseline. This suits the modest sample sizes typical of clinical studies [31] and keeps each channel auditable [32]. Third, **plan for heterogeneity**: collect data across languages, devices and settings, and validate externally, because acoustic and linguistic signals shift with context [33]. Fourth, **minimise data and burden**: gather each additional modality only when it is expected to add information, under separate consent and with strong privacy safeguards [34].

For research in multilingual regions such as North-East India, the cross-cultural fragility of language and speech models is a particular concern. Models developed in English-language or Western samples should not be assumed to transfer. Locally collected, consented and validated datasets are required, as are fairness checks across language groups [35].



1. Limitations

The coverage and appraisal ratings are qualitative syntheses and could vary with the evidence selected. The comparison does not pool performance metrics, because study designs, labels and populations differ too much. Empirical validation of the recommended late-fusion design on a locally collected dataset is the next step of the doctoral work.

VII. Conclusion

Questionnaire, record, language, speech and sensor data each provide a partial and complementary view of depression and anxiety. Combining them through a transparent, missing-data-tolerant fusion strategy and validating the result against clinical diagnosis across diverse populations offers the most credible path toward multimodal AI screening that supports, rather than replaces, professional judgement.

Conflict of Interest

The authors declare no conflict of interest.

References

1. GBD 2019 Mental Disorders Collaborators, "Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019," *Lancet Psychiatry*, vol. 9, no. 2, pp. 137–150, 2022, doi: 10.1016/S2215-0366(21)00395-3.
2. COVID-19 Mental Disorders Collaborators, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," *Lancet*, vol. 398, no. 10312, pp. 1700–1712, 2021, doi: 10.1016/S0140-6736(21)02143-7.
3. N. H. Kalin, "The critical relationship between anxiety and depression," *Am. J. Psychiatry*, vol. 177, no. 5, pp. 365–367, 2020, doi: 10.1176/appi.ajp.2020.20030305.
4. K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.
5. R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The GAD-7," *Arch. Intern. Med.*, vol. 166, no. 10, pp. 1092–1097, 2006, doi: 10.1001/archinte.166.10.1092.
6. D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal sensing: Understanding mental health using ubiquitous sensors and machine learning," *Annu. Rev. Clin. Psychol.*, vol. 13, pp. 23–47, 2017, doi: 10.1146/annurev-clinpsy-032816-044949.
7. T. R. Insel, "Digital phenotyping: Technology for a new science of behavior," *JAMA*, vol. 318, no. 13, pp. 1215–1216, 2017, doi: 10.1001/jama.2017.11295.
8. Y. Park, S. Park, and M. Lee, "Effectiveness of artificial intelligence in detecting and managing depressive disorders: Systematic review," *J. Affect. Disord.*, vol. 361, pp. 445–456, 2024, doi: 10.1016/j.jad.2024.06.035.



9. F. Zafar et al., "The role of artificial intelligence in identifying depression and anxiety: A comprehensive literature review," *Cureus*, vol. 16, no. 3, Art. no. e56472, 2024, doi: 10.7759/cureus.56472.
10. A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychol. Med.*, vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151.
11. B. Levis, A. Benedetti, and B. D. Thombs, "Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis," *BMJ*, vol. 365, Art. no. 11476, 2019, doi: 10.1136/bmj.11476.
12. C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, Art. no. 195, 2019, doi: 10.1186/s12916-019-1426-2.
13. S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cogn. Emot.*, vol. 18, no. 8, pp. 1121–1133, 2004, doi: 10.1080/02699930441000030.
14. S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: An integrative review," *Curr. Opin. Behav. Sci.*, vol. 18, pp. 43–49, 2017, doi: 10.1016/j.cobeha.2017.07.005.
15. J. C. Eichstaedt et al., "Facebook language predicts depression in medical records," *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 44, pp. 11203–11208, 2018, doi: 10.1073/pnas.1802331115.
16. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
17. N. Obradovich et al., "Opportunities and risks of large language models in psychiatry," *NPP—Digit. Psychiatry Neurosci.*, vol. 2, no. 1, Art. no. 8, 2024, doi: 10.1038/s44277-024-00010-z.
18. N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, 2015, doi: 10.1016/j.specom.2015.03.004.
19. M. Valstar et al., "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 3–10, doi: 10.1145/2988257.2988258.
20. D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investig. Otolaryngol.*, vol. 5, no. 1, pp. 96–116, 2020, doi: 10.1002/liv.2.354.
21. J.-P. Onnela and S. L. Rauch, "Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health," *Neuropsychopharmacology*, vol. 41, no. 7, pp. 1691–1696, 2016, doi: 10.1038/npp.2016.7.
22. S. Saeb et al., "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study," *J. Med. Internet Res.*, vol. 17, no. 7, Art. no. e175, 2015, doi: 10.2196/jmir.4273.



23. J. Torous et al., "The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality," *World Psychiatry*, vol. 20, no. 3, pp. 318–335, 2021, doi: 10.1002/wps.20883.
24. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
25. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
26. C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.
27. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
28. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
29. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.
30. Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013, doi: 10.1109/TPAMI.2013.50.
31. R. D. Riley et al., "Calculating the sample size required for developing a clinical prediction model," *BMJ*, vol. 368, Art. no. m441, 2020, doi: 10.1136/bmj.m441.
32. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.
33. S. G. Finlayson et al., "The clinician and dataset shift in artificial intelligence," *N. Engl. J. Med.*, vol. 385, no. 3, pp. 283–286, 2021, doi: 10.1056/NEJMc2104626.
34. W. N. Price II and I. G. Cohen, "Privacy in the age of medical big data," *Nat. Med.*, vol. 25, no. 1, pp. 37–43, 2019, doi: 10.1038/s41591-018-0272-7.
35. A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Ann. Intern. Med.*, vol. 169, no. 12, pp. 866–872, 2018, doi: 10.7326/M18-1990.