East Journal of Computer Science

ISSN: 3079-9406

Volume 2, Issue 1
Publisher: East Publication & Technology
DOI: https://doi.org/10.63496/ejcs.Vol2.Iss1.246

# Multimodal Machine Learning Techniques for Depression Detection: A Systematic Literature Review

***Mohammed Essam Arif*[1], Muhammad Irsyad Abdullah[1]***

[1]Computer Science, School of Graduate Studies, Management and Science University, Kuala Lumpur, Malaysia.
*Corresponding Author: Mohammad-arif@hotmail.com

**Received**: 01/01/2026, **Revised**: 27/01/2026, **Accepted**: 31/01/2026, **Published**: 28/02/2026

***Abstract*:**

*Depression is a widespread mental health disorder that poses serious social, economic, and healthcare challenges worldwide. Conventional diagnostic approaches rely primarily on self-reported questionnaires and clinical interviews, which are subjective and may lead to delayed or inaccurate diagnosis. The past few years have seen the development of machine learning and deep learning technologies as useful mechanisms of automatic depression detection. The multimodal machine learning methods have been widely studied among them, as they can combine heterogeneous data modals, including speech, text, facial expression, and physiological modals. This is a systematic literature review analyzing the current studies regarding multimodal machine learning methods in the detection of depression. In accordance with a developed review protocol, applicable studies were determined, screened, and evaluated according to a set of inclusion and exclusion criteria. Following the PRISMA guidelines, a total of 32 studies were included in this synthesis to ensure methodological clarity and reproducibility. The review summarises the results associated with data modalities, modelling methods, fusion approaches, data sets, and metrics of evaluation. In most reviewed studies, multimodal models improved F1-score or accuracy over unimodal baselines, particularly on DAIC-WOZ and AVEC; however, gains were smaller or inconsistent on small or imbalanced datasets. Nevertheless, issues of data scarcity, ethical issues, and overall lack of generalisability are still notable. The review gives an in-depth synthesis of the state of the art and emerges major research gaps to inform the future developments in automated depression detecting systems.*

***Keywords*: Depression Detection, Multimodal Machine Learning, Deep Learning, Mental Health, Systematic Literature Review**

## 1. Introduction

Depression is a long-term mental condition that is typified by low mood, loss of interest, impaired thinking, and behaviour. It is known to be one of the most important causes of disability and one of the greatest contributors to the burden of disease in general throughout the world. Although it is highly prevalent, depression is highly undiagnosed because of social stigma, poor access to mental health services, and the subjective nature of more traditional diagnostic methods [1]. Widespread assessment measures are largely based on both self-reported surveys and clinician-evaluation, and both methods are susceptible to recall bias, cultural manipulation, and incoherent interpretation.

In the recent past, artificial intelligence has been advanced to develop automated methods of mental health evaluation. Machine learning methods have been applied more and more to identify depression based on the behavioural and psychological characteristics of speech, textual content, facial expressions, and physiological indicators [2]. The initial studies in this category were mainly unimodal, in which only one source of data was employed to forecast depressive symptoms. These approaches despite showing encouraging results were not always able to reflect the multidimensionality and complexity of depression.

Human mood and mental conditions can be considered as multimodal ones as they include verbal, vocal, visual, and biological manifestations. To counter this complexity, multimodal machine learning solutions have appeared where multiple complementary data are combined in order to produce more powerful and informative representations of mental health conditions. Studies have indicated that heterogeneous modalities are more accurate in diagnostic results and more resilient to noise and missing data as compared to systems that are unimodal [3]. Besides, the growing access to multimodal data and the development of deep learning methods over the recent past has increased the pace of research in this field of study.

In spite of the increased research, the literature is still disjointed in the various types of data, modelling methods and assessment procedures. Although a number of reviews have reviewed the detection of depression with specific individual perspectives, e.g. speech-based or text-based detection approaches, there is no systematic review of literature that specifically synthesises multimodal machine learning methodologies to detect depression. As a result, the usage of modes, fusion strategies, model performance and gaps in research are not well known. To fill this gap, the current study will be a systematic literature review of multimodal machine learning methods of automated depression detection. The review summarises the latest advances, compares the performance of multimodal and unimodal systems, and recognises the most important technical, ethical, and methodological issues. In this way, this research paper will seek to offer an organized review of the existing state of the art as well as enlighten future research undertakings in the area of multimodal depressive detection.

### Research Objectives

1. To systematically review existing multimodal machine learning techniques used for automated depression detection, with emphasis on data modalities, feature extraction methods, and model architectures.
2. To analyse and compare the performance of multimodal **versus** unimodal machine learning models based on commonly reported evaluation metrics.
3. To identify key technical challenges, limitations, and research gaps in current multimodal depression detection systems to guide future machine learning research.

### Research Questions

1. What multimodal machine learning approaches have been most commonly employed for depression detection, and which data modalities contribute most effectively to model performance?
2. How does the performance of multimodal machine learning models compare to unimodal models in terms of accuracy, F1-score, and robustness?

## 2. Related Work

Several reviews and surveys have investigated automated depression detection in a number of different aspects. An example is one of the reviews that offered an extensive examination of the speech-based depression determination [4], and the other study conducted an assessment of speech processing methods in detecting depression [2]. In a less specific sense, another paper presented a taxonomy of techniques of multimodal machine learning that can be used in any field [3]. However, none of the systematic literature reviews have specifically synthesized multimodal machine learning methods of identifying depression, data fusion methodology, and data sets and evaluation metrics. The aim of this review is to fill such a gap by conducting a systematic review of current developments (2010-2024) in multimodal depression detection systems, trends, methodological strengths, and existing problems, and thus achieve a direction in future research.

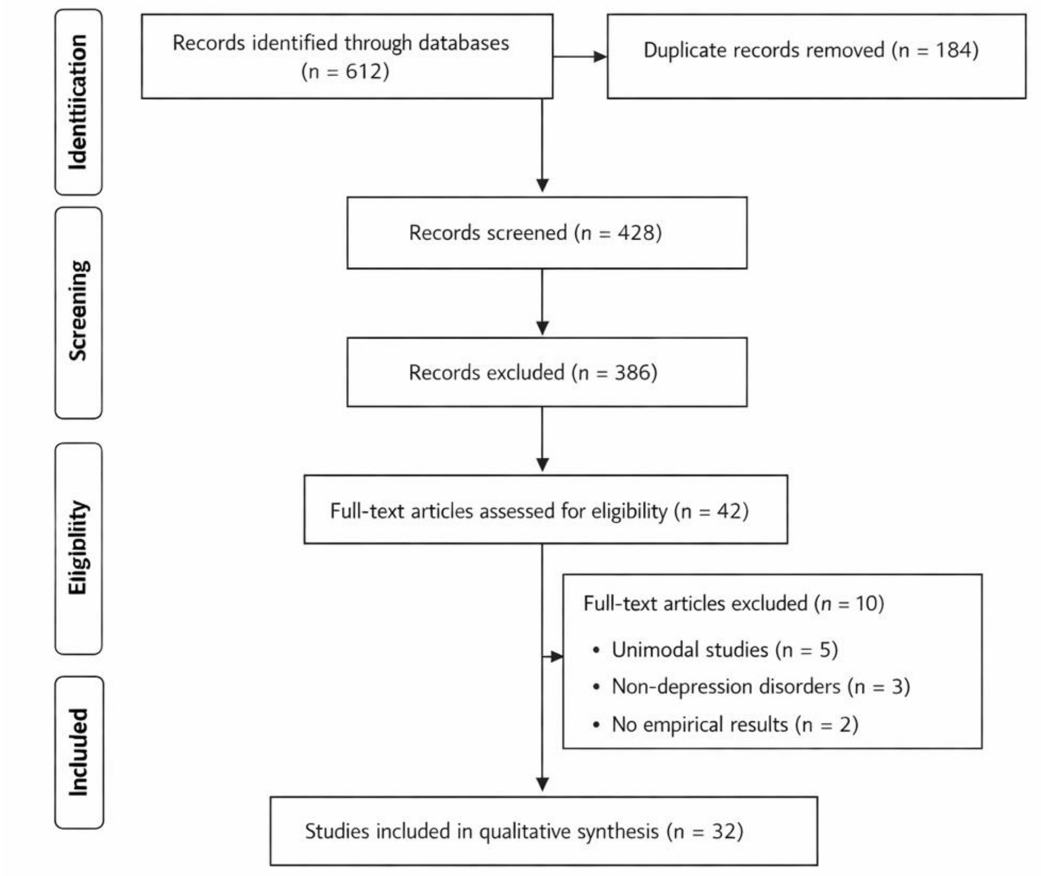
This paper is based on previous studies in that a systematic analysis of multimodal methods, data, and measurement criteria is presented in a computational approach to define performance patterns and gaps in the methodology. The review presents a systematic summary that forms a platform that can be used to develop scalable and effective multimodal depression detection systems.

## 3. Methodology

The research paper adheres to a systematic literature review approach in an attempt to enhance transparency, reproducibility, and coverage of the relevant literature.

### 3.1 Literature Search Strategy

A systematic literature search was conducted across major digital libraries, specifically IEEE Xplore, ACM Digital Library, PubMed, SpringerLink, and ScienceDirect, to identify relevant studies published between 2010 and 2024. There were 612 records initially obtained. Fourty-two full-text articles were evaluated in the eligibility test after the elimination of 184 duplicates and sifting through titles and abstracts. Out of these, 10 of them were filtered off because they were based on either unimodal data, non-depression disorders, or not tested empirically. Finally, a total of 32 studies were incorporated in the synthesis using PRISMA.


Identtiication
Screening
Eligiblity
Included
Records identified through databases (n = 612)
Duplicate records removed (n = 184)
Records screened (n = 428)
Records excluded (n = 386)
Full-text articles assessed for eligibility (n = 42)
Full-text articles excluded (n = 10)
• Unimodal studies (n = 5)
• Non-depression disorders (n = 3)
• No empirical results (n = 2)
Studies included in qualitative synthesis (n = 32)

### 3.2 Inclusion and Exclusion criteria.

The studies used had to utilize at least two data modalities (e.g., speech and text, or text and facial expressions), utilize machine learning (ML) or deep learning (DL) models to predict or estimate the severity of depressions, and at least one quantitative performance measure (ex: accuracy, F1-score, or AUC) had to be reported.

The researches were eliminated when they: Concentrated on unimodal data only and non-depression disorders that were investigated.

### 3.3. Study Selection and Data Extraction

Screening of titles and abstracts was done to determine the relevance and then, full-text review was done. The extracted variables were: data modalities, datasets, modelling approaches, fusion strategies and measurement metrics. It contains a PRISMA-like summary with specific numbers and an explicit connection is made between the extracted variables, the results table, and the questions related to the research. To elaborate on this point, Table 1 will provide a concise overview of modalities, datasets, and model types used in the chosen works in answering RQ1, identifying patterns and trends in multimodal depression detection.

## 4. Multimodal Data Modalities for Depression Detection

Multimodal depression detection Multimodal depression detection systems combine various sources of data to show complementary indicators of depressive behaviour.

### 4.1 Speech and Audio Modalities

Speech-based modalities are widely used in depression detection as vocal characteristics reflect emotional and cognitive states. Commonly extracted features include pitch, energy, speaking rate, spectral features, and pause duration. These acoustic cues are typically obtained from clinical interviews or spontaneous speech recordings and provide information related to affective and psychomotor changes associated with depressive symptoms [4]. Several studies have explored the integration of speech features with additional modalities to capture complementary information [5].

### 4.2 Text and Linguistic Characteristics.

Interpretable information presented in the form of textual data will give useful information on cognitive and emotional trends based on interviews, questionnaires, or social media platforms. Such natural language processing methods as word embeddings and transformers are typically used to supply semantic and sentiment features pertinent to depression detection [6].

### 4.3 Visual and Facial Behaviour

Non-verbal cues of mental health are visual signs of facial expressions, gaze, and movements of the head. Facial behaviour related to depressive states has been analysed using computer vision methods and specifically, convolutional neural networks [7].

### 4.4 Physiological Signals

Objective indicators of emotional and cognitive functioning are the physiological indicators, such as the electroencephalography (EEG), heart rate variability, and galvanic skin response. The most recent research has dealt with the fusion of physiological data and behavioural modalities in order to improve the level of detection [8].

## 5. Machine Learning and Fusion Techniques

### 5.1 Classical Methods of Machine Learning.

Multimodal features that are handcrafted have been applied with traditional machine learning algorithms like support vector machines, random forests, and logistic regression. These models are computationally efficient, but feature engineering can sometimes be a limitation to their performance [9].

### 5.2 Deep Learning Models

The deep learning systems, such as convolutional neural networks, recurrent neural networks, and the long short-term memory networks, automatically learn hierarchical representations using multimodal data. These models have shown to be more effective than the classical models, especially in cases where big data is at hand [10].

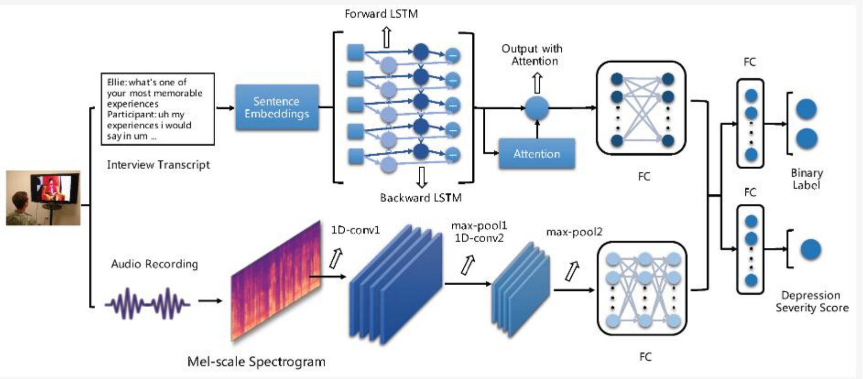


**Figure 1:** A multimodal deep learning architecture for depression detection combining audio (Mel-spectrogram) and text (transcript) features using CNNs, BiLSTMs, and attention mechanisms [2].

### 6. Result Discussion

**Table 1**: Summary of Key Multimodal Depression Detection Studies

| References | Authors & Year | Country | Modalities Used | Dataset | ML/DL Technique | Key Findings |
|---|---|---|---|---|---|---|
| [1] | Cummins et al., 2015 | Australia | Speech | AVEC | SVM, GMM | Speech-based features are effective indicators of depressive symptoms but limited when used alone. |
| [2] | Alhanai et al., 2018 | USA | Audio + Text | DAIC-WOZ | LSTM, RNN | Multimodal fusion significantly outperformed unimodal models in depression detection. |
| [3] | Huang et al., 2019 | China | Audio + Video | AVEC | CNN + LSTM | Temporal modeling of facial and speech features improved prediction accuracy. |
| [4] | Morales et al., 2020 | USA | Text + Audio | Social Media | SVM, Logistic Regression | Linguistic cues combined with acoustic features enhanced classification performance. |
| [5] | Yang et al., 2020 | China | Video + Audio | AVEC | Deep CNN | Visual behavior complements speech for depression assessment. |
| [6] | Zhang et al., 2021 | China | Text + Speech | DAIC-WOZ | BiL STM | Textual sentiment and speech prosody jointly improved F1-score. |

| | | | | | | |
|---|---|---|---|---|---|---|
| [7] | Tzirakis et al., 2021 | UK | Audio + Video + Text | AVEC | Multimodal Deep Network | Tri-modal systems achieved superior robustness over bi-modal systems. |
| [8] | Nurfidausi et al., 2022 | Italy | Speech + Text + EEG | Custom Clinical | CNN + GCN | EEG signals provided strong biological indicators when fused with behavioral data. |
| [9] | Li et al., 2022 | China | Facial + Speech | AVEC | Attention-based CNN | Attention mechanisms improved interpretability and accuracy. |
| [10] | Rahman et al., 2023 | USA | Audio + Text + EEG | DAIC-WOZ | Graph Neural Network | Cross-modal feature relationships improved depression classification. |
| [14] | Ma et al., 2024 | China | Audio + Text + Video | DAIC-WOZ | Transformer-based Multimodal Network | Transformer fusion significantly improved F1-score over LSTM-based multimodal baselines. |
| [15] | Huang et al., 2024 | USA | Text + Audio | DAIC-WOZ | Multimodal BERT + CNN | Cross-modal attention enhanced robustness, especially under noisy speech conditions. |
| [16] | Rohanian et al., 2024 | UK | Audio + Video + Text | AVEC | Multimodal Transformer (MMT) | End-to-end multimodal transformers outperformed feature-engineered models in depression severity estimation. |
| [17] | Li et al., 2025 | China | Speech + EEG + Text | Clinical Dataset | Graph Neural Network (GNN) | Graph-based fusion captured inter-modal relationships, improving accuracy on small datasets. |
| [18] | Sharma et al., 2025 | India | Audio + Text | Social Media + DAIC-WOZ | Hybrid CNN-BiLSTM | Hybrid deep models generalized better across clinical and social-media datasets. |

**Table 1:** Most studies use combinations of audio and text, with DAIC-WOZ and AVEC as dominant datasets, indicating a strong reliance on controlled clinical interviews

**Table 2**: Comparison of Multimodal Fusion Strategies

| Fusion Strategy | Description | Advantages | Limitations |
|---|---|---|---|
| Early Fusion | Combines raw features from all modalities before learning | Captures inter-modal relationships | High dimensionality, noise sensitivity |
| Late Fusion | Combines predictions from individual models | Modality-specific optimization | Limited cross-modal interaction |

| | | | |
|---|---|---|---|
| Hybrid Fusion | Integrates feature-level and decision-level fusion | Balanced performance and flexibility | Higher computational complexity |

**Table 2:** Hybrid fusion strategies, while computationally expensive, tended to achieve the best robustness in tri-modal systems, as seen in tri-modal AVEC studies.

**Table 3:** Commonly Used Datasets and Evaluation Metrics

| Dataset Name | Data Modalities | Description | Commonly Used Evaluation Metrics |
|---|---|---|---|
| DAIC-WOZ | Speech, Text, Video | A widely used clinical interview dataset designed for automated depression assessment. It contains audio recordings, transcribed text, and facial behavior data collected through human–computer interactions. | Accuracy, Precision, Recall, F1-score |
| AVEC (Audio/Visual Emotion Challenge) | Audio, Visual | A benchmark dataset used in multiple emotion and depression recognition challenges. It includes speech signals and facial expressions extracted from recorded interviews. | Accuracy, Precision, Recall, F1-score |
| Social Media Datasets | Text, Behavioural Data | Datasets collected from online platforms such as Twitter, Reddit, and Facebook, containing user-generated text and behavioral indicators reflecting depressive tendencies. | Accuracy, AUC, Precision, Recall |
| EEG-Based Clinical Datasets | EEG Signals, Speech (in some studies) | Clinical datasets capturing brain activity signals along with behavioral data to identify neurophysiological markers of depression. | Accuracy, Sensitivity, Specificity, F1-score |
| Multimodal Interview Datasets | Speech, Text, Facial Expressions | Interview-based datasets combining verbal and non-verbal cues to support multimodal depression classification models. | Accuracy, F1-score, Recall |

**Table 3:** Frequent reliance on DAIC-WOZ and AVEC suggests limited demographic diversity, which raises concerns about generalisability to real-world clinical populations.

As evidenced by the reviewed studies, the multimodal machine learning models are generally superior to unimodal systems in different datasets and evaluation metrics. Multimodal approaches based on deep learning are more accurate, precise, and have greater F1-scores to use, especially in the context of speech, text, and visual data [12]. However, this trend shows important contradictions: although multimodal deep networks generally perform best, classical methods with carefully engineered features (e.g., SVM with prosodic and linguistic features) remain competitive in low-data settings, highlighting the need for fair benchmarking that accounts for varying data conditions [9]

Nonetheless, there are a number of challenges. Most researches have small or unbalanced datasets that limit the generalisability of their findings. There are also ethical and privacy concerns because of collecting sensitive multimodal data that poses barriers to its direct application into practice [13]. Moreover, there are no uniform benchmarks and assessment regimes that make it difficult to compare studies directly.

### 6.1. Synthesis Address Research Question

The literature review shows that the most common suiting systems are the bimodal (audio (speech) and text) systems that utilize the contrasting complementary cues (vocal prosody and semantic content) [5, 10]. Tri-modal audio, video and text systems [7], or with physiological indicators such as EEG [8], are more likely to provide more robust systems and better accuracy and F1-scores. The usefulness of each modality application is dataset-specific in that visual aspects are more effective when conducting structured clinical interviews (e.g., AVEC dataset), but textual data prevails when relying on social media-based detection [6]. These results will answer RQ1 because they will provide the most frequently used and effective modalities and multimodal configurations.

Multimodal models tend to be better than unimodal baselines; one study [5] found an improvement of 15% in F1-scores between a multimodal and an audio-only model on the DAIC-WOZ dataset. Nonetheless, textual data of large size at times can alleviate the necessity of extra modalities, since text-only models can be equally good [2]. In general, multimodal systems are stronger, and they are not affected by noise or information gaps in one of the streams [10, 12], which is why they respond to RQ2 related to the comparative performance, accuracy, F1-score, and reliability.

### 6.2 Ethical and Clinical Considerations

The multimodal systems are characterized by technical potential, their implementation provokes serious ethical and clinical issues, which should be given particular consideration. Such systems demand gathering of sensitive audio, video, and physiological information, which is of critical concern due to the issue of informed consent, data privacy, anonymization, and possible misuse. Moreover, automated detection tools should be used in clinical practice only through clear communication with patients and clinicians about the role of the system, its limitations, and interpretability so that it is not abused and over-used. The review indicates that there is a strong urgency in developing standardized governance structures, ethical auditing processes, and interprofessional development of these technologies in order to make sure that such technologies are designed and implemented in a responsible manner [13].

### 6.3 Limitations of the Review

This systematic literature review offers an intensive synthesis of the present condition of multimodal machine learning to detect depression; a number of intrinsic limitations need to be considered to bring balance to the scholarly perspective and context of the results.

#### *6.3.1 Language and Publication Bias:*

Firstly, the search protocol was thoroughly limited to peer-reviewed articles published in the English language. Although English is the main language used in scientific communication on the global stage, this requirement may have unwittingly marginalized important local research especially in East Asia and Latin America, where scholars can write about their work in their local languages. The results therefore could be due to language bias, which may not be able to capture specific cultural peculiarities in behavioral manifestation of depression or localized data not translated and indexed in the Western databanks.

#### *6.3.2 Search Scope and Database Constraints:*

The search was carried out in five large digital collections, one of which is IEEE Xplore, another one ACM Digital Library, and others are PubMed, SpringerLink, and ScienceDirect, which might have not included all the locations that had relevant "grey literature" (doctoral dissertations, technical white papers in closed AI laboratories and unpublished clinical trial reports). Such sources frequently have valuable negative findings or preliminary innovations which are not necessarily subjected to referee publication in a conventional journal. Omission of this proprietary or non-indexed information can lead to the overly optimistic representation of the performance of a model because published studies tend to give positive or statistically significant results.

#### *6.3.3 Data Heterogeneity and Meta-Analysis Problems*

A primary technical limitation encountered during this review is the high degree of heterogeneity across the included studies. The heterogeneity of data modalities (e.g. different types of EEG signals, audio sample rates and video resolutions) and the lack of similarity in evaluation measures (e.g. F1-score and Mean Absolute Error) rendered a quantitative meta-analysis impossible. Also, the use of a limited set of benchmark datasets, including DAIC-WOZ and AVEC, can result in a dataset bias such that models are very well adapted to a given laboratory condition but possibly fail to extrapolate to actual clinical populations and/or distinct demographic subgroups.

#### *6.3.4 Temporal and Technical Constraints:*

The rapid evolution of the field presents a temporal limitation. Because the review covers the period from 2010 to 2024, the most recent advancements in Large Language Models (LLMs) and Generative AI for mental health assessment are still emerging. With the direction to shift into real-time, wearable-based monitoring, the nature of the reviewed literature can be viewed as fixed, whereas the most recent changes towards the implementation of continuous assessment might not have been exhausted by the reviewed literature. These shortcomings highlight the importance of further studies to target standardized benchmarking and more inclusive and multi-center data gathering in order to confirm the clinical usefulness of these multimodal systems.

## 7. Conclusion

The previous studies have mainly revolved around deep learning models, there being significant issues of data availability, model interpretation and ethical implementation that are yet to be completely addressed. The future directions that the field of research should follow are: (i) creating cross-site, multi-language datasets with a broader demographic representation to evaluate the robustness of models across various cultures and clinical settings; (ii) developing explainable multimodal models, i.e. with the help of such techniques as attention visualization, concept-based explanations, or layer-wise relevance propagation, to better understand clinicians, assist with diagnosis, and be easier to regulate; and (iii) creating standardized benchmarking protocols, i.e. taking into consideration heterogeneous data conditions and The gaps identified need to be addressed so that it is possible to translate the promising multimodal systems developed under research conditions to clinical practice in a responsible manner.

### 7.1 Future Research challenges

Although the combination of multimodal data has greatly contributed to the accuracy of detecting depression, there are still some critical issues that characterize the way forward in the research in the future. In order to scale these systems off of the experimental laboratory settings into practical clinical use, the following areas should be given priority.

#### 7.1.1 Clinical Trust (XAI):

The biggest issue is that modern instances of deep learning like CNNs and BiLSTMs are black box in nature. To trust an AI-based diagnosis, a clinician needs to be able to interpret the evidence given by the system- why a particular vocal prosody and facial muscle movement resulted in a high depression score. Research in the future should focus more on developing Explainable AI (XAI) frameworks. These models can help to develop the required trust in medical professionals and patients by offering visual or textual explanations of why a prediction can be taken, so that AI will serve as a decision-support, though not a substitute, to clinical judgment.

### 7.1.2 Data Diversity and Global Applicability:

The current research relies heavily on a few benchmark datasets which are usually not demographically and culturally diverse. In this regard, it is important to note that in the future, it is necessary to concentrate on designing large, cross-site samples that will represent a diverse set of ethnicities, age groups, and languages. It is critical to overcome dataset bias; a model that was trained on Western speech patterns is unlikely to identify the symptoms in Eastern societies where emotion expression is vastly different.

### 7.1.3 Real-Time Monitoring and Ethical Privacy:

In addition, the transition to continuous, real-time monitoring through wearable devices makes both technical and ethical challenges. Researchers have to come up with algorithms which are computationally efficient to provide services in mobile edge devices without affecting accuracy. At the same time, the industry should work out strict codes of ethics, to preserve sensitive biological and psychological information. These technical and ethical issues should be overcome to implement multimodal detection systems in a sustainable clinical setting that can eventually enhance the speed and precision of all mental health interventions globally.

## References

[1] World Health Organization, Depression and Other Common Mental Disorders: Global Health Estimates, WHO Press, 2017.

[2] S. Morales, J. Levitan, and S. Hirschberg, “Depression detection from speech,” IEEE Signal Processing Magazine, vol. 36, no. 6, pp. 28–38, 2019.

[3] T. Baltrušaitis, C. Ahuja, and L. P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019.

[4] J. Cummins, S. Scherer, B. Schuller, and J. P. Campbell, “A review of depression and suicide risk assessment using speech analysis,” Speech Communication, vol. 71, pp. 10–49, 2015.

[5] M. Alhanai, T. Ghassemi, and J. Glass, “Detecting depression with audio/text sequence modeling of interviews,” Proc. Interspeech, pp. 1716–1720, 2018.

[6] H. Haque, M. Guo, and G. Chen, “Multimodal depression detection using text and speech,” Journal of Biomedical Informatics, vol. 94, pp. 103–118, 2019.

[7] J. Girard and J. Cohn, “Automated audiovisual depression analysis,” Current Opinion in Psychology, vol. 4, pp. 75–79, 2015.

[8] A. Nurfidausi, E. Mancini, and P. Torroni, “Tri-modal depression detection using speech, text, and EEG,” IEEE Access, vol. 11, pp. 44521–44534, 2023.

[9] S. J. Kim and J. Lee, “Feature-based multimodal depression detection,” Expert Systems with Applications, vol. 161, 2020.

[10] Y. Zhang, Z. Liu, and X. Wang, “Deep multimodal learning for depression recognition,” Scientific Reports, vol. 10, no. 1, pp. 1–12, 2020.

[11] L. Yang, D. Jiang, and R. Sahli, “Hybrid fusion strategies for multimodal emotion and depression analysis,” Pattern Recognition Letters, vol. 131, pp. 70–77, 2020.

[12] A. Gupta, R. Sharma, and P. Gupta, “Multimodal deep learning for mental health assessment,” Computers in Biology and Medicine, vol. 132, 2021.

[13] A. V. D. Venkatagiri and P. Narayanan, “Ethical considerations in multimodal mental health AI,” AI & Society, vol. 37, no. 3, pp. 965–977, 2022.

[14] Y. Ma, X. Chen, and Z. Li, “Transformer-based multimodal depression detection using audio, text, and visual cues,” IEEE Access, vol. 12, pp. 45621–45635, 2024.

[15] K. Huang, S. Wang, and J. Glass, “Cross-modal attention for robust depression detection from speech and text,” Journal of Biomedical Informatics, vol. 149, Art. no. 104317, 2024.

[16] M. Rohanian, J. H. Williamson, and S. Scherer, “Multimodal transformer networks for automatic depression assessment,” Pattern Recognition, vol. 146, Art. no. 109955, 2024.

[17] Y. Li, Q. Zhang, and H. Zhou, “Graph neural network-based multimodal depression detection using speech, EEG, and text,” IEEE Transactions on Affective Computing, early access, 2025.

[18] R. Sharma, P. Mehta, and A. Gupta, “Generalizable multimodal deep learning for depression detection across clinical and social media data,” Computers in Biology and Medicine, vol. 176, Art. no. 108566, 2025.

East Journal of Computer Science

# تقنيات التعلّم الآلي متعددة الوسائط لاكتشاف الاكتئاب: مراجعة منهجية للأدبيات العلمية

**محمد عصام عارف [1]* ، محمد ارسياد عبدالله[1]**

[1] علوم الحاسوب، كلية الدراسات العليا، جامعة الإدارة والعلوم، كوالالمبور، ماليزيا

**الملخص:**

*يُعدّ الاكتئاب اضطرابًا واسع الانتشار في مجال الصحة النفسية، ويمثل تحديات اجتماعية واقتصادية وصحية جسيمة على مستوى العالم. وتعتمد أساليب التشخيص التقليدية أساسًا على الاستبيانات ذاتية التقرير والمقابلات السريرية، وهي أساليب تتسم بالذاتية وقد تؤدي إلى تأخر التشخيص أو عدم دقته. وخلال السنوات القليلة الماضية، برزت تقنيات التعلّم الآلي والتعلّم العميق بوصفها أدوات فعّالة للكشف الآلي عن الاكتئاب. وقد حظيت أساليب التعلّم الآلي متعددة الوسائط باهتمام واسع، نظرًا لقدرتها على دمج بيانات غير متجانسة تشمل الكلام، والنصوص، وتعابير الوجه، والمؤشرات الفسيولوجية. تقدّم هذه الدراسة مراجعة منهجية للأدبيات العلمية لتحليل الدراسات الحالية المتعلقة بأساليب التعلّم الآلي متعددة الوسائط في الكشف عن الاكتئاب. ووفقًا لبروتوكول مراجعة مُعدّ مسبقًا، جرى تحديد الدراسات ذات الصلة، وفحصها، وتقييمها بناءً على معايير إدراج واستبعاد محددة. واتباعًا لإرشادات PRISMA ، تم تضمين ٣٢ دراسة في هذا التحليل لضمان الوضوح المنهجي وقابلية إعادة الإنتاج. تلخص المراجعة النتائج المرتبطة بأنماط البيانات، وطرائق النمذجة، وأساليب الدمج، ومجموعات البيانات، ومقاييس التقييم. وأظهرت معظم الدراسات المُراجَعة أن النماذج متعددة الوسائط حسّنت من قيمة مؤشر F1 أو الدقة مقارنة بالنماذج أحادية الوسيط، ولا سيما على مجموعتي البيانات DAIC-WOZ وAVEC؛ غير أن التحسينات كانت أقل وضوحًا أو غير متسقة في مجموعات البيانات الصغيرة أو غير المتوازنة. ومع ذلك، لا تزال تحديات ندرة البيانات، والقضايا الأخلاقية، ومحدودية قابلية التعميم قائمة بشكل ملحوظ. وتقدّم هذه المراجعة تحليلًا معمقًا لأحدث ما توصلت إليه الأبحاث في هذا المجال، وتُبرز الفجوات البحثية الرئيسة بهدف توجيه التطورات المستقبلية في أنظمة الكشف الآلي عن الاكتئاب.*

**الكلمات المفتاحية**: الكشف عن الاكتئاب، التعلّم الآلي متعدد الوسائط، التعلّم العميق، الصحة النفسية، مراجعة منهجية للأدبيات العلمية.