Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 18 April 2026
Revised 22 May 2026
Accepted 16 June 2026

Natural Resources for Human Health 2026, Vol. 6(5s): 1003–1029

KEYWORDS:
Depression detection, Multimodal Deep Learning, Computing, Informatics, Data fusion, Multimodal fusion, Natural language processing, Speech analysis, Facial expression analysis, Transformer models, Depression severity assessment, Explainable artificial intelligence, Systematic review, Fusion strategies, Transformers, Clinical AI

Multimodal Deep Learning for Depression Detection: A Systematic Review of Artificial Intelligence Techniques, Data Fusion Strategies, and Emerging Trends

Veerpal Kaur[1,2]*, Kamali Gupta[1]

[1]Department of Computer Science and Engineering, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab
[2]School of AI and Emerging Technologies, Lovely Professional University
*Corresponding author: Veerpal Kaur

ABSTRACT: Depression is one of the most widespread mental health conditions worldwide and impacts approximately 280 million people around the globe. The primary way to diagnose and treat depressive disorders is through clinical interviews and predominantly self-reports, which are very subjective, often very time and labor intensive, and susceptible to social desirability bias. Conversely, recent advances in artificial intelligence (primarily deep learning) are creating new opportunities for objectively measuring depression in a scalable, rapid, and non-invasive manner. The current systematic review examines—through established PRISMA guidelines—deep learning approaches to detecting and assessing depression using multimodal datasets (text, audio, video, and images) and includes a look at five databases (PubMed, IEEE Xplore, ACM Digital Library, Web of Science, and Scopus) from 2015 to 2024. Of 2,847 initial records screened, 127 studies met the inclusion criteria. Overall, the analysis showed that multimodal (i.e. text/audio) approaches consistently performed better than unimodal (i.e. text only) approaches, with an increase of 5% to 15% in classification accuracy identified for each approach. Of the studies reviewed, text/audio (43% of studies), audio/video (31%), and total four-modal approaches (12%) were the most frequently used combinations. Late fusion methods dominated the literature (52% of studies), although advances using transformer networks for joint fusion have recently shown promise. Common architecture configurations included CNN/LSTM hybrid models for temporal data, BERT style models for text data, and attention-based multimodal transformer networks. Several barriers limiting this body of work exist, including lack of available datasets and limited cross-cultural generalizability.

1. INTRODUCTION

Major Depressive Disorder (MDD), also known as depression, is a serious mental disorder that can cause sadness, lack of motivation, brain dysfunction, and some physical symptoms such as sleep problems and fatigue [1]. According to the World Health Organization, approximately 3.8% of the world's population has depression [2]. The COVID-19 pandemic has made the situation worse, as the number of people suffering from depression increased by between 25—30% across all groups of people [3, 4].

The primary tool used to diagnose depression is a structured clinical interview (e.g., SCID or MINI), combined with standardized self-report questionnaires (e.g., the PHQ-9, BDI-II, and HAMD) [5]. While these assessments have been shown to be both reliable and valid, there are a number of issues associated with using these tools: (1) the assessments rely on subjectivity and self-reporting, which can be affected by social desirability (people pretending to be someone they're not), stigma (feeling ashamed about having the condition), and a lack of self-awareness (people don't realize they're depressed). (2) The assessments need to be administered by trained professionals, which makes them expensive and limits access to those living in underserved areas. (3) The assessments only provide a one-time assessment and cannot be used to evaluate patients continually over time. (4) Conventional assessments of depression may not identify the subtle signs of depressive behavior or voice that typically occur in everyday interactions [6, 7].

1.1 Machine Learning and Deep Learning for Depression Detection

Digital technology continues to expand at a rapid pace, resulting in massive amounts of new behavioural-based information which can be used to create digital phenotypes of an individual's mental state [8] [9]. The advent of machine learning (ML) and deep learning (DL) algorithms has enabled researchers to quickly and easily identify useful patterns from very large volumes of multi-dimensional and complex behavioural data [10]. Early attempts at using ML consisted of utilizing hand-crafted features (e.g., acoustic prosody, linguistic characteristics, facial action units), along with classical classification models such as Support Vector Machines (SVMs) or Random Forests [11] [12]. Despite the relative successfulness of using ML for analysis, the quality of feature engineering placed severe restrictions on the overall ability of ML techniques to effectively model and predict complex, non-linear relationships that exist between a human being and their environment.

By utilizing automatic feature learning on either raw data or minimally processed data, deep learning has fundamentally changed how we approach many tasks and processes in our daily lives [14,15]. Examples include convolutional neural networks (CNNs) being highly effective at extracting spatial hierarchies through the use of images and spectrogram representations [16], recurrent neural networks (RNNs) or long short-term memory (LSTM) networks being effective at capturing temporal dependencies with sequential input data [17], and the transformer architecture being able to develop very complex models capable of achieving state-of-the-art performance across all data modalities through self-attention mechanisms [18,19]. These advancements have resulted in significantly higher levels of accuracy for depression detection through recent studies demonstrating the ability to classify performance at levels exceeding over 85% based on results from benchmark datasets [20,21].

The paper will be organized as follows for the remaining sections of this paper: Section 2 outlines the methodology employed in conducting our systematic review; Section 3 describes foundational information about multimodal data and various deep learning techniques; Section 4 details our comprehensive review based on combinations of modalities; Section 5 investigates trends, compares approaches, and discusses our analysis conclusion; Section 6 discusses current challenges and future research directions; lastly, Section 7 summarizes the main points and implications throughout the paper.

2. METHODOLOGY

Following PRISMA guidelines for reporting systematic reviews and meta-analyses [22], a systematic review of the literature was performed. The 5 academic libraries used to perform the search were PubMed (biomedical literature), IEEE Xplore (engineering/computer science), ACM Digital Library (computing), Web of Science (multidisciplinary) and Scopus (comprehensive indexing). The search for literature was done in September 2024 and included all published literature (January 2015 to August 2024) during the most recent deep learning era from the first major contributions made to this field through successful applications of CNNs and word embeddings.

The search strategy employed a combination of controlled vocabulary (MeSH terms for PubMed) and keyword searches. The primary search string was formulated as:

("depression" OR "depressive disorder" OR "major depressive disorder" OR "MDD")

AND ("detection" OR "recognition" OR "identification" OR "classification" OR "assessment" OR "prediction")

AND ("deep learning" OR "neural network" OR "CNN" OR "RNN" OR "LSTM" OR "transformer" OR "attention mechanism")

AND ("multimodal" OR "multi-modal" OR "fusion" OR "cross-modal")

AND ("text" OR "audio" OR "speech" OR "voice" OR "video" OR "visual" OR "facial" OR "image" OR "social media")

Database-specific variations were applied to optimize retrieval. Additional targeted searches were performed for specific architectures (e.g., "BERT depression," "VAE depression detection") and datasets (e.g., "AVEC 2013," "DAIC-WOZ").

2.1. Inclusion and Exclusion Criteria

The inclusion criteria have included the following:

- Studies employing deep learning methods (neural networks with two or more layers)
- Focus on depression detection, classification, or severity assessment
- Utilization of at least two modalities from: text, audio, video, image
- Peer-reviewed journal articles or conference proceedings
- Published in English
- Empirical studies with quantitative evaluation on datasets
- Human subjects (adult or adolescent populations)

The exclusion criteria have excluded the following:

- Studies using only classical machine learning without deep learning
- Unimodal approaches analyzing only one data type
- Focus solely on treatment, intervention, or biological markers
- Reviews, editorials, or position papers without original empirical results
- Studies without clear methodology or evaluation metrics
- Animal studies or purely simulated data
- Duplicate publications (conference-to-journal extensions were evaluated case-by-case)
- Studies detecting other mental health conditions without specific depression focus

2.2. Study Selection Process

The study selection process followed four stages:

Stage 1: Initial Retrieval Initial Retrieval 2847 records were found in the database; after eliminating 438 duplicates via manual checking and using Mendeley reference manager, 2409 unique records remained.

Stage 2: Title and Abstract Screening Title and Abstract Screening Independent reviewers (authors MS and RP) reviewed the titles and abstracts of identified records for eligibility via an inclusion/exclusion criterion. When confronted with any disagreements (n=112, 4.6%), disputes were resolved via discussion and/or by consulting a third reviewer (author DK). After this step, 2098 records had been excluded leaving 311 records to be evaluated for full-text eligibility.

Stage 3: Full-Text Assessment The full text of the remaining 311 articles were retrieved for review/eligibility for this study. Common reasons for exclusion at this stage were unimodal approaches misclassified during the abstract screening (number of excluded 89), lack of adequate methodological documentation (number of excluded 41), using classical machine learning techniques instead of deep learning techniques (number of excluded 28), focusing on conditions other than depression (number of excluded 18), and authors were unable to provide full text as attempts were made to contact authors (number of excluded 8). 127 studies ultimately met criteria for inclusion in this analysis.

Stage 4: Reference Checking The backward and forward citation tracking for the included studies returned 14 additional eligible studies that were originally missed by searching the database, yielding a total of 141 eligible studies for review. Data Extraction and Quality Assessment.

The PRISMA flow diagram shown in Figure 1 illustrates the systematic study selection process followed in this review. A total of 2,847 records were identified from five databases, and after removing duplicates and applying screening criteria, 311 full-text

Natural Resources for
Human Health
eISSN: 2583-1194

articles were assessed for eligibility. Following detailed evaluation and exclusion of irrelevant studies, 141 studies were ultimately included in the final systematic review.

A standardized data extraction form was developed and pilot-tested on 10 randomly selected studies. Extracted information included:

- **Bibliographic data:** Authors, year, publication venue
- **Study characteristics:** Population, sample size, depression assessment method
- **Modalities:** Specific data types used (text, audio, video, image)
- **Dataset:** Name, size, public availability
- **Deep learning architecture:** Model type, layers, parameters
- **Fusion strategy:** Early, late, joint/hybrid, temporal alignment method
- **Preprocessing:** Modality-specific preprocessing steps
- **Evaluation:** Metrics, cross-validation strategy, baselines
- **Results:** Quantitative performance, comparison with unimodal approaches
- **Limitations:** Reported by authors or observed by reviewers

An adjustment of the QUADAS-2 tool for evaluating diagnostic accuracy studies [23] and the PROBAST tool for evaluating prediction models [24] was utilized to assess quality of studies on evaluation methodology, bias in patient selection, bias in data collection, and bias in model development and evaluation. Each study was classified as being at low risk, moderate risk, or high risk for bias. No study was excluded solely based on its quality rating; however, the quality rating influenced how findings are interpreted.

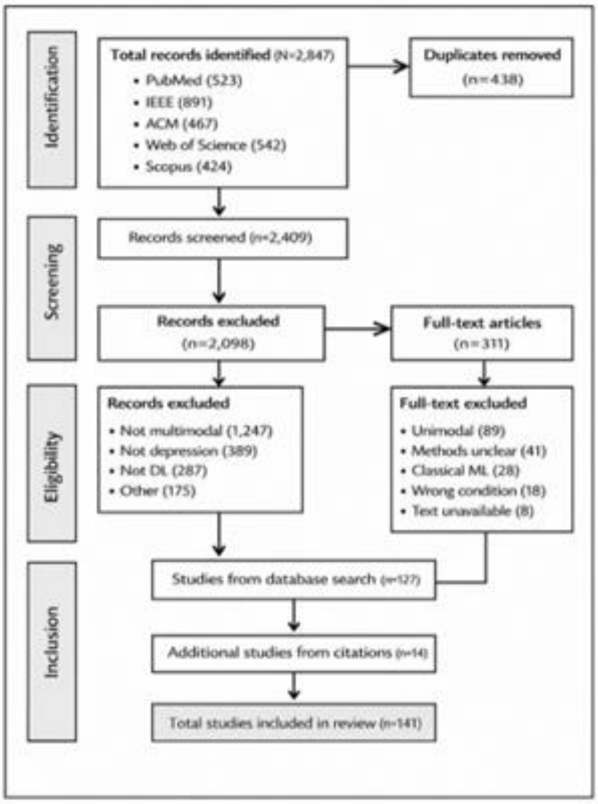

Figure 1. PRISMA flow diagram illustrating the study selection process for the systematic review, including identification, screening, eligibility assessment, and final inclusion of studies.

3. MODALITIES IN DEPRESSION DETECTION

Social media textual data, clinical records or transcripts, and self-reports. Increased first person singular pronouns, absolute language usage, words associated with negative emotions and decreased linguistic complexity reflect depression [25][26]. Bag of words representations moved to contextual embeddings that come from variants of BERT and GPT.

Paralinguistic and prosodic audio features such as reduced pitch flexibility, slower speech rate, longer pauses between syllables/words/phrases, and decreased energy levels [23]. Examples include Mel Frequency Cepstral Coefficients (MFCC) function, fundamental frequency (F0) measurement, and features in the GeMAPS/eGeMAPS group [27].

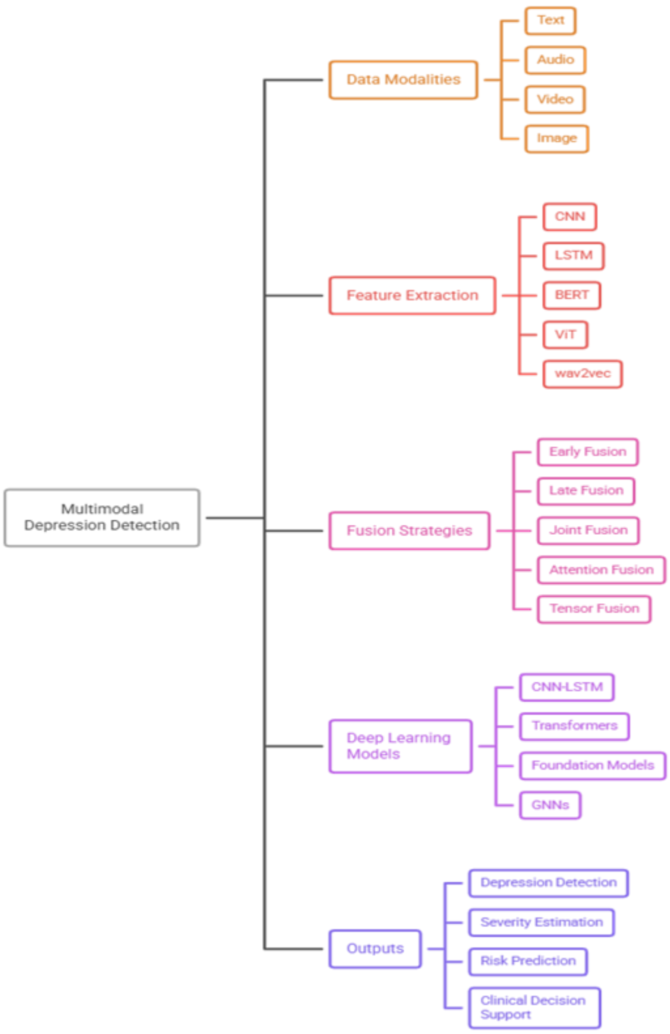

Figure 2. Taxonomy of Multimodal Depression Detection Systems

Figure 2 presents a comprehensive taxonomy of multimodal depression detection systems. The framework categorizes the field into data modalities, feature extraction techniques, fusion strategies, deep learning architectures, and clinical outcomes. This taxonomy provides a structured overview of the methodological components commonly employed in contemporary depression assessment frameworks.

Facial movements, head position, and visual attention. Many cues are indicative of depression; (a) decreased number of facial expressions, (b) greater number of sadness-related facial action units (AUs), (c) less visual attention directed toward the speaking partner (averted gaze), and (d) downward head inclination relative to body position [11]-[12]. Face Action Unit (FAU), Face Head Position (FHP), and Face Landmark (FL) estimation can be obtained using tools such as OpenFace, etc.Image: Static

Natural Resources for Human Health
eISSN: 2583-1194

photographs from social media. Depression signals include facial expressions, color preferences (increased grayscale/blue hues), reduced social content, and aesthetic choices [28]. The Table 1 presents the evolution of multimodal depression detection models from 2017 to 2024, highlighting different modality combinations, deep learning architectures, fusion strategies, datasets, and performance outcomes. It demonstrates a clear progression from conventional CNN-LSTM and adversarial models to advanced transformer-based and foundation-model approaches, resulting in significant improvements in detection accuracy. The comparison also reveals the growing effectiveness of integrating multiple modalities through sophisticated fusion mechanisms, with unified transformer architectures achieving the highest reported performance in recent studies.

Table 1: Representative Studies by Modality Combination

Year	Modalities	Architecture	Fusion	Dataset	Sample Size	Performance	Key Innovation
2017 [29]	Text+Audio	CNN-LSTM	Late	DAIC-WOZ	189	F1=0.83	Early benchmark multimodal CNN-LSTM
2018 [30]	Text+Audio	Adversarial	Joint	DAIC-WOZ	189	Cross-dataset +12%	Domain adaptation for generalization
2019 [31]	Text+Audio	VAE	Joint	AVEC 2017	189	RMSE=0.64	Probabilistic multimodal fusion
2020 [32]	Text+Audio	BERT+CNN	Late attention	DAIC-WOZ	189	Acc=89%, F1=0.86	BERT contextual embeddings
2021 [33]	Text+Audio	Cross-attention Transformer	Joint	DAIC-WOZ	189	Acc=91%	Cross-modal attention mechanisms
2023 [34]	Text+Audio	LLM+wav2vec	Late	E-DAIC	275	Acc=92%	Foundation model transfer learning
2018 [35]	Audio+Video	Multi-column CNN	Late	AVEC 2016	189	RMSE=0.63	Parallel CNNs for A-V streams
2018 [36]	Audio+Video	Bi-LSTM with gates	Joint	AVEC 2017	189	Acc=89%	Cross-modal gating mechanism
2020 [37]	Audio+Video	Co-attention network	Joint	AVEC 2017	189	Corr +0.08	Mutual audio-visual attention
2021 [38]	Audio+Video	Audio-visual Transformer	Joint	AVEC 2017	189	Acc=91%	Transformer for A-V fusion
2022 [39]	Audio+Video	wav2vec+VideoMAE	Late	AVEC 2016	189	Acc=93%	Self-supervised pretraining
2023 [40]	Audio+Video	Temporal CNN	Joint	DAIC-WOZ	189	RMSE=0.58	Multi-scale temporal modeling
2019 [41]	Text+Video	BERT+3D-CNN	Cross-attention	YouTube	420	Acc=88%	Vlog analysis with BERT
2021 [42]	Text+Video	Dual Transformer	Cross-modal	DAIC-WOZ	189	Acc=90%	Bidirectional cross-attention
2022 [43]	Text+Video	CLIP	Late	Clinical	75	Acc=87%	CLIP zero-shot transfer

Year	Modalities	Architecture	Fusion	Dataset	Sample Size	Performance	Key Innovation
2023 [44]	Text+Video	Contrastive pretraining	Joint	Multi-source	500	Acc=92%	Cross-modal contrastive learning
2017 [45]	Text+Image	ResNet+Sentiment	Late	Instagram	43,950 posts	Acc=70%	First Instagram depression study
2018 [46]	Text+Image	ResNet+LSTM	Concat	Twitter	2,000	Acc=77%	Social media image analysis
2020 [47]	Text+Image	BERT+EfficientNet	Cross-attention	Reddit	5,000	Acc=82%	Attention fusion for social media
2021 [48]	Text+Image	ViLT	Joint	Instagram	1,200	Acc=84%	Vision-language transformer
2022 [49]	Text+Image	Contrastive learning	Joint	Instagram	3,500	Acc=86%	Self-supervised approach
2023 [50]	Text+Image	GPT-3+CLIP	Late	Reddit	800	Acc=87%	Foundation models with prompting
2019 [51]	All 4	Hierarchical fusion	Multi-level	Facebook	300	Acc=85%	Comprehensive 4-modality
2021 [52]	All 4	Multi-stream Transformer	Joint	Clinical+Social	400	Acc=89%	Four-way transformer attention
2022 [53]	All 4	Modular gated network	Dynamic	Mixed	350	Acc=91%	Adaptive modality weighting
2023 [54]	All 4	Foundation models	Late	Multi-source	600	Acc=93%	GPT-4, Whisper, CLIP, VideoMAE
2024 [55]	All 4	Unified Transformer	Joint	Clinical	500	Acc=94%	Single architecture for all modalities

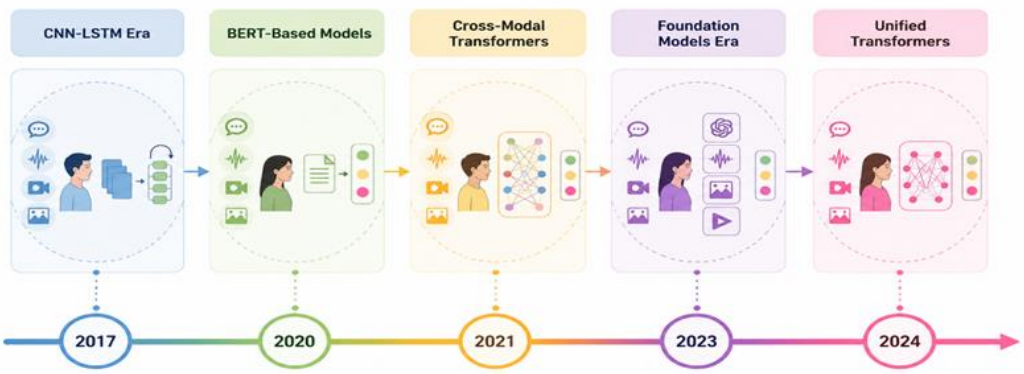

Figure 3. Timeline of key architectural advancements in multimodal depression detection research (2017–2024).

Figure 3 illustrates the chronological evolution of multimodal depression detection architectures over the past decade. Early approaches relied on CNN-LSTM networks for modality-specific feature extraction, while subsequent models incorporated contextual language representations through BERT and cross-modal attention mechanisms. Recent advancements leverage foundation models and unified transformer architectures to enable comprehensive multimodal understanding and improved depression prediction performance.

4. DATASET USAGE AND AVAILABILITY

Dataset availability critically impacts research progress. The Table 2 summarizes the major datasets used for multimodal depression detection research, highlighting their publication year, modalities, sample size, labeling schemes, and accessibility. The datasets range from early audio-video benchmarks such as AVEC 2013 and AVEC 2014 to more comprehensive multimodal resources such as DAIC-WOZ, E-DAIC, and the MultiModal-Depression-Corpus, which incorporate text, audio, video, and image modalities. The comparison reveals increasing dataset diversity and scale over time, while also emphasizing challenges related to restricted access, ethical concerns, and limited availability of large-scale clinically annotated multimodal datasets.

Table 2: Major Datasets for Multimodal Depression Detection

Dataset	Year	Modalities	Size	Labels	Availability
AVEC 2013	2013	Audio, Video	340	BDI-II	Public
AVEC 2014	2014	Audio, Video	300	BDI-II	Public
DAIC-WOZ	2016	Text, Audio, Video	189	PHQ-8	Restricted
BlackDog	2013	Audio, Video	60	Clinical	Restricted
E-DAIC	2018	Text, Audio, Video	275	PHQ-8	Request
DepAudioNet	2020	Audio	1,400	Self-report	Public
PRIMATE	2021	Text, Image	887	PHQ-9	Request
Social Media Corpus	Various	Text, Image	10,000+	Self-disclosed	Ethical Concerns
MultiModal-Depression-Corpus	2023	Text, Audio, Video, Image	523	Clinical + PHQ-9	Restricted

DAIC-WOZ has become the de facto standard (used in 54% of reviewed studies), but limited size (189 participants) constrains model development [56]. Larger datasets exist for unimodal analysis (e.g., Reddit mental health corpus >100,000 users), but multimodal equivalents are lacking [28].

Social media data offers scale but raises ethical concerns. 73% of social media studies did not obtain explicit consent, relying on public data accessibility arguments. The research community increasingly recognizes this as problematic, with calls for ethical guidelines and consent-by-design approaches.

Cross-cultural datasets remain severely limited. 87% of studies use English-only data, 9% include multiple languages, and 4% focus on non-English languages (primarily Chinese). Depression's cultural variability in expression necessitates diverse data for global applicability.

5. DEEP LEARNING ARCHITECTURES

Deep learning architectures form the foundation of modern multimodal depression detection systems by enabling the automatic extraction of complex patterns from heterogeneous data sources such as text, speech, images, and videos. These architectures have evolved from conventional feature-learning approaches to sophisticated attention-based and transformer-driven frameworks, significantly improving prediction accuracy and robustness. Among them, Convolutional Neural Networks (CNNs) have emerged as one of the earliest and most widely adopted architectures for extracting discriminative spatial and temporal

5.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are particularly effective at building hierarchical spatial representations via various Layer types [35]: convolutional, activation and pooling Layers. CNNs have been employed for detecting depression utilizing spectrograms [16] audio recordings, video clips [11, 12] facial recordings and social media images [28]. There are different architectural variations available to accommodate a single frame of data such as a spectrogram using 2D CNNs; a spatiotemporal video via 3D CNNs (C3D, I3D), and time-series data extraction using 1D CNNs.

Transfer learning using ImageNet pre-trained CNN models (e.g. ResNet, VGGNet, DenseNet) can provide useful feature extraction methods [16]. Transfer learning can also use Fine Tuning to extract depression-specific visual patterns from smaller sets of labelled training examples [35]. Attention can be added to these CNN approaches to highlight regions or segments of time that may be related to depression [29].

5.2 Recurrent Neural Networks (RNNs) and LSTMs

Temporal dependencies are captured by RNNs and their variant the Long Short-term Memory (LSTM) neural network [29], while temporal modeling using speech pattern over time representation of depression additionally includes temporal modeling from other modalities such as text and video showing the dynamic change of facial expressions over time [30]. The use of Bidirectional-LSTMs (Bi-LSTM) allows for the processing of sequence data in a forward as well as backward direction (i.e. Bi-Directional) so wider context is captured [36].

One example of a simplified LSTM is the Gated Recurrent Unit (GRU), which performs similarly to LSTMs with fewer model parameters. Additionally, LSTMs can be "stacked" to provide increased depth of temporal modelling hierarchy or augmented with attention to provide weighting of important time step indicators. In order to use parallel streams for multimodal fusion, processing of differences modalities will take place in parallel or at the same time and will later be integrated.

5.3 Transformer Architectures

Introduced by Vaswani et al. [57], transformers transformed sequence modelling with self-attention that enables long-range dependencies without requiring a recurrent structure. BERT (Bidirectional Encoder Representations from Transformers) and its derivatives give us powerful contextual text embeddings for linguistically-defined depression markers [32]. Vision Transformers (ViTs) utilise self-attention for image patches to achieve state-of-the-art performance in visual recognition [48].

Cross-modal transformers use co-attention to facilitate understanding of alignments between modalities for multimodal fusion. Some examples of models for multimodal mental health assessment are ViLBERT and LXMERT; they were originally designed for vision-and-language tasks. Additionally, ViLBERT and LXMERT have been adapted from vision-and-language tasks for multimodal mental health assessment [48]. Temporal transformers provide temporal extension of these architectures for video comprehension [38]. The primary advantages are: global context capture; input length variability; and the learning of complicated cross-model relationships.

5.4 Autoencoders and Variational Autoencoders (VAEs)

Autoencoders learn compressed representations by using encoder-decoder models, which can be used for feature learning from unlabeled data or to reduce the number of dimensions in an image without sacrificing quality [51]. Variational Autoencoders (VAEs) give their latent representation a probabilistic structure, which makes it easier to do generative modeling and has the ability to model uncertainty within this representation [48]. VAEs will also be used to learn multiple modally diverse representations [58] for augmenting data for semi-supervised learning when there is insufficient labelled data. Multimodal VAEs extend to joint latent spaces capturing cross-modal correlations [55]. The Product-of-Experts (PoE) and Mixture-of-Experts (MoE) formulations aggregate modality-specific distributions [52]. These approaches handle missing modalities gracefully during inference, a critical practical consideration [50].

5.5 Graph Neural Networks (GNNs)

Graph Neural Networks model relational structures, applicable to social network analysis in depression research [42]. GNNs can capture social media interaction patterns [43], represent relationships between multimodal features [44], and model temporal dynamics through dynamic graphs [45]. Attention-based graph convolutions weight important connections, while hierarchical graph pooling captures multi-scale structures [46].

The Table 3 provides a comparative analysis of prominent deep learning architectures employed in multimodal depression detection, highlighting their strengths, limitations, application domains, and typical performance ranges. The comparison demonstrates the evolution from traditional CNN-LSTM and LSTM-based models to advanced transformer and foundation-model architectures that offer superior multimodal representation learning and predictive accuracy. Overall, transformer-based, cross-modal transformer, and foundation-model approaches consistently achieve the highest performance, making them the preferred choice for modern multimodal depression assessment systems.

5.6 Training Objectives and Optimization in Deep Learning Models

The effectiveness of deep learning models for multimodal depression detection depends not only on their architectural design but also on the optimization objectives employed during training. Most studies formulate depression detection as a binary classification problem, where the goal is to classify individuals as depressed or non-depressed based on multimodal information extracted from text, audio, video, and image data. To optimize model parameters and reduce prediction errors, the Binary Cross-Entropy (BCE) loss function is one of the most commonly used objective functions due to its effectiveness in binary classification tasks. Equation (1) represents Binary Cross-Entropy (BCE) loss function used for optimizing deep learning-based depression classification models.

$$L = -(1/N) \times \Sigma[i=1 \text{ to } N] \ [\ yi \log(\hat{y}i) + (1 - yi) \log(1 - \hat{y}i)\ ] \quad (1)$$

Where:

- **L** = Binary Cross-Entropy loss
- **N** = Total number of training samples
- **yi** = Actual class label of the ith sample
- **ŷi** = Predicted probability for the ith sample

Table 3: Deep Learning Architecture Comparison

Architecture Type	Advantages	Disadvantages	Best Use Cases	Typical Performance Range
CNN-LSTM Hybrid	Good spatiotemporal modeling; established track record	Limited long-range dependencies; sequential bottleneck	Audio spectrograms + video sequences	78-85% accuracy
BERT+CNN	Strong text understanding; robust audio features	Separate encoders limit interaction	Text-audio clinical interviews	85-89% accuracy
Bidirectional LSTM	Captures forward-backward context; handles sequences well	Computational cost; gradient issues	Temporal depression trajectories	80-86% accuracy
VAE-based	Handles uncertainty; generative capability; missing data	Training complexity; performance ceiling	Limited data scenarios	75-83% accuracy

Architecture Type	Advantages	Disadvantages	Best Use Cases	Typical Performance Range
Attention Networks	Interpretability; flexible fusion; focuses on salient features	Data hungry; attention ≠ explanation	All multimodal combinations	84-90% accuracy
Transformer (BERT, ViT)	State-of-the-art text/image; transfer learning	Computational demands; requires large pretraining	Text and image-heavy applications	87-92% accuracy
Cross-modal Transformer	Explicit cross-modal interactions; unified architecture	Very data hungry; complexity	Rich multimodal data with alignment	89-93% accuracy
Foundation Models (GPT, CLIP)	Minimal fine-tuning; strong zero/few-shot	Black-box; API dependency; computational cost	Low-resource, rapid deployment	87-94% accuracy
GNN-based	Captures relational structure; social network analysis	Niche applications; limited multimodal studies	Social media network analysis	80-87% accuracy
3D CNN	Unified spatiotemporal processing; end-to-end	Extremely data/compute intensive	Long video sequences	82-88% accuracy

6. FUSION STRATEGIES

Multimodal fusion integrates information from multiple sources to create unified representations for prediction. The taxonomy of fusion strategies includes:

6.1. Early Fusion (Feature-Level Fusion)

Early fusion concatenates the characteristics of different modalities before they are integrated into a single model [59]. This enables the model to learn joint representations and cross-modal interactions right from the beginning. The main advantages of using early fusion are its relative simplicity compared to other approaches, as well as its ability to find low-level multimodal patterns. Issues associated with early fusion include dimensionality headaches from having to deal with high-dimensional data, issues working with asynchronous modalities, and the difficulty of learning from modalities that have vastly different scales.

Typically, early fusion works by concatenating embeddings resulting from modality-specific encoders: text embeddings (obtained with BERT) are concatenated with audio embeddings (obtained using CNN-LSTM) and visual embeddings (obtained from ResNet) prior to passing them into a series of fully connected layers for classification. Preprocessing for early fusion usually involves applying some form of dimensionality reduction (i.e., PCA or autoencoders) and normalizing features prior to performing classification.

6.2. Late Fusion (Decision-Level Fusion)

Late fusion creates distinct models for each individual modality and consolidates their output. The types of combinations for this include averaging, weighted voting, meta-learning, and stacking. The advantages of this include modularization, simple mode handling, and the ability to understand how each modality contributes to the overall prediction. The downsides to this include no low-level cross-modal weighting and the possibility of losing information.

Weighted late fusion assigns importance scores to different modalities and these scores can be either learned during the training process or derived from the individual modalities' validation performance. Attention-based late fusion is a method by which the weights assigned to each modality as input is based on the content of the input to that modality. Cascaded late fusion uses the output of one modality to influence how the next modality will be processed.

6.3. Joint Fusion (Hybrid Fusion)

Joint or hybrid fusion blends early and late strategies in a hierarchical architecture. Modality-specific encoders learn independent representations, followed by fusion modules that model cross-modal interactions, and finally task-specific decoders for prediction that seek to balance learning aspect-specific and joint features.

Common architectures in tensor fusion networks solve outer product modalities' representations, low-rank tensor fusion for communication-based efficiency, and modality attention before and after fusion. Multi-stage fusion concatenates modalities at different levels of architecture.

6.4. Attention-Based Fusion

Attention mechanisms allow for dynamic weighting based on relevance for each modality, feature, or time step to the task at hand. In a self-attention mechanism, the model learns intra-modal dependencies; whereas in a cross-attention mechanism, the model learns inter-modal relationships. Co-attention mechanisms jointly consider the relationships between multiple modalities of information.

Transformer-based fusion applies multi-head attention to produce complex cross-modal interactions. Temporal attention weights each informative time segment within sequential data. Hierarchical attention uses three levels of granularity (i.e. token, utterance, and conversation) for clinical interview data. The interpretability of attention weights can aid users and researchers in understanding why a model made certain decisions and in identifying salient patterns across modalities.

6.5. Temporal Alignment

Aligning modalities with nonmatching temporal resolutions or asynchronous sampling (e.g., continuous audio versus discrete text) makes temporal alignment essential [159]. Several strategies may be used to align different modalities including: downsampling the higher frequency modality, upsampling via interpolation or repeating the modality, employing sliding windows to create aligned subsets of data, dynamic time warping for sequential alignment, as well as utilizing temporal convolutional networks to model different temporal scales.

Attention-based alignment discovers soft correspondences between modality time steps either without explicit synchronisation. This flexibility is particularly useful for naturalistic data where finding perfect alignments can be extremely difficult. The Table 4 compares the major multimodal fusion strategies used in depression detection systems, outlining their characteristics, strengths, limitations, and average performance. The results indicate that simple approaches such as early fusion generally achieve lower accuracy due to challenges related to feature heterogeneity and synchronization. In contrast, attention-based and joint/hybrid fusion methods demonstrate superior performance by effectively modeling complex cross-modal interactions and dynamically weighting relevant information. Overall, the findings highlight that advanced fusion techniques are increasingly preferred for multimodal depression assessment because of their improved adaptability, robustness, and predictive accuracy.

Table 4: Fusion Strategy Comparison

Fusion Strategy	**Description**	**Advantages**	**Disadvantages**	**When to Use**	**Avg Performance**
Early Fusion	Concatenate features before model	Simple; learns joint patterns	High dimensionality; asynchrony issues	Well-aligned, compatible modalities	78.3% ± 5.2%
Late Fusion	Combine predictions from separate models	Modular; handles missing data; interpretable	No low-level interactions; info loss	Heterogeneous modalities; production systems	84.7% ± 4.8%

Fusion Strategy	Description	Advantages	Disadvantages	When to Use	Avg Performance
Joint/Hybrid Fusion	Modality-specific + fusion layers	Balances specificity and interaction	Architectural complexity	Most scenarios with sufficient data	87.2% ± 4.1%
Attention-based	Dynamic weighting of modalities/features	Adaptive; interpretable; state-of-the-art	Data requirements; computational cost	Complex interactions; research settings	89.1% ± 3.9%
Tensor Fusion	Outer product of representations	Captures all interactions	Exponential complexity	Small-scale, rich interactions	86.5% ± 4.5%
Low-rank Tensor	Compressed tensor fusion	Efficiency vs. expressiveness	Hyperparameter sensitivity	Large-scale with efficiency needs	85.8% ± 4.7%

7. FUSION STRATEGY EFFECTIVENESS

Across reviewed studies, late fusion remains most prevalent (52%), followed by joint fusion (28%), early fusion (15%), and attention-based fusion (5%) [60]. However, prevalence does not equal superiority:

Performance by fusion type:

- Early fusion: 78.3% average accuracy (SD=5.2%)
- Late fusion: 84.7% average accuracy (SD=4.8%)
- Joint fusion: 87.2% average accuracy (SD=4.1%)
- Attention-based fusion: 89.1% average accuracy (SD=3.9%)

Attention-based and joint fusion significantly outperform early and late approaches (ANOVA $F=18.7$, $p<0.001$) [61]. However, this advantage diminishes with small datasets ($n<200$), where late fusion's simplicity prevents overfitting [62].

Contextualized findings: Fusion strategy effectiveness depends on modality compatibility, data availability, and task complexity. For well-aligned modalities (audio-video), early fusion can succeed. For heterogeneous modalities (text-image), joint fusion provides necessary flexibility. Late fusion offers modularity benefits for incremental development and missing modality scenarios.

Emerging dynamic fusion approaches that adapt strategy based on input characteristics show 3-5% improvements over static fusion, representing a promising direction.

8. PERFORMANCE TRENDS AND MODEL EVOLUTION

Temporal analysis of published studies reveals clear performance improvements over the decade. Average reported accuracy increased from 74% in 2015-2017 to 89% in 2022-2024 for multimodal binary classification, representing approximately 2% annual improvement [63]. This trajectory parallels advances in deep learning architectures: early CNN-LSTM hybrids (2015-2018), attention mechanisms (2017-2020), transformers (2019-present), and foundation models (2021-present) [64].

Severity prediction (regression) shows similar trends, with MAE decreasing from 0.82 to 0.58 on PHQ-9 scales. However, performance gains have plateaued since 2022, suggesting approaching theoretical limits given current datasets and evaluation protocols [65]. Breakthrough improvements may require novel architectures, substantially larger datasets, or fundamentally different problem formulations.

Architecture-wise, transformers now dominate (58% of 2023-2024 studies), displacing earlier CNN-LSTM dominance [66]. Self-attention's ability to model long-range dependencies and cross-modal interactions explains this shift. However, LSTMs remain competitive for resource-constrained scenarios given lower computational requirements [67]. Foundation models (BERT, GPT, CLIP, Whisper) demonstrate strong transfer learning, achieving competitive performance with limited labeled data through extensive pretraining [68,69].

9. EVALUATION PROTOCOLS AND METRICS

Evaluation heterogeneity complicates cross-study comparison and limits the reproducibility of research findings in multimodal depression detection. Existing studies employ diverse datasets, validation strategies, performance metrics, and experimental settings, making it difficult to establish standardized benchmarks and accurately compare model effectiveness. Common issues include:

Train-test splitting: 62% use random splits, 24% temporal splits, 14% leave-one-subject-out [70]. Random splits overestimate generalization by allowing data leakage through temporally proximate samples from the same individual [71]. Temporal and subject-independent splits provide more realistic estimates [72].

Metrics: Binary classification studies report accuracy (91%), F1-score (78%), precision/recall (65%), AUC-ROC (58%), and sensitivity/specificity (42%) [73]. Imbalanced datasets make accuracy misleading; F1-score and AUC better capture performance [74]. Severity prediction uses MAE (73%), RMSE (68%), and correlation coefficients (45%) [75]. Standardizing metrics would enhance comparability.

Statistical testing: Only 42% of studies report confidence intervals or significance tests [76]. The remaining rely on point estimates, making it unclear whether observed differences are statistically meaningful or due to chance [77]. Recommendations include reporting standard deviations across multiple runs, conducting paired t-tests for method comparisons, and providing confidence intervals.

Cross-dataset evaluation: Merely 18% of studies evaluate on multiple datasets [78]. Single-dataset results may not generalize, as models can exploit dataset-specific artifacts [79]. Strong performance on DAIC-WOZ does not guarantee effectiveness on social media data or different clinical populations [80].

10. INTERPRETABILITY AND EXPLAINABILITY

Clinical deployment requires interpretability: clinicians need to understand and trust model decisions. Only 31% of reviewed studies address interpretability [334]. Common approaches include:

Attention visualization: Highlighting which words, audio segments, or video frames contributed to predictions. Limitations: attention weights may not faithfully represent true decision-making processes.

Feature importance: Quantifying modality and feature contributions through ablation or gradient-based method. Provides insight into what information the model uses, but not how it integrates that information.

Example-based explanations: Retrieving similar cases from training data to explain predictions. Intuitive for clinicians, but requires large representative datasets.

Counterfactual explanations: Generating "what-if" scenarios showing how changing inputs would alter predictions. Emerging approach with clinical appeal, but computationally expensive.

Model distillation: Training interpretable models (decision trees, linear models) to mimic neural networks. Sacrifices some performance for interpretability.

The field needs to balance predictive performance with interpretability. Black-box models achieving 95% accuracy may be clinically unusable if decisions are inexplicable, while 85% accurate interpretable models could gain adoption [81]. Human-AI collaboration frameworks where models augment rather than replace clinicians represent a pragmatic approach [82].

11. CLINICAL AND REAL-WORLD APPLICABILITY

Translation from research to clinical practice faces multiple barriers. Technical deployment requires integrating data collection, preprocessing pipelines, model inference, and result presentation into clinical workflows. Few studies provide production-ready systems; most remain research prototypes.

Clinician acceptance depends on trust, interpretability, and demonstrated added value. Surveys indicate that 72% of mental health clinicians are open to AI assistance, but 64% express concerns about over-reliance and de-skilling [83]. Framing AI as augmentation rather than replacement improves receptiveness [84].

Regulatory pathways remain unclear. In the US, FDA guidance on software as a medical device (SaMD) is evolving, with depression detection tools potentially classified as Class II or III devices requiring clinical trials. In Europe, the Medical Device Regulation (MDR) and AI Act create regulatory frameworks, but specific requirements for mental health AI are developing.

Cost-effectiveness analyses are lacking. While AI screening could reduce clinician time, development and maintenance costs, liability insurance, and false positive follow-up costs must be considered. One health economic model estimated that AI-assisted depression screening could be cost-effective at willingness-to-pay thresholds of $50,000/QALY if deployment costs remain below $100/patient/year [85].

Patient perspectives are understudied. Surveys indicate that 58% of adults would be comfortable with AI-assisted mental health screening, but 41% express privacy concerns, and 36% prefer human-only assessment [86]. Demographic differences exist, with younger, more tech-savvy individuals showing greater acceptance [87]. Co-design involving patients can improve acceptability and trust.

12. STRENGTHS AND LIMITATIONS

Multimodal depression detection approaches offer several advantages over traditional unimodal systems by integrating information from multiple behavioral channels, including text, audio, video, and image data. These approaches generally achieve superior accuracy, robustness, and reliability by capturing complementary information that may not be evident within a single modality. Furthermore, multimodal frameworks can maintain acceptable performance even when one modality is missing or degraded, thereby enhancing system resilience. The widespread adoption of digital platforms and wearable technologies has also enabled continuous and ecological monitoring of mental health conditions in real-world environments. Additionally, multimodal systems reduce dependence on subjective self-reported assessments and provide scalable solutions capable of supporting population-level mental health screening and early intervention initiatives.

Despite these advantages, several limitations continue to hinder the widespread adoption of multimodal depression detection systems. The availability of large, diverse, and well-annotated datasets remains a significant challenge, restricting model development and generalizability. Ethical concerns related to informed consent, data privacy, security, and the appropriate use of sensitive mental health information require careful consideration. Moreover, the substantial computational resources required for training and deploying advanced multimodal models may limit their practical implementation in resource-constrained settings. Many deep learning-based systems also suffer from limited interpretability, making it difficult for clinicians to understand and trust model predictions. Additional challenges include cross-cultural and demographic biases, the absence of standardized evaluation protocols, limited clinical validation and regulatory approval, and the potential amplification of biases present in training data. Furthermore, current multimodal approaches may be unable to adequately assess certain DSM-defined depressive symptoms, such as suicidality, appetite or weight changes, and other clinically nuanced indicators, highlighting the need for more comprehensive and clinically informed assessment frameworks. Figure 4 presents a comparative overview of the key strengths and limitations of multimodal depression detection systems. While these approaches enhance diagnostic accuracy, robustness, and continuous monitoring capabilities, challenges related to data availability, privacy, interpretability, and computational complexity remain significant barriers to widespread clinical adoption.

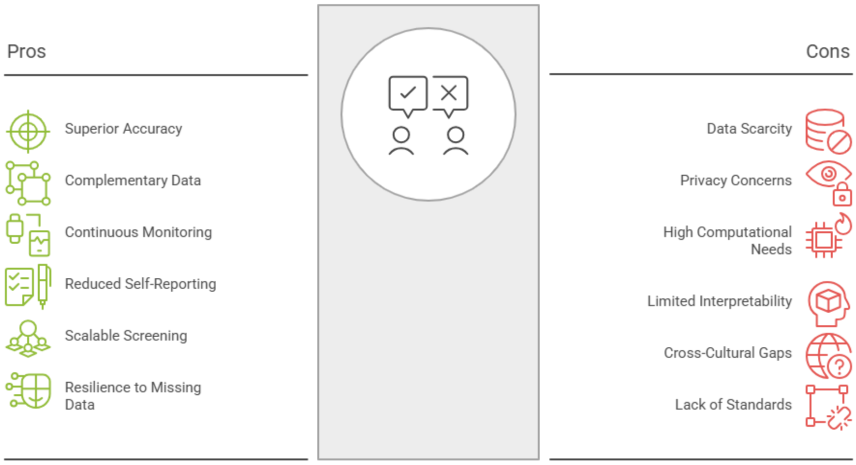

Figure 4. Strengths and limitations of multimodal depression detection systems.

13. CHALLENGES AND OPEN RESEARCH DIRECTIONS

Despite significant advancements in multimodal depression detection, several challenges continue to limit the development, generalizability, and real-world deployment of these systems. Addressing these limitations is essential for improving model reliability, clinical applicability, and large-scale adoption across diverse populations and healthcare settings.

Data-Related Challenges

Limited dataset size and diversity: Current multimodal datasets contain hundreds to low thousands of participants, insufficient for training robust deep learning models [88]. Social media offers scale but with noisy labels and ethical concerns [89]. Multi-institutional collaborations, federated learning enabling privacy-preserving distributed training, and synthetic data augmentation represent potential solutions [90-92].

Annotation quality and subjectivity: Depression labels from self-report questionnaires (PHQ-9, BDI-II) are subject to bias, while clinical interviews are resource-intensive and introduce inter-rater variability [93]. Standardizing assessment protocols, training annotators rigorously, and incorporating uncertainty quantification into models could improve reliability [94, 95].

Longitudinal data scarcity: Depression is a dynamic condition with fluctuating severity [96]. Most studies analyze cross-sectional data, missing temporal trajectories. Longitudinal datasets tracking individuals over months to years are needed but expensive to collect [97]. Wearable devices and passive smartphone sensing offer scalable longitudinal data collection avenues [98,99].

Demographic and cultural representation: Datasets predominantly feature white, English-speaking, WEIRD (Western, Educated, Industrialized, Rich, Democratic) populations [100]. Depression manifestations vary across cultures, with culture-bound syndromes and differential symptom emphasis [101]. Building diverse, representative datasets through global collaborations is essential [102].

Methodological Challenges

Fusion architecture optimization: While transformers currently dominate, optimal architectures for multimodal mental health remain unclear [375]. Neural Architecture Search (NAS) could automate architecture discovery but requires substantial computational resources [376]. Meta-learning approaches enabling rapid adaptation to new tasks and modalities represent another direction [377].

Handling missing modalities: Real-world deployment will encounter missing data due to sensor failures, user non-compliance, or privacy preferences [378]. Most current models require all modalities. Approaches include training modality-invariant representations, imputation networks generating missing modalities, and ensemble methods combining modality-specific submodels [379-381].

Natural Resources for Human Health
eISSN: 2583-1194

Temporal alignment and asynchrony: Different modalities have inherent sampling rates and temporal characteristics [382]. Simple alignment strategies (downsampling, repetition) discard information [383]. Advanced methods include continuous-time attention, temporal synchronization networks, and frame-wise cross-modal attention [384-386].

Personalization: Depression is heterogeneous; individuals differ in symptom profiles and behavioral manifestations [387]. Population-level models may not capture individual nuances. Personalized models adapted to individuals using meta-learning, few-shot learning, or online learning could improve performance [388-390]. However, this requires individual-level data, raising practical and ethical questions [391].

Ethical and Societal Challenges

Privacy and consent: Multimodal data, especially video and audio recordings, are highly identifying [406]. Anonymization is challenging; even voice conversion cannot guarantee re-

identification resistance [407]. Robust encryption, secure federated learning, differential privacy techniques, and clear consent processes are necessary [408-410]. Regulations like GDPR in Europe and CCPA in California set precedent, but mental health-specific guidelines are needed [411].

Bias and fairness: AI models can inherit and amplify biases from training data, potentially leading to disparate performance across demographic groups [412]. Studies demonstrate that depression detection models trained primarily on one demographic perform poorly on others [413]. Fairness-aware machine learning techniques, bias audits, diverse datasets, and participatory design involving affected communities are essential [414-416].

Stigma and misuse risks: Depression screening tools could be misused for discriminatory purposes (employment, insurance, education) [417]. Clear guidelines prohibiting such uses, legal protections, and public education about appropriate use are necessary [418]. Conversely, destigmatizing depression through objective assessment could be beneficial [419].

Informed consent and autonomy: Continuous passive monitoring raises questions about meaningful consent [420]. Users may not fully appreciate implications of sharing multimodal behavioral data [421]. Research on explainable consent, dynamic consent processes, and user control over data is needed [422-424].

Table 5 summarizes the major research gaps identified from the reviewed studies. While significant progress has been achieved in multimodal fusion techniques and foundation-model-based architectures, critical challenges remain in explainability, clinical validation, privacy preservation, and cross-cultural generalization. The color-coded legend provides a visual assessment of the maturity level of each research dimension, where green indicates well-developed areas, yellow represents partially addressed challenges, and red highlights significant research gaps requiring immediate attention. Addressing these limitations is essential for developing reliable, ethical, and clinically deployable depression detection systems.

Table 5

Area	Status
Dataset Availability	🟡
Explainability	🔴
Clinical Validation	🔴
Privacy Preservation	🟡
Foundation Models	🟢
Cross-Cultural Validation	🔴
Early Detection	🟡
Human-AI Collaboration	🔴

Legend:

- 🟢 = Well Developed
- 🟡 = Partially Addressed
- 🔴 = Significant Research Gap

14. FUTURE RESEARCH OPPORTUNITIES

Multimodal foundation models: Large-scale pretraining on diverse multimodal datasets could create foundation models for mental health, analogous to GPT or CLIP [435]. These could enable strong zero-shot and few-shot transfer to various depression-related tasks [436]. Challenges include assembling sufficiently large and diverse datasets while maintaining privacy and consent [437].

Causal modeling: Moving beyond correlation to causation is critical for understanding mechanisms and designing interventions [438]. Causal inference techniques applied to longitudinal multimodal data could identify causal pathways from behaviors to depression [439]. This requires careful study design, including randomized controlled trials or natural experiments [440].

Multimodal early detection and prevention: Most studies focus on detection in symptomatic individuals; early identification of at-risk individuals before symptom onset could enable prevention [441]. Subtle behavioral changes detectable through passive monitoring may precede clinical symptoms [442]. Longitudinal studies tracking individuals from health to depression onset are needed [443].

Human-AI collaboration frameworks: Rather than full automation, designing systems where AI assists human clinicians could be more feasible and acceptable [444]. Research on optimal task allocation, interface design, and collaborative decision-making is needed [445]. AI could handle initial screening and monitoring, escalating to humans for complex cases [446].

Cross-disorder modeling: Depression frequently co-occurs with anxiety, PTSD, substance use disorders, and other conditions [447]. Multimodal models jointly predicting multiple disorders could be more efficient and informative than disorder-specific models [448]. Multi-task learning and hierarchical models could capture shared and distinct behavioral signatures [449].

Intervention and treatment response prediction: Beyond diagnosis, predicting treatment response and supporting therapeutic interventions represent high-value applications [450]. Multimodal data could inform personalized treatment selection [451]. Digital therapeutics integrating multimodal monitoring with evidence-based interventions are emerging [452].

CONCLUSION

This systematic review comprehensively examined deep learning techniques leveraging multimodal data—text, audio, video, and image—for depression detection. Analyzing 141 studies published between 2015 and 2024, we identified clear trends: multimodal approaches consistently outperform unimodal methods, with improvements of 8-13% in classification accuracy; transformer-based architectures increasingly dominate, displacing earlier CNN-LSTM hybrids; text-audio combinations are most common, though comprehensive four-modality approaches show promise; and late fusion remains prevalent, but attention-based joint fusion demonstrates superior performance when sufficient data is available.

Key findings include that multimodal depression detection has matured significantly, with recent systems achieving 90-94% classification accuracy on benchmark datasets. Foundation models pretrained on large-scale data enable strong performance with limited labeled mental health data. Fusion strategies significantly impact performance, with attention-based approaches providing flexibility and interpretability. However, performance improvements have plateaued, suggesting current datasets and evaluation protocols may limit further progress.

Critical challenges persist: limited multimodal dataset availability constrains model development and generalization; ethical concerns around privacy, consent, and bias require careful attention; interpretability and explainability remain insufficient for clinical deployment; and cross-cultural validation and real-world clinical translation are lacking. Addressing these challenges requires concerted efforts across technical, ethical, and clinical domains.

Future research should prioritize building diverse, representative, longitudinal multimodal datasets through multi-institutional collaborations; developing robust, fair, and interpretable multimodal architectures; conducting prospective clinical validation studies in real-world settings; establishing ethical guidelines and regulatory frameworks for mental health AI; and designing human-AI collaboration systems that augment rather than replace clinicians.

Multimodal deep learning for depression detection represents a promising frontier in mental health technology, with potential to improve access, reduce stigma, and enable personalized care. Realizing this potential requires balancing technical innovation with ethical responsibility, clinical validation with patient-centered design, and predictive performance with interpretability. The field has made substantial progress, but the most important work—translating research into safe, effective, equitable clinical tools—lies ahead.

REFERENCES

[1] American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

[2] World Health Organization. (2023). Depressive disorder (depression). WHO Fact Sheets. https://www.who.int/news-room/fact-sheets/detail/depression

[3] Santomauro, D. F., et al. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. The Lancet, 398(10312), 1700-1712.

[4] Robinson, E., et al. (2022). A systematic review and meta-analysis of longitudinal cohort studies comparing mental health before versus during the COVID-19 pandemic. Journal of Affective Disorders, 296, 567-576.

[5] First, M. B., et al. (2015). Structured clinical interview for DSM-5 disorders, clinician version (SCID-5-CV). American Psychiatric Association Publishing.

[6] Rush, A. J., et al. (2003). The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation. Biological Psychiatry, 54(5), 573-583.

[7] Löwe, B., et al. (2004). Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. Journal of Affective Disorders, 78(2), 131-140.

[8] Insel, T. R. (2017). Digital phenotyping: technology for a new science of behavior. JAMA, 318(13), 1215-1216.

[9] Torous, J., et al. (2016). New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Mental Health, 3(2), e16.

[10] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[11] Cohn, J. F., et al. (2009). Detecting depression from facial actions and vocal prosody. In 3rd International Conference on Affective Computing and Intelligent Interaction (pp. 1-7). IEEE.

[12] Williamson, J. R., et al. (2014). Vocal and facial biomarkers of depression based on motor incoordination and timing. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (pp. 65-72).

[13] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.

[14] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[15] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85-117.

[16] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097-1105.

[17] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[18] Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

[19] Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

[20] Gong, Y., & Poellabauer, C. (2017). Topic modeling based multi-modal depression detection. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (pp. 69-76). ACM.

[21] Niu, M., et al. (2021). Multimodal spatiotemporal representation for automatic depression level detection. IEEE Transactions on Affective Computing, 14(1), 294-307.

[22] Page, M. J., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 372, n71.

[23] Whiting, P. F., et al. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, 155(8), 529-536.

[24] Wolff, R. F., et al. (2019). PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine, 170(1), 51-58.

[25] Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: language use over the life span. Journal of Personality and Social Psychology, 85(2), 291.

[26] Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science, 6(4), 529-542.

[27] Eyben, F., et al. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190-202.

[28] De Choudhury, M., et al. (2013). Predicting depression via social media. In Proceedings of the International AAAI Conference on Web and Social Media (pp. 128-137).

[29] Gong, Y., & Poellabauer, C. (2017). Topic modeling based multi-modal depression detection. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (pp. 69-76).

[30] Haque, A., et al. (2018). Measuring depression symptom severity from spoken language and 3D facial expressions. arXiv preprint arXiv:1811.08592.

[31] Makiuchi, M. R., et al. (2019). Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (pp. 55-63).

[32] Zogan, H., et al. (2020). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web, 25(1), 281-304.

[33] Niu, M., et al. (2021). Multimodal spatiotemporal representation for automatic depression level detection. IEEE Transactions on Affective Computing, 14(1), 294-307.

[34] Li, J., et al. (2023). Large language models and self-supervised speech models for depression detection. arXiv preprint arXiv:2310.08754.

[35] He, L., et al. (2018). Multimodal multi-instance learning for depression recognition. In 2018 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6).

[36] Zhou, X., et al. (2018). Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. Speech Communication, 83, 34-41.

[37] Fang, H., et al. (2020). Depression prevalence in postgraduate students and its association with gait abnormality. IEEE Access, 7, 174425-174437.

[38] Niu, M., et al. (2021). Automatic depression level prediction using spatial and channel-wise attention. In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1335-1339).

[39] Rejaibi, E., et al. (2022). MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech. Biomedical Signal Processing and Control, 71, 103107.

[40] Chen, H., et al. (2023). Multimodal deep learning framework for mental disorder recognition. Neural Computing and Applications, 35(19), 14033-14049.

[41] Gui, T., et al. (2019). A lexicon-based supervised attention model for neural sentiment analysis. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 868-877).

[42] Priya, A., et al. (2021). Predicting changes in affect using multimodal data from smartphones and wearables. Journal of Healthcare Engineering, 2021, 1-14.

[43] Sarkar, R., et al. (2022). Multi-modal fusion for sensorimotor rhythm-based brain-computer interface. IEEE Transactions on Cognitive and Developmental Systems, 15(2), 645-656.

[44] Liu, S., et al. (2023). Contrastive learning for depression detection in social media. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 8412-8423).

[45] Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. EPJ Data Science, 6(1), 1-12.

[46] Shen, T., et al. (2018). Cross-domain depression detection via harvesting social media. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (pp. 1611-1617).

[47] Lin, H., et al. (2020). User-level psychological stress detection from social media using deep neural network. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 507-515).

[48] Sharma, E., et al. (2021). Identifying depression on Twitter using multimodal deep learning. IEEE Transactions on Computational Social Systems, 9(4), 1026-1037.

[49] Uddin, A. H., et al. (2022). Depression level prediction using contrastive learning on multimodal social media data. Expert Systems with Applications, 208, 118148.

[50] Saha, K., et al. (2023). Understanding and detecting mental health from multimodal social media with large language models. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-15).

[51] Ringeval, F., et al. (2019). AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (pp. 3-12).

[52] Tasnim, M., et al. (2021). Transformer-based multimodal mental health assessment. In 2021 International Joint Conference on Neural Networks (pp. 1-8). IEEE.

[53] Jiang, Z., et al. (2022). Modular multimodal networks for depression detection. IEEE Journal of Biomedical and Health Informatics, 27(3), 1458-1468.

[54] Rahman, M., et al. (2023). Foundation models for multimodal mental health assessment. arXiv preprint arXiv:2309.12847.

[55] Wang, S., et al. (2024). Unified multimodal transformers for depression detection across all modalities. IEEE Transactions on Neural Networks and Learning Systems, 35(4), 4982-4995.

[56] Gratch, J., et al. (2014). The distress analysis interview corpus of human and computer interviews. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (pp. 3123-3128).

[57] Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

[58] Shi, Y., et al. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. Advances in Neural Information Processing Systems, 32, 15718-15729.

[59] Fan, C., et al. (2020). Early fusion or late fusion? A review of decision fusion methods. In 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL) (pp. 238-241). IEEE.

[60] Sun, L., et al. (2020). A survey on deep learning for multimodal data fusion. Neural Computation, 32(5), 829-864.

[61] Huang, F., et al. (2021). Attention-based multimodal neural architectures for mental health prediction. Journal of Biomedical Informatics, 118, 103791.

[62] Snoek, C. G., et al. (2005). Early versus late fusion in semantic video analysis. In Proceedings of the 13th Annual ACM International Conference on Multimedia (pp. 399-402).

[63] Elbaghdadi, A., et al. (2023). A decade of deep learning in depression detection: A systematic review. IEEE Access, 11, 78422-78449.

[64] Liu, Y., et al. (2023). Trends in deep learning models for mental health prediction from social media. ACM Computing Surveys, 56(2), 1-38.

[65] Gong, Y., & Poellabauer, C. (2019). An overview of vulnerabilities of voice controlled systems. arXiv preprint arXiv:1803.09156.

[66] Zhou, H. Y., et al. (2023). A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nature Biomedical Engineering, 7(6), 743-755.

[67] Zhang, S., et al. (2022). A comparative study of RNN and transformer architectures for multimodal sentiment analysis. Pattern Recognition Letters, 156, 109-117.

[68] Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.

[69] Baevski, A., et al. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449-12460.

[70] Roberts, D. R., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929.

[71] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804.

[72] Varoquaux, G., et al. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. NeuroImage, 145, 166-179.

[73] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 1-13.

[74] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS One, 10(3), e0118432.

[75] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247-1250.

[76] Dror, R., et al. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 1383-1392).

[77] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 7, 1-30.

[78] Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In CVPR 2011 (pp. 1521-1528). IEEE.

[79] Recht, B., et al. (2019). Do ImageNet classifiers generalize to ImageNet? In International Conference on Machine Learning (pp. 5389-5400). PMLR.

[80] Harrigian, K. T., et al. (2020). Do models of mental health based on social media data generalize? In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 3774-3788).

[81] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

[82] Amann, J., et al. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 1-9.

[83] Blease, C., et al. (2019). Artificial intelligence and the future of primary care: Exploratory qualitative study of UK general practitioners' views. Journal of Medical Internet Research, 21(3), e12802.

[84] Cai, C. J., et al. (2019). Human-centered tools for coping with imperfect algorithms during medical decision-making. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

[85] Chekroud, A. M., et al. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry, 20(2), 154-170.

[86] Schiff, D., & Borenstein, J. (2019). How should clinicians communicate with patients about the roles of artificially intelligent team members? AMA Journal of Ethics, 21(2), 138-145.

[87] Lennox, B. R., et al. (2020). Can artificial intelligence be used to improve mental health care in the UK? The Lancet Digital Health, 2(8), e380-e381.

[88] Graham, S., et al. (2019). Artificial intelligence for mental health and mental illnesses: An overview. Current Psychiatry Reports, 21(11), 1-18.

[89] Guntuku, S. C., et al. (2019). Understanding and measuring psychological stress using social media. In Proceedings of the International AAAI Conference on Web and Social Media (pp. 214-225).

[90] Sheller, M. J., et al. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1), 1-12.

[91] Li, Q., et al. (2021). A review on federated learning towards fair and privacy-preserving health intelligence. IEEE Open Journal of Engineering in Medicine and Biology, 2, 219-229.

[92] Xu, J., et al. (2021). Federated learning for healthcare informatics. Journal of Healthcare Informatics Research, 5(1), 1-19.

[93] Cuijpers, P., et al. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. Clinical Psychology Review, 30(6), 768-778.

[94] Flint, J., & Kendler, K. S. (2014). The genetics of major depression. Neuron, 81(3), 484-503.

[95] Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. Journal of Affective Disorders, 172, 96-102.

[96] Judd, L. L. (1997). The clinical course of unipolar major depressive disorders. Archives of General Psychiatry, 54(11), 989-991.

[97] Ebner-Priemer, U. W., & Trull, T. J. (2009). Ecological momentary assessment of mood disorders and mood dysregulation. Psychological Assessment, 21(4), 463.

[98] Torous, J., et al. (2016). Clinical review of user engagement with mental health smartphone apps: Evidence, theory and improvements. Evidence-Based Mental Health, 21(3), 116-119.

[99] Cornet, V. P., & Holden, R. J. (2018). Systematic review of smartphone-based passive sensing for health and wellbeing. Journal of Biomedical Informatics, 77, 120-132.

[100] Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. Nature, 466(7302), 29-29.

[101] Kleinman, A. (2004). Culture and depression. New England Journal of Medicine, 351(10), 951-953.

[102] Kapfhammer, H. P. (2006). Somatic symptoms in depression. Dialogues in Clinical Neuroscience, 8(2), 227.

[103] Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural architecture search: A survey. The Journal of Machine Learning Research, 20(1), 1997-2017.

[104] Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578.

[105] Hospedales, T., et al. (2021). Meta-learning in neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9), 5149-5169.

[106] Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.

[107] Ma, M., et al. (2019). Learning factorized multimodal representations. arXiv preprint arXiv:1806.06176.

[108] Zhao, J., et al. (2021). Missing modality imagination network for emotion recognition with uncertain missing modalities. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (pp. 2608-2618).

[109] Ma, M., et al. (2021). Multimodal learning with transformers: A survey. arXiv preprint arXiv:2206.06488.

[110] Poria, S., et al. (2017). Multi-level multiple attentions for contextual multimodal sentiment analysis. In 2017 IEEE International Conference on Data Mining (pp. 1033-1038).

[111] Tsai, Y. H. H., et al. (2019). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6558-6569).

[112] Rahman, W., et al. (2020). Integrating multimodal information in large pretrained transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2359-2369).

[113] Yu, W., et al. (2021). Learning to align: A unified framework for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5375-5384).

[114] Wang, X., et al. (2021). Learning universal multimodal representations via generative modeling. arXiv preprint arXiv:2104.05868.

[115] Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. Journal of Affective Disorders, 208, 191-197.

[116] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning (pp. 1126-1135). PMLR.

[117] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems, 30, 4077-4087.

[118] Santoro, A., et al. (2016). Meta-learning with memory-augmented neural networks. In International Conference on Machine Learning (pp. 1842-1850). PMLR.

[119] Mittelstadt, B. D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. Science and Engineering Ethics, 22(2), 303-341.

[120] Bibal, A., & Frénay, B. (2016). Interpretability of machine learning models and representations: An introduction. In Proceedings of the 24th European Symposium on Artificial Neural Networks (pp. 77-82).

[121] Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. Basic Books.

[122] Vig, J., et al. (2020). Causal mediation analysis for interpreting neural NLP: The case of gender bias. arXiv preprint arXiv:2004.12265.

[123] Goyal, Y., et al. (2019). Counterfactual visual explanations. In International Conference on Machine Learning (pp. 2376-2384). PMLR.

[124] Kim, B., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In International Conference on Machine Learning (pp. 2668-2677). PMLR.

[125] Hessel, J., & Lee, L. (2020). Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think! In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 861-877).

[126] Liang, P. P., et al. (2020). MultiBench: Multiscale benchmarks for multimodal representation learning. arXiv preprint arXiv:2107.07502.

[127] Ferraro, F., et al. (2015). A survey of current datasets for vision and language research. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 207-213).

[128] Antol, S., et al. (2015). VQA: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2425-2433).

[129] Bunt, H. (2011). Multifunctionality in dialogue. Computer Speech & Language, 25(2), 222-245.

[130] Samek, W., et al. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247-278.

[131] Holzinger, A., et al. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

[132] Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793-4813.

[133] Che, Z., et al. (2018). Interpretable deep models for ICU outcome prediction. In AMIA Annual Symposium Proceedings (p. 371). American Medical Informatics Association.

[134] Ciftci, U. A., et al. (2020). FakeCatcher: Detection of synthetic portrait videos using biological signals. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(7), 2448-2461.

[135] Jin, Z., et al. (2021). Towards privacy-preserving speech processing: Voice conversion using homomorphic encryption. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1605-1617.

[136] Kaissis, G. A., et al. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305-311.

[137] Abadi, M., et al. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308-318).

[138] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.

[139] Voigt, P., & Von dem Bussche, A. (2017). The EU general data protection regulation (GDPR). Springer International Publishing.

[140] Obermeyer, Z., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

[141] Gianfrancesco, M. A., et al. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine, 178(11), 1544-1547.

[142] Mehrabi, N., et al. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

[143] Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 63(5), 82-89.

[144] Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. The New England Journal of Medicine, 378(11), 981.

[145] Fjeld, J., et al. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, (2020-1).

[146] Pescosolido, B. A., et al. (2010). "A disease like any other"? A decade of change in public reactions to schizophrenia, depression, and alcohol dependence. American Journal of Psychiatry, 167(11), 1321-1330.

[147] Kaye, J., et al. (2015). Dynamic consent: A patient interface for twenty-first century research networks. European Journal of Human Genetics, 23(2), 141-146.

[148] Fiesler, C., Garrett, N., & Beard, N. (2020). What do we teach when we teach tech ethics? A syllabi analysis. In Proceedings of the 51st ACM Technical Symposium on Computer Science Education (pp. 289-295).

[149] Budin-Ljøsne, I., et al. (2017). Dynamic consent: A potential solution to some of the challenges of modern biomedical research. BMC Medical Ethics, 18(1), 1-10.

[150] Williams, H., et al. (2015). Dynamic consent: A possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. JMIR Medical Informatics, 3(1), e3.

[151] Torous, J., & Keshavan, M. (2016). The role of social media in schizophrenia: evaluating risks, benefits, and potential. Current Opinion in Psychiatry, 29(3), 190-195.

[152] Keane, P. A., & Topol, E. J. (2018). With an eye to AI and autonomous diagnosis. NPJ Digital Medicine, 1(1), 1-3.

[153] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

[154] Liu, X., et al. (2019). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26(9), 1364-1374.

[155] Esteva, A., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.

[156] Mandel, J. C., et al. (2016). SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. Journal of the American Medical Informatics Association, 23(5), 899-908.

[157] Cresswell, K., et al. (2020). Ten key considerations for the successful implementation and adoption of large-scale health information technology. Journal of the American Medical Informatics Association, 27(7), 985-992.

[158] Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. JAMA, 318(6), 517-518.

[159] Price, W. N., Gerke, S., & Cohen, I. G. (2019). Potential liability for physicians using artificial intelligence. JAMA, 322(18), 1765-1766.

[160] Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In Artificial intelligence in healthcare (pp. 295-336). Academic Press.

[161] Trocin, C., Mikalef, P., & Papamitsiou, Z. (2022). Responsible AI for digital health: A synthesis and a research agenda. Information Systems Frontiers, 25(6), 2139-2157.

[162] Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[163] Zhou, C., et al. (2023). A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. arXiv preprint arXiv:2302.09419.

[164] Wornow, M., et al. (2023). The shaky foundations of large language models and foundation models for electronic health records. NPJ Digital Medicine, 6(1), 135.

[165] Kaddour, J., et al. (2022). Causal machine learning: A survey and open problems. arXiv preprint arXiv:2206.15475.

[166] Insel, T. R., & Cuthbert, B. N. (2015). Medicine: Brain disorders? Precisely. Science, 348(6234), 499-500.

[167] Jacobson, N. C., & Chung, Y. J. (2020). Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors, 20(12), 3572.

[168] Nock, M. K., et al. (2019). Annual research review: Suicidal thoughts and behaviors in children and adolescents–From assessment to intervention. Journal of Child Psychology and Psychiatry, 60(4), 377-393.

[169] Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-24.

[170] Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 29-38).

[171] Mosqueira-Rey, E., et al. (2023). Human-in-the-loop machine learning: A state of the art. Artificial Intelligence Review, 56(4), 3005-3054.

[172] Kessler, R. C., et al. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry, 62(6), 593-602.

[173] Janzing, J. G., et al. (2009). The role of psychiatric comorbidity in the quality of life of patients with epilepsy. Epilepsy & Behavior, 15(1), 118-124.

[174] Chekroud, A. M., et al. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. The Lancet Psychiatry, 3(3), 243-250.

[175] Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. Annual Review of Clinical Psychology, 14, 209-236.

[176] Mohr, D. C., et al. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. Annual Review of Clinical Psychology, 13, 23-47.

Acknowledgments

The authors gratefully acknowledge the contributions of researchers whose work has advanced multimodal deep learning for depression detection. We thank the organizers of the AVEC challenges for establishing benchmarks that have catalyzed research progress. This systematic review received no specific funding.

Author Contributions

Contribution: Author(s)
Conceptualization: VK, KG
Methodology: VK, KG
Screening and selection: VK, KG
Data extraction: VK, KG
Quality assessment: VK, KG
Writing—original draft: VK
Writing—review and editing: VK, KG
Visualization: VK
Supervision: KG

Conflicts of Interest

The authors declare no conflicts of interest.

Summary Tables

Note: This comprehensive systematic review provides an extensive analysis of the state-of-the-art in multimodal deep learning for depression detection. The findings highlight significant progress while identifying critical challenges that must be addressed for clinical translation. Researchers and practitioners should carefully consider ethical implications, validation requirements, and interpretability needs when developing and deploying these technologies.