Beyond Accuracy: A Review of Model and Data Trustworthiness in Audio-Visual Depression Recognition

Yuchen Pan[1], Hongxun Yao*[1], Wuxin Shen[1], and Lang He[2]

[1] Harbin Institute of Technology, Harbin, 150001, China
[2] Xi'an university of posts and communications, Xi'an, 710121, China
h.yao@hit.edu.cn

Abstract. Depression is a pervasive global mental health disorder with significant societal impact. While AI-based audio-visual assessment tools offer a promising, objective, and scalable alternative to traditional diagnostics, their translation to clinical practice is blocked by a critical hurdle: **trust**. These tools must be demonstrably reliable, fair, explainable, and private. This paper provides a comprehensive review of the automated depression recognition field, uniquely framed through the lens of Trustworthy AI. We first survey the primary technical frameworks from unimodal handcrafted features to end-to-end multimodal fusion, establishing the technical foundation. We then conduct an in-depth analysis of the core pillars of trustworthiness, including **explainability** (moving from post-hoc visualization to interpretable-by-design methods), **reliability and fairness** (analyzing Uncertainty Quantification and the challenge of equitable reliability), and the non-trivial challenge of **privacy** (surveying approaches from intermediate features, Federated Learning, and hardware-level Deep Optics). We conclude that a holistic focus on trustworthiness, including generalization, multimodal explainability, and the privacy-efficacy trade-off, is the central challenge and the most critical direction for the field's future.

Keywords: Depression Recognition · Multimodal · Trustworthy Artificial Intelligence · Explainability · Privacy Preserving

1 Introduction

Depression disorder represents a significant public health challenge with profound implications for individual well-being and societal productivity. It is recognized as one of the most pervasive and debilitating mental health disorders globally. According to the World Health Organization (WHO), depression affects over 5.7% of people worldwide; furthermore, recent data reveals a concerning 28% growth rate in severe depression cases, manifesting a notable increase in incidence. Nevertheless, a mere 10% of afflicted individuals actively seek medical intervention [1]. The principal impediment is often the pervasive "stigma" associated with this condition, which can lead to fear of social judgment, discrimination, and a reluctance to self-identify. Furthermore, traditional diagnostic methods often rely on subjective patient self-reporting and clinical interviews. This process can be time-consuming, resource-intensive, and limited by the availability of trained mental health professionals, highlighting a critical need for more accessible and objective assessment tools.

In recent years, scholars have shown increasing interest in the fusion of AI healthcare, propelled by the rapid advancement of multimodal algorithms and deep learning-based technologies. Among these, the analysis of facial expressions and vocal patterns has emerged as a particularly promising avenue. The human face is a primary visual channel for non-verbal behavioral communication, broadcasting a rich stream of affective signals through muscle movements, gaze patterns, and expression dynamics [102]. Similarly, acoustic patterns offer powerful, language-independent insights. While linguistic content (*what* is said) is a powerful indicator, the non-verbal and para-verbal signals (*how* it is said) provide a rich, parallel stream of information, often considered as acoustic signals [132]. In depressive disorders, these signals are often characteristically altered, manifesting as blunted affect, reduced expressivity, or depressed facial postures and monotone vocal characteristics.

AI-driven solutions that can capture and interpret these subtle cues possess the potential to enhance the precision of clinical assessment and broaden access to early mental health support. They can offer objective, quantifiable metrics to supplement clinical judgment, potentially facilitating earlier and more accurate diagnosis. By enabling accessible, private, and automated screening (e.g., via a smartphone app), they can help mitigate the apprehension and societal stigma that deter individuals from seeking help. Concurrently, AI-enabled digital monitoring equips healthcare providers with more comprehensive, longitudinal patient data from naturalistic settings, tracking subtle changes over time and thereby augmenting the quality and personalization of treatment.

Despite this promise, the development of depression recognition algorithms faces significant hurdles, whether relying on audio signals [134,55,27], visual data [56,33,85], or audio-visual solutions [79,57]. The medical field and the public at large place a considerable and necessary emphasis on ensuring privacy and reliability. The sensitive nature of both biometric data and mental health information creates a significant ethical and technical challenge. Consequently, there is a growing demand for innovative approaches that can simultaneously execute depression recognition while safeguarding user privacy, and the given output should be explainable, rather than only a prediction number. The genesis of depression recognition as a substantial research issue can be traced back to the Audio-Visual Emotion Challenge and Workshop (AVEC). This competition, held over several years, featured sub-challenges related to audio-visual depression recognition in 2013 [112], 2014 [111], 2016 [110], 2017 [95], and 2019 [94]. In 2013 and 2014, the dataset was composed of video recordings of participants for depression recognition, first in a long and mixed interviewing task in 2013, which was then categorized into *Northwind* and *Freeform* tasks in 2014. In 2016 and 2017, the competition used the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [23], which was extended to the Extended Distress Analysis Interview Corpus (E-DAIC) [94] in 2019 with more samples. This series of challenges clearly illustrates the roadmap: in the earlier editions, recognition predominantly relied on complete facial images. However, the organizers adopted a heightened focus on privacy protection from 2016. Consequently, the DAIC dataset provided only essential biometric facial information, such as VGG facial features as visual data, thereby substituting full facial images. Similar to the AVEC data collection and organization, there have been several other open-source datasets that contribute to this task, such as D-Vlog [125], LMVD [31], and CMDC [139].

This review paper surveys the landscape of audio-visual based depression recognition, navigating its core methodologies and critical challenges. We place a particular focus on the components of Trustworthy AI, especially model explainability and privacy risk. We begin in Section 2 by examining the primary technical frameworks, from traditional handcrafted feature extraction to modern deep learning-based spatiotemporal and fusion models. In Section 3, we delve into the crucial challenge of model explainability, reliability, and fairness, which are vital components for clinical trust and adoption. Section 4 addresses the contemporary and non-trivial concern of privacy, exploring methods that attempt to balance diagnostic efficacy with robust data protection. Finally, Section 5 concludes the paper by summarizing the key challenges and outlining promising directions for future research in this impactful domain.

2 Audio-Visual Depression Recognition

The automated depression recognition from behavioral and acoustic cues is a fundamentally multimodal problem. This section reviews the primary technical frameworks for this task, which are typically divided into three domains: analysis of the audio signal, analysis of the visual signal, and methods for fusing these two modalities.

2.1 Audio-based Recognition

Research on speech-based depression recognition has progressed from handcrafted descriptors to end-to-end deep representations, with a growing emphasis on attention, self-supervision, and multimodal fusion. Early work, grounded in the clinical sign of psychomotor retardation, treated depression as an altered coordination of the vocal tract. These approaches captured slowed and less synchronized articulation via articulatory trajectory estimates, acoustic-phonetic descriptors, and prosody [119,68]. Benchmark systems combined such features with traditional learners, including Support Vector Regression (SVR), Gaussian Mixture Models (GMM), and decision trees, to establish strong baselines for both classification and severity regression [112,111].

The shift to deep models introduced architectures like CNN-LSTM stacks, which jointly model local time-frequency patterns and long-range temporal dynamics, improving over shallow, hand-crafted pipelines [60]. Hybrid approaches subsequently combined learned and engineered cues, often using PCA or sparsity for compactness and generalization [29,66]. Concurrently, methodological studies addressed critical issues of aggregation, elicitation, and dataset reliability, such as using ℓ_p-norm pooling for evidence integration [77], optimizing speech-input length for speaker-independent inference [96], and auditing demographic or severity balance [127] to improve generalizability.

From 2020 onward, attention mechanisms and hierarchical designs became central. Examples include hierarchical attention transfer with attention autoencoders, and self-attention 3D log-Mel hybrids whose concatenated features are pooled and fed to SVR [132,133]. Time-domain architectures using dilated convolutions enlarged receptive fields, while efficient channel attention enhanced sensitivity to depression-related cues [6].

To mitigate challenges in small-data and cost-sensitive settings, some studies fused pretrained Speech Recognition (SR) and Speech Emotion Recognition (SER) embeddings in a hierarchical pipeline, which first classifies and then regresses severity, leveraging complementary vocal and affective information [17]. Parallel efforts emphasized robustness and regularization, alongside multi-resolution modulation-filtered cochleagrams to strengthen temporal modeling [82,88]. Traditional speech-technology baselines also remained informative, including i-vector/SVDA frameworks and MFCC-based RNNs, which provide competitive and interpretable points of comparison [69,93].

Recent work continues to push toward raw-waveform end-to-end learning and refined temporal-frequency fusion, exemplified by WavDepressionNet [76], TTFNet [10], attention-guided bi-directional models [55], and dense coordinate channel attention networks [134]. Meanwhile, self-supervised embeddings are being used to improve label efficiency. Emerging studies are also enriching acoustic data with transcribed speech, affect, and personality traits for severity estimation, signaling a shift toward an *acoustics-semantics-individual traits* paradigm [128,26].

Overall, the field has evolved from MFCC/prosody/articulatory proxies to attention-augmented representations, from shallow classifiers to hierarchical, multi-branch networks, and from preprocessed, handcrafted acoustic feature design to the end-to-end pipeline. With the recent progress of Large Language Models (LLMs), audio-based methods are also beginning to integrate semantic content with acoustic cues. However, audio-based depression recognition still faces open challenges in cross-corpus generalization, elicitation protocols, explainability, privacy, and principled fusion with lexical and person-centric information.

2.2 Visual-based Recognition

In visual depression recognition approaches, the extraction of facial features encompasses two categories: handcrafted features and deep features, both of which have been extended to capture temporal dynamics.

Handcrafted features are typically designed by researchers who leverage domain knowledge, employing specialized operators to process images. Examples include geometric features (distances and angles between facial landmarks), appearance features (Local Binary Patterns (LBP) or Histogram of Oriented Gradients (HOG)), and specific metrics like the frequency and duration of Action Units (AUs) [116,38,32]. However, these features are often customized for specific low-level image patterns or for general-purpose facial recognition. Consequently, when applied to depression recognition, generic feature extraction algorithms may produce suboptimal results. While designing new operators specifically for depression is an option, it presents substantial challenges. Furthermore, even when combined with deep learning models [99,101], handcrafted features require a distinct pre-processing step. The necessary parameters must be manually set, making them difficult to integrate into an auto-differentiable, end-to-end learning framework.

Deep feature extractors possess the capability to autonomously learn particular patterns from extensive datasets. Typically, these are deep learning models, such as pre-trained CNN or ViT image encoders, paired with a regression head to predict depression severity. Early approaches applied 2D CNNs (e.g., VGG, ResNet) to single-frame facial images. These models were usually pretrained on large-scale face datasets to obtain basic facial representations and then fine-tuned on depression-related datasets to learn task-relevant features [136,30]. A major limitation of such static 2D models is their neglect of temporal information. For example, a person might exhibit a brief distressed expression in response to an unpleasant scene, but this may not be enduring. Individuals with depression, however, tend to maintain more long-term characteristic facial expressions. Integrating continuous video frames thus provides crucial temporal capability [33,56,114]. Beyond motion, facial videos can also be a source of physiological signals; notably, remote photoplethysmography (rPPG) has been studied as an effective signal for depression recognition [7,114].

In summary, visual-based depression recognition approaches can be divided into two main categories:

1. Employing handcrafted spatiotemporal feature extraction operators. In this approach, features are first extracted and then used for depression recognition via statistical analysis [116,37] or a deep learning model [126,101]. This branch is usually pretrain-free, as the expert-based feature extractor is the main component and the learning model often serves only as a data regressor.
2. Using end-to-end deep models. This category includes 2D-CNNs or 3D-models combined with spatiotemporal models and attention mechanisms. These models are often pretrained on large-scale facial and dynamic video datasets for the feature extractor, then fine-tuned on depression-specific datasets [85,3]. In this branch, the feature encoder and the regression head are learned simultaneously and automatically.

The video-based approach still faces several challenges. Some subtle micro-expressions related to depression require high-quality video capture, which can be costly for real-world deployment. Furthermore, illumination and head pose can cause significant changes in visual features, making model robustness in varied environments a persistent challenge. Finally, as will be discussed in the following sections, explainability and privacy remain critical concerns.

2.3 Audio-Visual Fusion

Given that both vocal and facial channels carry complementary affective information, a logical progression is to fuse them. Fusion models aim to create a more robust and accurate representation than either modality could provide alone. The fusion method is a critical design choice. According to the typical taxonomy, fusion strategies can be categorized as feature-level, decision-level, and model-level. However, it should be noted that the boundaries between these strategies are not always distinct.

Feature-Level (Early Fusion): In this approach, features from audio and visual streams are combined at the beginning of the pipeline. For example, handcrafted audio features (MFCCs, Low-level descriptors (LLDs)) and visual features (AUs, facial landmarks) are concatenated into a single, large vector. This vector is then fed into a single classifier [40,22,4]. While simple, this method can be problematic as the modalities have different time scales and statistical properties, and concatenation can be a naive way of combining them which is challenging on the information discovery in single modality.

Decision-Level (Late Fusion): This method involves training separate, modality-specific models (e.g., one model for audio, one for video). The outputs or decisions from these models (e.g., their predicted depression scores) are then combined at the very end, often through simple averaging, a weighted sum, or a final integrated regression head [59,123,109,98]. This is a robust and common baseline, but it typically fails to capture any complex, non-linear interactions between the modalities (e.g., the synchrony of a smiling expression and a laugh).

Model-Level (Hybrid Fusion): This is the most complex and powerful approach, representing the focus of modern deep learning. Here, fusion occurs within the architecture of the model itself, allowing the model to learn complex inter-modal relationships. This category includes methods like Cross-Modal Attention, where one modality's representation is used to interact with and attend to the most relevant parts of the other modality [83,46,28,12]. For instance, the audio model might learn to pay more attention to the visual features of the mouth region when speech is detected. Multimodal Transformers, which are now state-of-the-art, convert both audio and visual features into a common latent space (often represented as tokens) and process them jointly, allowing for deep, complex, and temporally-aware fusion. This approach is also a promising avenue for integrating semantic (linguistic) information from LLMs [130,20,106]. This could help address performance degradation caused by language differences. However, the integration of LLMs with behavioral and acoustic cues is still under development.

Multimodal approaches represent the main trend in depression recognition, often combining state-of-the-art findings from both audio and visual-based methods. However, the main challenges revolve around high computational cost, the

Table 1: A systematic comparison of recent automated depression recognition methods on the AVEC dataset test sets. **M.**: Modality. **A/V**: Audio/Visual. +**T**: Includes text modality. **Dev set**: Reported on the development set.

(a) AVEC 2013

M.	Method	MAE↓	RMSE↓
A	AVEC 2013 Baseline [112]	10.35	14.12
	Two-Stage [59]	10.88	14.49
	PLSR [67]	9.14	11.19
	DCNN [29]	8.20	10.00
	Lp-norm [77]	7.48	9.79
	SAN-DCNN [133]	7.38	9.65
	Dis2DR (A) [83]	7.32	9.56
	FVCM [17]	7.32	8.73
	AGBiTNet [55]	7.29	9.45
	MFDS-VAN [87]	7.29	9.43
	MAFF [78]	7.14	9.50
	TTFNet [10]	7.08	8.93
	TDCA-Net [6]	6.90	9.22
	DCCANet [134]	6.78	8.47
	TFCAV [75]	6.26	8.32
	WavDepressionNet [76]	6.14	8.20
	DSDD [79]	6.09	8.27
	GMM [119] (Dev set)	5.75	7.42
V	AVEC 2013 Baseline [112]	10.88	13.61
	Sparse Coding [116]	8.22	10.27
	CCA [44]	7.86	9.72
	AD-DCNN [138]	7.58	9.82
	DPFV [32]	7.55	9.20
	RNN-C3D [3]	7.37	9.28
	MAFF [78]	7.32	8.97
	OpticalDR [84]	7.53	8.48
	VLDN-LSTM [108]	7.04	8.93
	DJ-LDML [137]	6.63	8.37
	DLGA-CNN [30]	6.59	8.39
	C3D-GAP [62]	6.40	8.26
	rPPG [7]	6.43	8.01
	LQGDNet [99]	6.38	8.20
	CNN-DDL [63]	6.30	8.25
	DepressNet [136]	6.20	8.28
	DAER [81]	6.28	8.13
	LMB [126]	6.28	7.54
	MDN [65]	6.24	7.55
	Behavior Primitives [101]	6.16	8.10
	STA-DRN [85]	6.15	7.98
	MTDAN [129]	6.14	8.08
	DMSN [16]	6.14	7.66
	LMTformer [36]	6.12	7.75
	Depressioner [72]	6.12	7.49
	EGDC [5]	6.09	8.05
	SE-TOV [80]	6.09	7.42
	PRA-Net [58]	6.08	7.59
	HMHN [51]	6.05	7.38
	LMS-VDR [124]	6.04	7.68
	MM-CRN [54]	6.04	7.61
	Dis2DR (V) [83]	6.04	7.50
	CFGMamba [56]	6.01	7.59
	MSN [64]	5.98	7.90
	Hi-Lo [131]	5.97	7.36
	Two-Stream [61]	5.96	7.97
	LSCAformer [34]	5.89	7.69
	LDBM [114]	5.71	6.99
	PTN [73]	5.62	7.36
	Depressformer [33]	5.49	7.47
	MLM-EOE [52]	5.49	6.84
	DepressionMLP [74]	5.43	7.49
	STE-Mamba [53]	5.27	7.26
	DSDD [79]	5.12	7.21
A-V	A-V System [41]	9.09	11.19
	MHH + PLSR [13]	-	10.62
	Fusion [67]	8.72	10.96
	CCA [44]	7.68	9.44
	FedDAAM [35]	6.78	8.61
	Two-Stage [59]	6.75	8.29
	TMFE-GFN [19]	6.60	9.00
	MAFF [78]	6.14	8.16
	AVA-DepressNet [86]	6.23	7.99
	FDFNet [46]	6.22	7.58
	Dis2DR (A-V) [83]	6.12	7.97
	MFMamba [57]	6.11	7.05
	VLDSP-TAP-MFB [109]	5.38	6.83
	DSDD [79]	5.19	6.48
	STE-Mamba [53]	5.08	6.94

(b) AVEC 2014

M.	Method	MAE↓	RMSE↓
A	AVEC 2014 Baseline [111]	10.04	12.57
	PLSR [39]	9.10	11.30
	SRI audio system [68]	8.83	11.10
	DCNN [29]	8.19	10.00
	Lp-norm [77]	8.02	9.66
	SAN-DCNN [133]	7.94	9.57
	MAFF [78]	7.65	9.13
	Dis2DR (A) [83]	7.63	9.28
	TFCAV [75]	7.49	9.25
	MFDS-VAN [87]	7.33	9.44
	AGBiTNet [55]	7.30	9.47
	TTFNet [10]	7.13	8.96
	TDCA-Net [6]	7.08	8.90
	i-vector + SVDA [69]	6.99	8.90
	FVCM [17]	6.80	8.82
	WavDepressionNet [76]	6.60	8.61
	DSDD [79]	6.22	8.14
	DCCANet [134]	6.17	7.54
V	AVEC 2014 Baseline [111]	8.86	10.86
	PLSR [39]	8.44	10.50
	AD-DCNN [138]	7.47	9.55
	OpticalDR [84]	7.89	8.82
	RNN-C3D [3]	7.22	9.20
	DPFV [32]	7.21	9.01
	VLDN-LSTM [108]	6.86	8.78
	ERBMA-Net [45]	6.80	8.18
	MAFF [78]	6.43	8.60
	rPPG [7]	6.57	8.49
	C3D-GAP [62]	6.59	8.31
	DJ-LDML [137]	6.59	8.30
	DLGA-CNN [30]	6.51	8.30
	DepressNet [136]	6.21	8.39
	MTDAN [129]	6.35	7.93
	Two-Stream [61]	6.20	7.94
	DeepFusion [9]	6.16	8.13
	CNN-DDL [63]	6.15	8.23
	DAER [81]	6.14	8.07
	LMB [126]	6.14	7.58
	LQGDNet [99]	6.08	7.84
	MDN [65]	6.06	7.65
	LMTformer [36]	6.05	7.97
	PRA-Net [58]	6.04	7.98
	HMHN [51]	6.01	7.60
	Depressioner [72]	6.01	7.56
	STA-DRN [85]	6.00	7.75
	MM-CRN [54]	5.99	7.59
	LMS-VDR [124]	5.98	7.59
	EGDC [5]	5.94	7.98
	Behavior Primitives [101]	5.95	7.15
	FacialPulse [113]	5.92	7.60
	Dis2DR (V) [83]	5.92	7.09
	LSCAformer [34]	5.91	7.55
	CFGMamba [56]	5.96	7.52
	SE-TOV [80]	5.87	7.39
	Hi-Lo [131]	5.85	7.23
	MSN [64]	5.82	7.61
	TSFFM [47]	5.75	7.91
	DMSN [16]	5.69	7.50
	DepressionMLP [74]	5.63	7.27
	Depressformer [33]	5.56	7.22
	DSDD [79]	5.52	6.80
	MLM-EOE [52]	5.51	7.22
	PTN [73]	5.41	7.12
	STE-Mamba [53]	5.38	7.24
	LDBM [114]	5.36	6.93
	DepMGNN [120]	4.99	6.28
A-V	AVEC 2014 Baseline [111]	7.89	9.89
	Fusion System [89]	8.99	10.82
	Model Fusion [98]	8.33	10.43
	Fisher Vector [37]	8.40	10.25
	PLSR fusion [39]	8.30	10.26
	LSTM-RNN [8]	7.91	9.98
	CCA [43]	7.69	9.61
	Meta knowledge [42]	7.10	9.19
	TMFE-GFN [19]	7.05	9.45
	Feature selection [25]	-	8.99
	FedDAAM [35]	6.77	8.59
	GMM + ELM [118]	6.31	8.12
	PLSR + LR [40]	6.14	7.43
	M-BAM [12]	5.78	7.47
	Dis2DR (A-V) [83]	5.45	6.61
	AVA-DepressNet [86]	5.32	6.83
	FAU-GF [21]	5.26	6.80
	MAFF [78]	5.21	7.03
	FDFNet [46]	5.21	6.49
	MFMamba [57]	5.16	6.71
	STE-Mamba [53]	5.10	6.77
	VLDSP-TAP-MFB [109]	5.03	6.16
	DSDD [79]	4.89	5.99

(c) DAIC-WOZ (E-DAIC)

M.	Method	MAE↓	RMSE↓
A	DAIC-WOZ Baseline [95]	5.72	7.78
	E-DAIC Baseline [94] (E-DAIC)	-	8.19
	CNN-GAN [115]	7.32	8.56
	Acoustic + RF [127] (E-DAIC)	5.77	6.78
	AFN [90]	5.67	6.55
	Acoustic + LR [127]	5.49	7.18
	DEPA [128] (Dev set)	5.48	6.31
	STFN [27] (E-DAIC)	5.38	6.29
	STFN [27]	5.38	6.36
	GSM [117] (Dev set)	5.36	6.74
	LLD + Fisher Vector [105]	5.30	6.34
	STE-Mamba [53]	5.27	6.79
	Random Forest [107]	5.22	6.17
	LSTM [2]	5.13	6.50
	TTFNet [10]	5.09	6.01
	TTFNet [10] (E-DAIC)	5.00	5.76
	Dis2DR (A) [83]	4.88	5.60
	STE-Mamba [53] (E-DAIC)	4.80	5.80
	DSDD [79]	4.62	5.61
	AGBiTNet [55]	4.27	5.35
	MFDS-VAN [87]	4.27	5.34
	HATN [132]	4.20	5.51
	MLAtt [92] (Dev set, E-DAIC)	-	5.11
	REPT [103]	4.11	4.94
	EmoAudioNet [82]	-	4.14
	HCAG [71] (+T)	2.94	3.80
V	DAIC-WOZ Baseline [95]	6.12	6.97
	E-DAIC Baseline [94]	-	8.01
	GSM [117] (Dev set)	5.88	7.13
	REPT [103]	5.36	6.72
	STE-Mamba [53] (E-DAIC)	4.89	6.28
	HOG-PCA [104]	4.89	6.23
	FDR + LDA [91]	4.64	5.98
	DepArt-Net [18]	4.61	5.78
	STE-Mamba [53]	4.58	5.99
	Dis2DR (V) [83]	4.51	5.88
	Behaviour-based [102] (Dev set)	4.37	5.84
	MLAtt [92] (Dev set, E-DAIC)	-	5.38
	DepressionMLP [74]	4.11	5.03
	DSDD [79]	4.09	5.40
	PTN [73]	3.84	5.08
A-V	DAIC-WOZ Baseline [95]	5.66	7.05
	E-DAIC Baseline [94]	-	6.37
	GSM [117] (Dev set)	5.52	6.62
	ANEW + GSR [15] (+T)	5.30	6.52
	STE-Mamba [53]	5.18	6.24
	DCNN-DNN-1 [122] (+T)	5.16	5.97
	STE-Mamba [53] (E-DAIC)	5.05	6.21
	AVA-DepressNet [86]	4.62	5.78
	Dis2DR (A-V) [83]	4.69	5.49
	FPT-Former [48] (E-DAIC)	4.58	4.80
	FDFNet [46] (E-DAIC)	4.41	5.10
	DCNN-DNN-2 [123] (+T)	4.36	5.40
	FDFNet [46]	4.25	5.34
	Two-stage [14] (+T)	3.98	5.11
	Topic Modeling [22] (+T)	3.96	4.99
	FAU-GF [21] (E-DAIC)	3.77	4.95
	FedDAAM [35]	3.68	4.71
	C-CNN [28] (Dev set)	3.67	-
	DSDD [79]	3.53	4.76

need for high-quality and complete multimodal data, and the need for explainable predictions, especially regarding cross-modal interactions.

Finally, these various approaches, from single-modality handcrafted features to deep multimodal fusion models, are systematically cataloged and their performance comprehensively compared in Table 1.

3 Model Trustworthiness: Explainability, Reliability, and Fairness Depression Recognition

Can we trust the model's output?
Is it explainable, reliable, and fair
in its judgments?

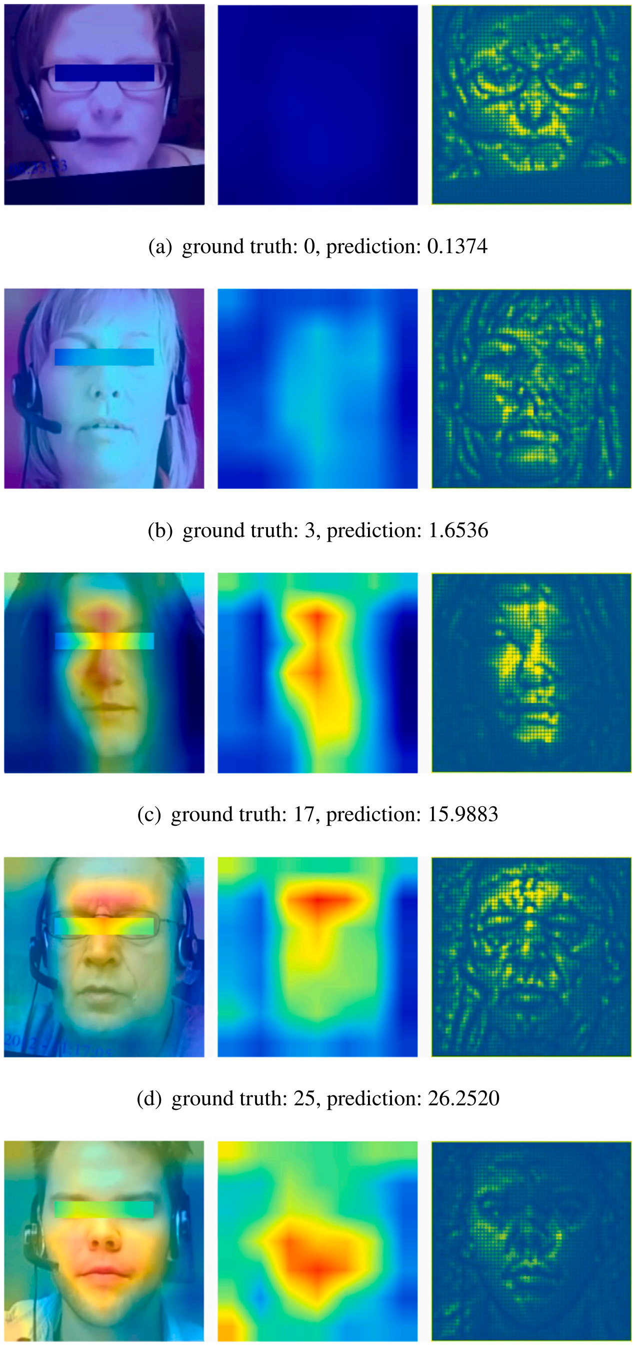
The adoption of complex, deep learning models in a high-stakes clinical domain like mental health diagnostics is contingent not only on their accuracy but also on their transparency and reliability. For a clinician to trust and act upon an AI-generated assessment, the model's "black box" nature must be addressed. This need has given rise to a focus on Explainable Artificial Intelligence (XAI), which seeks to render model decisions comprehensible to human users. In the context of audio-visual depression recognition, this translates to a critical question: Which specific cues is the model using to arrive at its depression score, and do these cues align with clinical knowledge?

This section reviews the key pillars of building trust in these models: 1) post-hoc explainability, 2) interpretable-by-design models, and 3) the quantification of reliability and fairness.

3.1 Post-hoc Visualization and Saliency Methods

The most common approach to XAI in this domain is to apply post-hoc visualization techniques, which generate saliency or attention maps. These methods, such as Class Activation Mapping (CAM) [135] and its derivatives Grad-CAM [97], identify which parts of the input data most strongly influenced the model's final decision.

Several studies [136,65,85,3] have endeavored to use these techniques to visualize the features learned by deep learning depression recognition models. For visual data, this typically produces heatmaps that highlight specific facial regions. These efforts have successfully demonstrated that models learn to focus on clinically-relevant areas, such as the periorbital (eyes, eyebrows) or perioral (mouth) regions, to identify significant facial actions associated with depression. For instance, Figure 1 shows the visualization results from STA-DRN [85] as a typical example of this analysis. Similarly, for the audio modality, attention weights can reveal which speech segments or acoustic events (e.g., pauses, specific word-level prosody) most contributed to the final score [76,78].



(a) ground truth: 0, prediction: 0.1374

(b) ground truth: 3, prediction: 1.6536

(c) ground truth: 17, prediction: 15.9883

(d) ground truth: 25, prediction: 26.2520

(e) ground truth: 44, prediction: 43.5324

Fig. 1: Visualization analysis from STA-DRN [85]. *(Left)* CAM highlight broad facial regions contributing to the depression score (red indicates strong activation, blue indicates weak). *(Right)* Guided backpropagation maps reveal the specific pixel-level features (e.g., edges around the eyes and mouth, shown in yellow) that the model relies on within those regions.

However, these visualization methods have a significant limitation. While they excel at answering where the model is looking, they provide little insight into why or what it has learned. A heatmap might highlight a key feature or region, but it cannot, by itself, explain the intricate relationship between a lack of smile dynamics or a specific vocal pitch, and the model's depression score. As such, a more nuanced understanding of the relationship between these regions and the model's reasoning remains an area of limited development. This refinement is essential to move beyond simple model validation and provide genuine, actionable insights for clinical advancement.

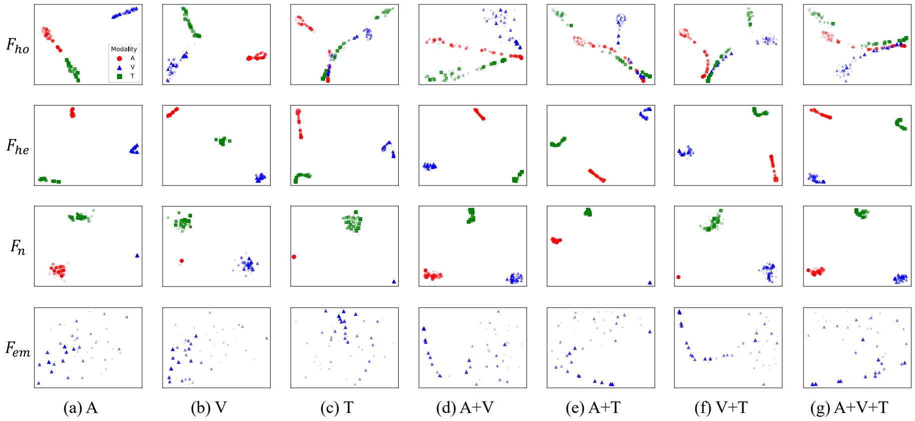
3.2 Interpretable-by-Design with Feature Disentanglement

A more recent and complex path to enhancing model explainability is through feature disentanglement. Rather than treating the model as a black box to be explained later, this approach aims to build models whose internal representations are inherently meaningful and separated by concept. In the context of multimodal depression recognition, this is a particularly powerful but challenging idea. An individual's audio-visual stream contains a massive entanglement of information:

1. **Identity**: The unique shape of a person's face or the fundamental pitch of their voice. This includes attributes such as gender and other personal identity information.
2. **Content**: The linguistic information being spoken, which is often independent of the acoustic features.
3. **Affect/Sentiment**: The emotional and physical state of the speaker (e.g., depression, fatigue, joy).

A robust depression model should, ideally, be invariant to identity and content, focusing only on the affective cues. Disentanglement aims to force the model to learn separate, non-overlapping representations for these factors. Recent research has begun to explore modality-level disentanglement [83,46,70], but this remains a nascent field.

The study of the interaction mechanism between modalities is even more scarce. A truly explainable system would not only disentangle features within each modality (e.g., separate speech content from vocal prosody) but also explain the fusion process itself. For example, how does the model learn to weigh the information from a flat, monotone voice against the information from a simultaneous subtle facial expression? In Dis2DR [83], this question is answered to a certain extent by revealing the modality homogeneous and heterogeneous mechanisms of depression recognition. Figure 2 shows an example of such research, visualizing the disentangled features. The experimental results indicate that the text and visual modalities are the most crucial inputs, with the homogeneous feature (F_{ho}) space demonstrating a critical sensitivity to their absence. While F_{ho} effectively learns identical cross-modal representations that correlate strongly with depression severity, manifesting as tight clustering for severe cases



(a) A (b) V (c) T (d) A+V (e) A+T (f) V+T (g) A+V+T

Fig. 2: Visualization of disentangled features from Dis2DR [83]. The model learns to separate modality-heterogeneous patterns from modality-homogeneous patterns across different modalities.

and divergence for mild cases, the heterogeneous feature (F_{he}) space provides superior robustness. Specifically, F_{he} maintains proper representations even when a modality is missing, suggesting a powerful capability for handling incomplete data, while a high entanglement between audio and text modalities allows audio to implicitly compensate for missing textual information in the F_{ho} representation. However, research on this novel topic remains limited. Answering these questions is the next frontier for XAI and is crucial for understanding the complex, multimodal signature of depression.

3.3 Reliability and Fairness in Uncertainty Quantification

A distinct but complementary component of trustworthiness is predictive reliability. A model that provides a score without expressing its own confidence is clinically incomplete. A clinician needs to know if a depression score of "15" is a confident "15 ± 2" or a highly uncertain "15 ± 10". This is the domain of Uncertainty Quantification (UQ). UQ is vital as it provides clinicians with the necessary confidence measure for the model's prediction, enabling them to safely override uncertain assessments and preventing the model from being overconfident in a potentially life-altering diagnosis like severe depression.

Recently, achieving reliable depression predictions through UQ has attracted increasing attention [49,50,11]. The goal is to produce not just a single-point estimate (e.g., a score) but a valid, calibrated range or confidence interval. Methods such as Conformal Prediction [49] are gaining traction as they can provide such intervals with theoretical guarantees on their coverage.

Furthermore, it is crucial to investigate the algorithmic fairness of UQ. Fairness is paramount because diagnostic bias (e.g., against specific demographics like gender, age, or ethnicity) can lead to systematic underdiagnosis or misdiagnosis, denying vulnerable populations necessary care and worsening existing health disparities. Thus UQ in depression recognition should not only provide statistically valid confidence intervals globally but also ensure fair coverage rates across different demographic groups. Current UQ methods often produce disparate coverage rates across groups. This can lead to unfair predictions where majority groups are over-covered, resulting in excessively broad, clinically unhelpful intervals. While minority groups are under-covered, yielding dangerously overconfident and misleadingly narrow intervals. This is a subtle but dangerous form of bias, as the average reliability across the entire dataset may appear high, masking these severe group-level disparities. The work on Fair Uncertainty Quantification (FUQ) addresses this by introducing concepts like the Equal Opportunity of Coverage (EOC) [50]. This approach uses group-based analysis by sensitive attributes (e.g., gender, age) and a fairness-aware optimization strategy that works to equalize coverage rates, ensuring the model's reliability is equitable across demographic groups. This pursuit of fair and reliable predictions is essential for the ethical and responsible deployment of these technologies in diverse clinical populations.

4 Data Trustworthiness: Privacy-Preserving Depression Recognition

> Can we trust the system with our data? Is it private? Will it protect my identity?

The deployment of automated mental health assessment tools carries a profound privacy challenge. First, the depression assessment output of the model constitutes sensitive personal health information. Second, the raw video and audio of an individual input to the model is itself highly sensitive biometric data that can be used for re-identification. A robust framework must therefore protect both the user's diagnosis and their biometric identity.

With the growing emphasis on safeguarding data, research has moved beyond models that require raw audio-visual data. These efforts can be broadly categorized into methods that use intermediate with less-identifiable features, and methods that are structurally designed for privacy from the ground up.

4.1 Privacy-by-Processing: Using Intermediate Representations

A primary strategy is to remove identifiable information by processing the raw data into abstract and intermediate representations before analysis. The assumption is that these representations can preserve diagnostically relevant behavioral patterns while discarding biometric identifiers.

For visual data, this involves avoiding full face images. Instead, researchers have explored the utilization of non-image structural data, including facial landmarks (only 2D/3D coordinates of key facial points), Action Units (codified muscle activations), and gaze or head pose direction. This approach directly aligns with the later AVEC challenges after 2016, which provided only these features. However, current research predominantly centers on the temporal variations of these features, treating them as simple temporal data for analysis [101,28,18]. This can be a significant limitation, as it may miss crucial spatial relationships. In contrast, AVA-DepressNet [86] delved into the spatiotemporal characteristics of facial landmarks, considering the dynamic changes in their spatial interrelationships, and demonstrated the utility of this richer yet still privacy preserving representation.

A parallel approach exists for the audio modality. Instead of processing the raw waveform, which contains a unique and identifiable voiceprint [87], models can be trained on pre-extracted, low-level acoustic feature sets. These include prosodic features (pitch, energy) and spectral features (e.g., MFCCs). Like facial landmarks, these features are considered privacy-friendlier as they are more difficult to reverse engineer into a recognizable voice, thus preventing user identity or conversation content from being leaked. It is noteworthy that early research predominantly used such audio features for machine learning analysis [119,118]. With the rise of sequential deep learning models, end-to-end training on the raw waveform [27,76] has become common, which re-introduces the privacy risk of loading entire audio signals into an online or cloud-based AI assessment system.

4.2 Privacy-by-Design: Architectural and Hardware Solutions

While intermediate features offer a partial solution, a more robust line of inquiry focuses on building depression recognition systems that are inherently private by design, either at the software or hardware level.

Federated Learning (FL): One of the most prominent architectural solutions is FL. In this paradigm, the raw audio-visual data never leaves the user's local device (e.g., smartphone, laptop). A global model is sent to the device, trained locally on the user's private data, and only the resulting model updates (gradients) are sent back to a central server to be aggregated [121,24,35], rather than uploading any user data itself. This approach is a powerful solution for protecting raw data, though challenges in data heterogeneity and communication efficiency remain, which are also common issues in the broader FL field.

Hardware-Level Anonymization: A novel and cutting-edge paradigm for privacy protection is to intervene at the point of data capture itself. Given that digital data is vulnerable to leaks at any point in the processing pipeline, the core idea is to prevent the creation of digital data containing private information in the first place. This idea has led to the development of Deep Optics [100], which explores the joint design of a physical camera lens and a deep learning model. For depression recognition, this has led to frameworks like OpticalDR [84] as illustrated in Figure 3, where the camera lens is optically optimized to perform a task-specific encryption. It is trained to physically distort the light from the scene, erasing identifiable facial features *before* the light ever hits the sensor, while simultaneously preserving the subtle, depression-related features. The resulting blurred image is incomprehensible not only to a human but also to state-of-the-art facial recognition systems, yet it is perfectly decodable by the jointly optimized neural network. This method provides an exceptionally strong, irreversible privacy guarantee at the hardware level.

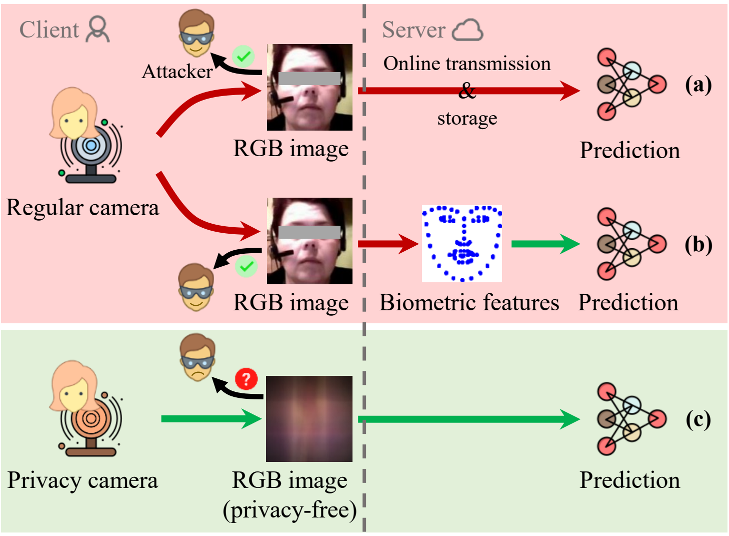


Fig. 3: The conceptual difference between traditional systems and the OpticalDR [84] framework. Red arrows represent data flows containing sensitive private information. Green arrows represent data flows where the private information has been removed.

4.3 The Privacy-Efficacy Trade-Off

Despite these advancements, research regarding privacy in audio-visual depression recognition remains relatively limited. The central challenge is the trade-off between the robustness of privacy protection and the effectiveness of depression recognition. Every privacy-enhancing step, from using abstract landmarks to applying optical distortion, carries an inherent risk of information loss.

This trade-off is the central emerging problem in privacy-preserving depression recognition. For instance, while a deep optics solution like OpticalDR [84] can effectively balance performance and privacy, making private information almost entirely unrecoverable while maintaining usable depression recognition performance. But it is not without costs. The resulting performance, while usable, still has a gap compared to the best state-of-the-art models that use full data. Furthermore, such end-to-end hardware solutions are often highly specific. They are hard to generalize, meaning that if the task or predictive labels change, the entire system must be re-trained or even re-designed from scratch. This lack of flexibility is a significant limitation for real-world deployment.

5 Conclusion

The field of automated depression recognition from audio-visual signals has witnessed several years of development, and has achieved notable performance. Research has evolved from static, single-modality feature engineering to complex, end-to-end spatiotemporal and multimodal fusion models. However, as this review has cataloged, performance on benchmark datasets is only the first step. Several emerging challenges must be addressed before these technologies can be responsibly translated into clinical practice.

1) Model robustness and generalization remain a fundamental challenge. The field's heavy reliance on small, homogeneous, and lab-collected datasets (AVEC and DAIC) limits the real-world applicability of current models. Future work must prioritize the collection of larger, more diverse, in-the-wild datasets [31,125] that capture a wider range of demographics and cultures. This is also a prerequisite for developing personalized models, which learn to track an individual's behavioral changes against their own baseline, rather than a population-level average.

2) The trustworthiness of these models is a critical and previously neglected area of research. For clinical adoption, a model must be transparent, reliable, and fair. For Explainability, the field is moving beyond simple CAM-based visualizations, which are limited to *where* a model is looking. Promising new directions, such as feature disentanglement [83,46,70], are beginning to explore the *why*, but this remains a nascent and highly complex area. For Reliability and Fairness, recent studies on UQ [49] and Fair UQ [50,11] represent a significant step forward. Ensuring that a model's confidence is both statistically valid and equitable across demographic groups is essential for ethical deployment.

3) Privacy and security necessitate a comprehensive, "privacy-by-design" approach related to data trustworthiness. Existing solutions such as using intermediate features [86], FL frameworks [121,24,35], and hardware-level solutions [84], all present a difficult trade-off between privacy robustness and model performance. This balance remains a central research problem. Moreover, it is worth noting that while visual privacy has received some attention, the specific challenges of audio privacy (e.g., voiceprint and content leakage from raw waveforms) and the complex privacy implications of multimodal fusion are almost entirely unstudied.

In conclusion, the future of automated depression recognition lies not in a singular pursuit of accuracy, but in a holistic approach that balances model performance with generalization, clinical trustworthiness, and robust privacy guarantees.

References

1. Mental health atlas 2024. World Health Organization, Geneva (2025)
2. Al Hanai, T., Ghassemi, M., Glass, J.: Detecting Depression with Audio/Text Sequence Modeling of Interviews. In: Interspeech. pp. 1716–1720 (2018)
3. Al Jazaery, M., Guo, G.: Video-based depression level analysis by encoding deep spatiotemporal features. IEEE Transactions on Affective Computing **12**(1), 262–268 (2021). https://doi.org/10.1109/TAFFC.2018.2870884
4. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Hyett, M., Parker, G., Breakspear, M.: Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors. IEEE Transactions on Affective Computing **9**(4), 478–490 (2018). https://doi.org/10.1109/TAFFC.2016.2634527
5. Bargshady, G., Goecke, R.: Estimating depression severity from long-sequence face videos via an ensemble global diverse convolutional model. In: 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 296–303 (2023). https://doi.org/10.1109/DICTA60407.2023.00048
6. Cai, C., Niu, M., Liu, B., Tao, J., Liu, X.: TDCA-Net: Time-Domain Channel Attention Network for Depression Detection. In: Interspeech 2021. pp. 2511–2515 (2021). https://doi.org/10.21437/Interspeech.2021-1176
7. Casado, C.A., Cañellas, M.L., López, M.B.: Depression recognition using remote photoplethysmography from facial videos. IEEE Transactions on Affective Computing **14**(4), 3305–3316 (2023). https://doi.org/10.1109/TAFFC.2023.3238641
8. Chao, L., Tao, J., Yang, M., Li, Y.: Multi task sequence learning for depression scale prediction from video. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 526–531 (2015). https://doi.org/10.1109/ACII.2015.7344620
9. Chen, Q., Chaturvedi, I., Ji, S., Cambria, E.: Sequential fusion of facial appearance and dynamics for depression recognition. Pattern Recognition Letters **150**, 115–121 (2021). https://doi.org/https://doi.org/10.1016/j.patrec.2021.07.005, https://www.sciencedirect.com/science/article/pii/S0167865521002397
10. Chen, X., Shao, Z., Jiang, Y., Chen, R., Wang, Y., Li, B., Niu, M., Chen, H., Hu, Q., Wu, J., Yang, C., Shang, Y.: TTFNet: Temporal-frequency features fusion network for speech based automatic depression recognition and assessment. IEEE Journal of Biomedical and Health Informatics **29**(10), 7536–7548 (2025). https://doi.org/10.1109/JBHI.2025.3574864
11. Cheong, J., Bangar, A., Kalkan, S., Gunes, H.: U-Fair: Uncertainty-based Multimodal Multitask Learning for Fairer Depression Detection. In: Hegselmann, S., Zhou, H., Healey, E., Chang, T., Ellington, C., Mhasawade, V., Tonekaboni, S., Argaw, P., Zhang, H. (eds.) Proceedings of the 4th Machine Learning for Health Symposium. Proceedings of Machine Learning Research, vol. 259, pp. 203–218. PMLR (15–16 Dec 2025), https://proceedings.mlr.press/v259/cheong25a.html
12. Cholet, S., Paugam-Moisy, H., Regis, S.: Bidirectional associative memory for multimodal fusion: a depression evaluation case study. In: International Joint Conference on Neural Networks (IJCNN). pp. 1–6 (2019)
13. Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., Epps, J.: Diagnosis of depression by behavioural signals: a multimodal approach. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 11–20. Association for Computing Machinery, New York, NY, USA (2013)

14. Dai, Z., Zhou, H., Ba, Q., Zhou, Y., Wang, L., Li, G.: Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. Journal of Affective Disorders **295**, 1040–1048 (2021). https://doi.org/https://doi.org/10.1016/j.jad.2021.09.001, https://www.sciencedirect.com/science/article/pii/S0165032721009599
15. Dang, T., Stasak, B., Huang, Z., Jayawardena, S., Atcheson, M., Hayat, M., Le, P., Sethu, V., Goecke, R., Epps, J.: Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in AVEC 2017. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 27–35. Association for Computing Machinery, New York, NY, USA (2017)
16. de Melo, W.C., Granger, E., Lopez, M.B.: Facial expression analysis using decomposed multiscale spatiotemporal networks. Expert Systems with Applications **236**, 121276 (2024). https://doi.org/https://doi.org/10.1016/j.eswa.2023.121276, https://www.sciencedirect.com/science/article/pii/S0957417423017785
17. Dong, Y., Yang, X.: A hierarchical depression detection model based on vocal and emotional cues. Neurocomputing **441**, 279–290 (2021)
18. Du, Z., Li, W., Huang, D., Wang, Y.: Encoding visual behaviors with attentive temporal convolution for depression prediction. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–7 (2019). https://doi.org/10.1109/FG.2019.8756584
19. Fan, H., Zhang, X., Xu, Y., Fang, J., Zhang, S., Zhao, X., Yu, J.: Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. Information Fusion **104**, 102161 (2024). https://doi.org/https://doi.org/10.1016/j.inffus.2023.102161, https://www.sciencedirect.com/science/article/pii/S1566253523004773
20. Feng, S., Sun, G., Lubis, N., Wu, W., Zhang, C., Gasic, M.: Affect recognition in conversations using large language models. In: Kawahara, T., Demberg, V., Ultes, S., Inoue, K., Mehri, S., Howcroft, D., Komatani, K. (eds.) Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 259–273. Association for Computational Linguistics, Kyoto, Japan (Sep 2024). https://doi.org/10.18653/v1/2024.sigdial-1.23, https://aclanthology.org/2024.sigdial-1.23/
21. Fu, C., Qian, F., Su, Y., Su, K., Song, S., Niu, M., Shi, J., Liu, Z., Liu, C., Ishi, C.T., Ishiguro, H.: Facial action units guided graph representation learning for multimodal depression detection. Neurocomputing **619**, 129106 (2025). https://doi.org/https://doi.org/10.1016/j.neucom.2024.129106, https://www.sciencedirect.com/science/article/pii/S0925231224018770
22. Gong, Y., Poellabauer, C.: Topic Modeling Based Multi-modal Depression Detection. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. p. 69–76. AVEC '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3133944.3133945, https://doi.org/10.1145/3133944.3133945
23. Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., Morency, L.P.: The distress analysis interview corpus of human and computer interviews. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 3123–3128.

European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), https://aclanthology.org/L14-1421/
24. Gupta, C., Khullar, V.: Modality independent federated multimodal classification system detached eeg, audio and text data for iid and non-iid conditions. Biomedical Signal Processing and Control **108**, 107938 (2025). https://doi.org/https://doi.org/10.1016/j.bspc.2025.107938, https://www.sciencedirect.com/science/article/pii/S1746809425004495
25. Gupta, R., Malandrakis, N., Xiao, B., Guha, T., Van Segbroeck, M., Black, M., Potamianos, A., Narayanan, S.: Multimodal prediction of affective dimensions and depression in human-computer interactions. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. p. 33–40. AVEC '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2661806.2661810, https://doi.org/10.1145/2661806.2661810
26. Gönç, K., Dibeklioğlu, H.: Affect and personality aided modeling of transcribed speech for depression severity estimation. IEEE Transactions on Affective Computing **16**(3), 2334–2351 (2025). https://doi.org/10.1109/TAFFC.2025.3560476
27. Han, Z., Shang, Y., Shao, Z., Liu, J., Guo, G., Liu, T., Ding, H., Hu, Q.: Spatial–temporal feature network for speech-based depression recognition. IEEE Transactions on Cognitive and Developmental Systems **16**(1), 308–318 (2024). https://doi.org/10.1109/TCDS.2023.3273614
28. Haque, A., Guo, M., Miner, A.S., Fei-Fei, L.: Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions. arXiv e-prints arXiv:1811.08592 (Nov 2018). https://doi.org/10.48550/arXiv.1811.08592
29. He, L., Cao, C.: Automated depression analysis using convolutional neural networks from speech. Journal of Biomedical Informatics **83**, 103–111 (2018)
30. He, L., Chan, J.C.W., Wang, Z.: Automatic depression recognition using cnn with attention mechanism from videos. Neurocomputing **422**, 165–175 (2021). https://doi.org/https://doi.org/10.1016/j.neucom.2020.10.015, https://www.sciencedirect.com/science/article/pii/S0925231220315101
31. He, L., Chen, K., Zhao, J., Wang, Y., Pei, E., Chen, H., Jiang, J., Zhang, S., Zhang, J., Wang, Z., He, T., Tiwari, P.: LMVD: A large-scale multimodal vlog dataset for depression detection in the wild. Information Fusion **126**, 103632 (2026). https://doi.org/https://doi.org/10.1016/j.inffus.2025.103632, https://www.sciencedirect.com/science/article/pii/S1566253525007043
32. He, L., Jiang, D., Sahli, H.: Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding. IEEE Transactions on Multimedia **21**(6), 1476–1486 (2019). https://doi.org/10.1109/TMM.2018.2877129
33. He, L., Li, Z., Tiwari, P., Cao, C., Xue, J., Zhu, F., Wu, D.: Depressformer: Leveraging video swin transformer and fine-grained local features for depression scale estimation. Biomedical Signal Processing and Control **96**, 106490 (2024)
34. He, L., Li, Z., Tiwari, P., Zhu, F., Wu, D.: LSCAformer: Long and short-term cross-attention-aware transformer for depression recognition from video sequences. Biomedical Signal Processing and Control **98**, 106767 (2024). https://doi.org/https://doi.org/10.1016/j.bspc.2024.106767, https://www.sciencedirect.com/science/article/pii/S1746809424008255
35. He, L., Yang, W., Zhao, J., Chen, H., Jiang, D.: FedDAAM: Federated domain adversarial learning with attention mechanism for privacy preserving multimodal depression assessment. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2025). https://doi.org/10.1109/TCSVT.2025.3609776

36. He, L., Zhao, J., Zhang, J., Jiang, J., Qi, S., Wang, Z., Wu, D.: LMTformer: facial depression recognition with lightweight multi-scale transformer from videos. Applied Intelligence **55**(3) (Dec 2024)
37. Jain, V., Crowley, J.L., Dey, A.K., Lux, A.: Depression estimation using audio-visual features and fisher vector encoding. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 87–91. Association for Computing Machinery, New York, NY, USA (2014)
38. Jaiswal, S., Song, S., Valstar, M.: Automatic prediction of depression and anxiety from behaviour and personality attributes. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 1–7 (2019). https://doi.org/10.1109/ACII.2019.8925456
39. Jan, A., Meng, H., Gaus, Y.F.A., Zhang, F., Turabzadeh, S.: Automatic depression scale prediction using facial expression dynamics and regression. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 73–80. Association for Computing Machinery, New York, NY, USA (2014)
40. Jan, A., Meng, H., Gaus, Y.F.B.A., Zhang, F.: Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. IEEE Transactions on Cognitive and Developmental Systems **10**(3), 668–680 (2018)
41. Kächele, M., Glodek, M., Zharkov, D., Meudt, S., Schwenker, F.: Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. In: International Conference on Pattern Recognition Applications and Methods (ICPRAM). pp. 671–678 (2014)
42. Kächele, M., Schels, M., Schwenker, F.: Inferring depression and affect from application dependent meta knowledge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. p. 41–48. AVEC '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2661806.2661813, https://doi.org/10.1145/2661806.2661813
43. Kaya, H., Çilli, F., Salah, A.A.: Ensemble CCA for continuous emotion prediction. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 19–26. Association for Computing Machinery, New York, NY, USA (2014)
44. Kaya, H., Salah, A.A.: Eyes Whisper Depression: A CCA based Multimodal Approach. In: Proceedings of the 22nd ACM International Conference on Multimedia. p. 961–964. MM '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2647868.2654978, https://doi.org/10.1145/2647868.2654978
45. Khan, M.T., Cao, Y., Shafait, F., Jun, W.: ERBMA-Net: Enhanced random binary multilevel attention network for facial depression recognition. IEEE Transactions on Computational Social Systems pp. 1–19 (2025). https://doi.org/10.1109/TCSS.2025.3596047
46. Li, S., Shao, Z., Qin, R., Huang, Y., Liang, P., Li, X., Jiang, Y., Deng, Y., Liu, T., Tan, X.: Audio-visual Feature Disentanglement and Fusion Network for Automatic Depression Severity Prediction. IEEE Transactions on Affective Computing pp. 1–15 (2025). https://doi.org/10.1109/TAFFC.2025.3611238
47. Li, X., Yi, X., Lu, L., Wang, H., Zheng, Y., Han, M., Wang, Q.: TSFFM: Depression detection based on latent association of facial and body expressions. Computers in Biology and Medicine **168**, 107805 (2024). https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.107805, https://www.sciencedirect.com/science/article/pii/S0010482523012702
48. Li, Y., Yang, X., Zhao, M., Wang, Z., Yao, Y., Qian, W., Qi, S.: FPT-Former: A Flexible Parallel Transformer of Recognizing Depression by Using Audiovisual Expert-Knowledge-Based Multimodal Measures. International Journal of

Intelligent Systems **2024**(1), 1564574 (2024). https://doi.org/https://doi.org/10.1155/2024/1564574, https://onlinelibrary.wiley.com/doi/abs/10.1155/2024/1564574

49. Li, Y., Qu, S., Zhou, X.: Conformal Depression Prediction. IEEE Transactions on Affective Computing **16**(3), 1814–1824 (2025). https://doi.org/10.1109/TAFFC.2025.3542023
50. Li, Y., Zhang, Z., Zhou, X.: Fair Uncertainty Quantification for Depression Prediction (2025), https://arxiv.org/abs/2505.04931
51. Li, Y., Liu, Z., Zhou, L., Yuan, X., Shangguan, Z., Hu, X., Hu, B.: A facial depression recognition method based on hybrid multi-head cross attention network. Frontiers in Neuroscience **Volume 17 - 2023** (2023). https://doi.org/10.3389/fnins.2023.1188434, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2023.1188434
52. Lin, Z., Wang, Y., Zhou, Y., Du, F., Yang, Y.: MLM-EOE: Automatic depression detection via sentimental annotation and multi-expert ensemble. IEEE Transactions on Affective Computing pp. 1–18 (2025). https://doi.org/10.1109/TAFFC.2025.3585599
53. Lin, Z., Wang, Y., Zhou, Y., Du, F., Yang, Y.: STE-Mamba: Automated multimodal depression detection through emotional analysis and spatio-temporal information ensemble. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10889512
54. Liu, J., Shang, Y., Yang, M., Lu, J., Shao, Z., Ding, H., Liu, T.: A multi-level and multi-scale context refinement network for video-based depression recognition. In: 2025 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2025). https://doi.org/10.1109/IJCNN64981.2025.11228686
55. Liu, J., Shang, Y., Yang, M., Shao, Z., Ding, H., Liu, T.: Attention-guided bi-direction temporal-aware network for speech-based depression recognition. Digital Signal Processing **166**, 105359 (2025). https://doi.org/https://doi.org/10.1016/j.dsp.2025.105359, https://www.sciencedirect.com/science/article/pii/S1051200425003811
56. Liu, J., Shang, Y., Yang, M., Shao, Z., Ding, H., Liu, T.: CFG-Mamba: Cross frame group mamba for video-based depression recognition. Biomedical Signal Processing and Control **110**, 108113 (2025). https://doi.org/https://doi.org/10.1016/j.bspc.2025.108113, https://www.sciencedirect.com/science/article/pii/S174680942500624X
57. Liu, J., Shang, Y., Yang, M., Shao, Z., Lu, J., Liu, T.: MFMamba: A multimodal fusion state space model for depression recognition. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10888951
58. Liu, Z., Yuan, X., Li, Y., Shangguan, Z., Zhou, L., Hu, B.: PRA-Net: Part-and-relation attention network for depression recognition from facial expression. Computers in Biology and Medicine **157**, 106589 (2023). https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.106589, https://www.sciencedirect.com/science/article/pii/S0010482523000549
59. Ma, X., Huang, D., Wang, Y., Wang, Y.: Cost-sensitive two-stage depression prediction using dynamic visual clues. In: Asian Conference on Computer Vision (ACCV). pp. 338–351. Springer International Publishing, Cham (2017)
60. Ma, X., Yang, H., Chen, Q., Huang, D., Wang, Y.: DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge.

p. 35–42. AVEC '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2988257.2988267, https://doi.org/10.1145/2988257.2988267

61. Carneiro de Melo, W., Granger, E., Lopez, M.B.: Encoding temporal information for automatic depression recognition from facial analysis. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1080–1084 (2020). https://doi.org/10.1109/ICASSP40776.2020.9054375
62. de Melo, W.C., Granger, E., Hadid, A.: Combining Global and Local Convolutional 3D Networks for Detecting Depression from Facial Expressions. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–8 (2019). https://doi.org/10.1109/FG.2019.8756568
63. de Melo, W.C., Granger, E., Hadid, A.: Depression detection based on deep distribution learning. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 4544–4548 (2019). https://doi.org/10.1109/ICIP.2019.8803467
64. de Melo, W.C., Granger, E., Hadid, A.: A deep multiscale spatiotemporal network for assessing depression from facial dynamics. IEEE Transactions on Affective Computing **13**(3), 1581–1592 (2022)
65. de Melo, W.C., Granger, E., López, M.B.: MDN: A deep maximization-differentiation network for spatio-temporal depression detection. IEEE Transactions on Affective Computing **14**(1), 578–590 (2023)
66. Mendiratta, A., Scibelli, F., Esposito, A.M., Capuano, V., Likforman-Sulem, L., Maldonato, M.N., Vinciarelli, A., Esposito, A.: Automatic Detection of Depressive States from Speech, pp. 301–314. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-56904-8_29, https://doi.org/10.1007/978-3-319-56904-8_29
67. Meng, H., Huang, D., Wang, H., Yang, H., AI-Shuraifi, M., Wang, Y.: Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). p. 21–30. Association for Computing Machinery, New York, NY, USA (2013)
68. Mitra, V., Shriberg, E., McLaren, M., Kathol, A., Richey, C., Vergyri, D., Graciarena, M.: The SRI AVEC-2014 Evaluation System. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 93–101. Association for Computing Machinery, New York, NY, USA (2014)
69. Mobram, S., Vali, M.: Depression detection based on linear and nonlinear speech features in I-vector/SVDA framework. Computers in Biology and Medicine **149**, 105926 (2022). https://doi.org/https://doi.org/10.1016/j.compbiomed.2022.105926, https://www.sciencedirect.com/science/article/pii/S0010482522006679
70. Mou, L., Zhen, S., Mao, S., Ma, N.: Disentangled Representation Learning via Transformer with Graph Attention Fusion for Depression Detection. In: Proceedings of the 1st International Workshop on Cognition-Oriented Multimodal Affective and Empathetic Computing. p. 20–29. CogMAEC '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746277.3760407, https://doi.org/10.1145/3746277.3760407
71. Niu, M., Chen, K., Chen, Q., Yang, L.: HCAG: A Hierarchical Context-Aware Graph Attention Model for Depression Detection. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4235–4239 (2021). https://doi.org/10.1109/ICASSP39728.2021.9413486

72. Niu, M., He, L., Li, Y., Liu, B.: Depressioner: Facial dynamic representation for automatic depression level prediction. Expert Systems with Applications **204**, 117512 (2022)
73. Niu, M., Li, M., Fu, C.: PointTransform networks for automatic depression level prediction via facial keypoints. Knowledge-Based Systems **297**, 111951 (2024). https://doi.org/https://doi.org/10.1016/j.knosys.2024.111951, https://www.sciencedirect.com/science/article/pii/S0950705124005859
74. Niu, M., Li, Y., Tao, J., Zhou, X., Schuller, B.W.: DepressionMLP: A multi-layer perceptron architecture for automatic depression level prediction via facial keypoints and action units. IEEE Transactions on Circuits and Systems for Video Technology **34**(9), 8924–8938 (2024). https://doi.org/10.1109/TCSVT.2024.3382334
75. Niu, M., Liu, B., Tao, J., Li, Q.: A time-frequency channel attention and vectorization network for automatic depression level prediction. Neurocomputing **450**, 208–218 (2021). https://doi.org/https://doi.org/10.1016/j.neucom.2021.04.056, https://www.sciencedirect.com/science/article/pii/S0925231221005981
76. Niu, M., Tao, J., Li, Y., Qin, Y., Li, Y.: WavDepressionNet: Automatic Depression Level Prediction via Raw Speech Signals. IEEE Transactions on Affective Computing **15**(1), 285–296 (2024). https://doi.org/10.1109/TAFFC.2023.3272553
77. Niu, M., Tao, J., Liu, B., Fan, C.: Automatic Depression Level Detection via ℓ_p-Norm Pooling. In: Interspeech 2019. pp. 4559–4563 (2019). https://doi.org/10.21437/Interspeech.2019-1617
78. Niu, M., Tao, J., Liu, B., Huang, J., Lian, Z.: Multimodal spatiotemporal representation for automatic depression level detection. IEEE Transactions on Affective Computing **14**(1), 294–307 (2023)
79. Niu, M., Wang, X., Gong, J., Liu, B., Tao, J., Schuller, B.W.: Depression scale dictionary decomposition framework for multimodal automatic depression level prediction. IEEE Transactions on Circuits and Systems for Video Technology **35**(6), 6195–6210 (2025). https://doi.org/10.1109/TCSVT.2025.3533480
80. Niu, M., Zhao, Z., Tao, J., Li, Y., Schuller, B.W.: Selective element and two orders vectorization networks for automatic depression severity diagnosis via facial changes. IEEE Transactions on Circuits and Systems for Video Technology **32**(11), 8065–8077 (2022). https://doi.org/10.1109/TCSVT.2022.3182658
81. Niu, M., Zhao, Z., Tao, J., Li, Y., Schuller, B.W.: Dual attention and element recalibration networks for automatic depression level prediction. IEEE Transactions on Affective Computing **14**(3), 1954–1965 (2023). https://doi.org/10.1109/TAFFC.2022.3177737
82. Othmani, A., Kadoch, D., Bentounes, K., Rejaibi, E., Alfred, R., Hadid, A.: Towards robust deep neural networks for affect and depression recognition from speech. In: Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R. (eds.) Pattern Recognition. ICPR International Workshops and Challenges. pp. 5–19. Springer International Publishing, Cham (2021)
83. Pan, Y., Jiang, J., Jiang, K., Liu, X.: Disentangled-multimodal privileged knowledge distillation for depression recognition with incomplete multimodal data. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 5712–5721. MM '24, Association for Computing Machinery, New York, NY, USA (2024)

84. Pan, Y., Jiang, J., Jiang, K., Wu, Z., Yu, K., Liu, X.: OpticalDR: A deep optical imaging model for privacy-protective depression recognition. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1303–1312 (2024). `https://doi.org/10.1109/CVPR52733.2024.00130`
85. Pan, Y., Shang, Y., Liu, T., Shao, Z., Guo, G., Ding, H., Hu, Q.: Spatial–temporal attention network for depression recognition from facial videos. Expert Systems with Applications **237**, 121410 (2024)
86. Pan, Y., Shang, Y., Shao, Z., Liu, T., Guo, G., Ding, H.: Integrating deep facial priors into landmarks for privacy preserving multimodal depression recognition. IEEE Transactions on Affective Computing pp. 1–8 (2023)
87. Pan, Y., Shang, Y., Wang, W., Shao, Z., Han, Z., Liu, T., Guo, G., Ding, H.: Multi-feature deep supervised voiceprint adversarial network for depression recognition from speech. Biomedical Signal Processing and Control **89**, 105704 (2024)
88. Peng, Z., Dang, J., Unoki, M., Akagi, M.: Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech. Neural Networks **140**, 261–273 (2021). `https://doi.org/https://doi.org/10.1016/j.neunet.2021.03.027`, `https://www.sciencedirect.com/science/article/pii/S0893608021001155`
89. Pérez Espinosa, H., Escalante, H.J., Villaseñor Pineda, L., Montes-y Gómez, M., Pinto-Avedaño, D., Reyez-Meza, V.: Fusing affective dimensions and audio-visual features from segmented video for depression recognition: INAOE-BUAP's participation at AVEC'14 challenge. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 49–55. Association for Computing Machinery, New York, NY, USA (2014)
90. Qureshi, S.A., Saha, S., Hasanuzzaman, M., Dias, G.: Multitask representation learning for multimodal estimation of depression level. IEEE Intelligent Systems **34**(5), 45–52 (2019)
91. Rathi, S., Kaur, B., Agrawal, R.K.: Enhanced depression detection from facial cues using univariate feature selection techniques. In: Pattern Recognition and Machine Intelligence. pp. 22–29. Springer International Publishing, Cham (2019)
92. Ray, A., Kumar, S., Reddy, R., Mukherjee, P., Garg, R.: Multi-level Attention Network using Text, Audio and Video for Depression Prediction. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. p. 81–88. AVEC '19, Association for Computing Machinery, New York, NY, USA (2019). `https://doi.org/10.1145/3347320.3357697`, `https://doi.org/10.1145/3347320.3357697`
93. Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., Othmani, A.: MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. Biomedical Signal Processing and Control **71**, 103107 (2022). `https://doi.org/https://doi.org/10.1016/j.bspc.2021.103107`, `https://www.sciencedirect.com/science/article/pii/S1746809421007047`
94. Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.M., Song, S., Liu, S., Zhao, Z., Mallol-Ragolta, A., Ren, Z., Soleymani, M., Pantic, M.: AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 3–12. Association for Computing Machinery, New York, NY, USA (2019)
95. Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., Pantic, M.: AVEC 2017: Real-Life Depression, and

Affect Recognition Workshop and Challenge. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 3–9. Association for Computing Machinery, New York, NY, USA (2017)

96. Rutowski, T., Harati, A., Lu, Y., Shriberg, E.: Optimizing speech-input length for speaker-independent depression classification. In: Interspeech 2019. pp. 3023–3027 (2019). `https://doi.org/10.21437/Interspeech.2019-3095`
97. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). `https://doi.org/10.1109/ICCV.2017.74`
98. Senoussaoui, M., Sarria-Paja, M., Santos, J.a.F., Falk, T.H.: Model fusion for multimodal depression classification and level detection. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. p. 57–63. AVEC '14, Association for Computing Machinery, New York, NY, USA (2014). `https://doi.org/10.1145/2661806.2661819`, `https://doi.org/10.1145/2661806.2661819`
99. Shang, Y., Pan, Y., Jiang, X., Shao, Z., Guo, G., Liu, T., Ding, H.: LQGDNet: A local quaternion and global deep network for facial depression recognition. IEEE Transactions on Affective Computing **14**(3), 2557–2563 (2023)
100. Sitzmann, V., Diamond, S., Peng, Y., Dun, X., Boyd, S., Heidrich, W., Heide, F., Wetzstein, G.: End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. ACM Trans. Graph. **37**(4) (Jul 2018). `https://doi.org/10.1145/3197517.3201333`, `https://doi.org/10.1145/3197517.3201333`
101. Song, S., Jaiswal, S., Shen, L., Valstar, M.: Spectral representation of behaviour primitives for depression analysis. IEEE Transactions on Affective Computing **13**(2), 829–844 (2022)
102. Song, S., Shen, L., Valstar, M.: Human Behaviour-Based Automatic Depression Analysis Using Hand-Crafted Statistics and Deep Learned Spectral Features. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 158–165 (2018). `https://doi.org/10.1109/FG.2018.00032`
103. Stepanov, E.A., Lathuilière, S., Chowdhury, S.A., Ghosh, A., Vieriu, R.L., Sebe, N., Riccardi, G.: Depression severity estimation from multiple modalities. In: 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom). pp. 1–6 (2018)
104. Sun, B., Zhang, Y., He, J., Yu, L., Xu, Q., Li, D., Wang, Z.: A random forest regression method with selected-text feature for depression assessment. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 61–68. Association for Computing Machinery, New York, NY, USA (2017)
105. Syed, Z.S., Sidorov, K., Marshall, D.: Depression severity prediction based on biomarkers of psychomotor retardation. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 37–43. Association for Computing Machinery, New York, NY, USA (2017)
106. Tao, Y., Yang, M., Shen, H., Yang, Z., Weng, Z., Hu, B.: Classifying anxiety and depression through llms virtual interactions: A case study with chatgpt. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2259–2264 (2023). `https://doi.org/10.1109/BIBM58861.2023.10385305`
107. Tasnim, M., Stroulia, E.: Detecting depression from voice. In: Advances in Artificial Intelligence. pp. 472–478. Springer International Publishing, Cham (2019)
108. Uddin, M.A., Joolee, J.B., Lee, Y.K.: Depression Level Prediction Using Deep Spatiotemporal Features and Multilayer Bi-LTSM. IEEE Transactions on Affec-

tive Computing **13**(2), 864–870 (2022). https://doi.org/10.1109/TAFFC.2020.2970418

109. Uddin, M.A., Joolee, J.B., Sohn, K.A.: Deep Multi-Modal Network Based Automated Depression Severity Estimation. IEEE Transactions on Affective Computing **14**(3), 2153–2167 (2023). https://doi.org/10.1109/TAFFC.2022.3179478
110. Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M.: AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. p. 3–10. AVEC '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2988257.2988258, https://doi.org/10.1145/2988257.2988258
111. Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M.: AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 3–10. Association for Computing Machinery, New York, NY, USA (2014)
112. Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M.: AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 3–10. Association for Computing Machinery, New York, NY, USA (2013)
113. Wang, R., Huang, J., Zhang, J., Liu, X., Zhang, X., Liu, Z., Zhao, P., Chen, S., Sun, X.: FacialPulse: An Efficient RNN-based Depression Detection via Temporal Facial Landmarks. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 311–320. MM '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664647.3681546, https://doi.org/10.1145/3664647.3681546
114. Wang, Y., Lin, Z., Yang, C., Zhou, Y., Yang, Y.: Automatic depression recognition with an ensemble of multimodal spatio-temporal routing features. IEEE Transactions on Affective Computing **16**(3), 1855–1872 (2025). https://doi.org/10.1109/TAFFC.2025.3543226
115. Wang, Z., Chen, L., Wang, L., Diao, G.: Recognition of audio depression based on convolutional neural network and generative antagonism network model. IEEE Access **8**, 101181–101191 (2020)
116. Wen, L., Li, X., Guo, G., Zhu, Y.: Automated depression diagnosis based on facial dynamic analysis and sparse coding. IEEE Transactions on Information Forensics and Security **10**(7), 1432–1441 (2015). https://doi.org/10.1109/TIFS.2015.2414392
117. Williamson, J.R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., Kung, H.T., Dagli, C., Quatieri, T.F.: Detecting depression using vocal, facial and semantic communication cues. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). p. 11–18. Association for Computing Machinery, New York, NY, USA (2016)
118. Williamson, J.R., Quatieri, T.F., Helfer, B.S., Ciccarelli, G., Mehta, D.D.: Vocal and facial biomarkers of depression based on motor incoordination and timing. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 65–72. Association for Computing Machinery, New York, NY, USA (2014)
119. Williamson, J.R., Quatieri, T.F., Helfer, B.S., Horwitz, R., Yu, B., Mehta, D.D.: Vocal biomarkers of depression based on motor incoordination. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion

Challenge. p. 41–48. Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2512530.2512531, https://doi.org/10.1145/2512530.2512531

120. Wu, Z., Zhou, L., Li, S., Fu, C., Lu, J., Han, J., Zhang, Y., Zhao, Z., Song, S.: DepMGNN: matrixial graph neural network for video-based automatic depression assessment. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25, AAAI Press (2025). https://doi.org/10.1609/aaai.v39i2.32153, https://doi.org/10.1609/aaai.v39i2.32153
121. Xu, X., Peng, H., Bhuiyan, M.Z.A., Hao, Z., Liu, L., Sun, L., He, L.: Privacy-preserving federated depression detection from multisource mobile health data. IEEE Transactions on Industrial Informatics **18**(7), 4788–4797 (2022). https://doi.org/10.1109/TII.2021.3113708
122. Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M.C., Sahli, H.: Multimodal measurement of depression using deep learning models. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 53–59. Association for Computing Machinery, New York, NY, USA (2017)
123. Yang, L., Sahli, H., Xia, X., Pei, E., Oveneke, M.C., Jiang, D.: Hybrid depression classification and estimation from audio video and text information. In: ACM International Workshop on Audio/Visual Emotion Challenge (AVEC). pp. 45–51. Association for Computing Machinery, New York, NY, USA (2017)
124. Yang, M., Shang, Y., Liu, J., Shao, Z., Liu, T., Ding, H., Li, H.: LMS-VDR: Integrating landmarks into multi-scale hybrid net for video-based depression recognition. In: Lin, Z., Cheng, M.M., He, R., Ubul, K., Silamu, W., Zha, H., Zhou, J., Liu, C.L. (eds.) Pattern Recognition and Computer Vision. pp. 299–312. Springer Nature Singapore, Singapore (2025)
125. Yoon, J., Kang, C., Kim, S., Han, J.: D-vlog: Multimodal vlog dataset for depression detection. Proceedings of the AAAI Conference on Artificial Intelligence **36**(11), 12226–12234 (Jun 2022). https://doi.org/10.1609/aaai.v36i11.21483, https://ojs.aaai.org/index.php/AAAI/article/view/21483
126. Yuan, X., Liu, Z., Chen, Q., Li, G., Ding, Z., Shangguan, Z., Hu, B.: Combining informative regions and clips for detecting depression from facial expressions. Cognitive Computation (Jun 2023). https://doi.org/10.1007/s12559-023-10157-0, https://doi.org/10.1007/s12559-023-10157-0
127. Zhang, L., Driscol, J., Chen, X., Hosseini Ghomi, R.: Evaluating acoustic and linguistic features of detecting depression sub-challenge dataset. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. p. 47–53. AVEC '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3347320.3357693, https://doi.org/10.1145/3347320.3357693
128. Zhang, P., Wu, M., Dinkel, H., Yu, K.: DEPA: Self-supervised audio embedding for depression detection. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 135–143. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3479236, https://doi.org/10.1145/3474085.3479236
129. Zhang, S., Zhang, X., Zhao, X., Fang, J., Niu, M., Zhao, Z., Yu, J., Tian, Q.: MTDAN: A Lightweight Multi-Scale Temporal Difference Attention Networks for Automated Video Depression Detection. IEEE Transactions on Affective Computing **15**(3), 1078–1089 (2024). https://doi.org/10.1109/TAFFC.2023.3312263

130. Zhang, X., Liu, H., Xu, K., Zhang, Q., Liu, D., Ahmed, B., Epps, J.: When LLMs meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 146–158. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.8, https://aclanthology.org/2024.emnlp-main.8/
131. Zhao, J., Zhang, L., Cui, Y., Shi, J., He, L.: A novel image-data-driven and frequency-based method for depression detection. Biomedical Signal Processing and Control **86**, 105248 (2023). https://doi.org/https://doi.org/10.1016/j.bspc.2023.105248, https://www.sciencedirect.com/science/article/pii/S174680942300681X
132. Zhao, Z., Bao, Z., Zhang, Z., Deng, J., Cummins, N., Wang, H., Tao, J., Schuller, B.: Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders. IEEE Journal of Selected Topics in Signal Processing **14**(2), 423–434 (2020). https://doi.org/10.1109/JSTSP.2019.2955012
133. Zhao, Z., Li, Q., Cummins, N., Liu, B., Wang, H., Tao, J., Schuller, B.W.: Hybrid Network Feature Extraction for Depression Assessment from Speech. In: Interspeech. pp. 4956–4960 (2020)
134. Zhao, Z., Liu, S., Niu, M., Wang, H., Schuller, B.W.: Dense coordinate channel attention network for depression level estimation from speech. In: Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part XIII. p. 402–413. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-78201-5_26, https://doi.org/10.1007/978-3-031-78201-5_26
135. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319
136. Zhou, X., Jin, K., Shang, Y., Guo, G.: Visually interpretable representation learning for depression recognition from facial images. IEEE Transactions on Affective Computing **11**(3), 542–552 (2020)
137. Zhou, X., Wei, Z., Xu, M., Qu, S., Guo, G.: Facial Depression Recognition by Deep Joint Label Distribution and Metric Learning. IEEE Transactions on Affective Computing **13**(3), 1605–1618 (2022). https://doi.org/10.1109/TAFFC.2020.3022732
138. Zhu, Y., Shang, Y., Shao, Z., Guo, G.: Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. IEEE Transactions on Affective Computing **9**(4), 578–584 (2018). https://doi.org/10.1109/TAFFC.2017.2650899
139. Zou, B., Han, J., Wang, Y., Liu, R., Zhao, S., Feng, L., Lyu, X., Ma, H.: Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. IEEE Transactions on Affective Computing **14**(4), 2823–2838 (2023). https://doi.org/10.1109/TAFFC.2022.3181210