Multi-modal Depression Detection: A Survey and Quad Modal -DepNet Framework

Kaveri[1*], NagnathBiradar[2]

[1]*Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Bidar, Karnataka, India Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India*

Email:{kaverigadgikar11,nagnathbiradar}@gmail.com

Abstract

Major Depressive Disorder (MDD) remains among the most under-diagnosed conditions globally, largely because existing automated screening tools capture only a narrow slice of observable symptoms. Contemporary Automated Depression Detection (ADD) pipelines overwhelmingly confine themselves to textual and acoustic cues, leaving the rich behavioral information embedded in facial morphology and temporal visual dynamics unexploited. This paper introduces QuadModal-DepNet, an end-to- end deep learning architecture that jointly processes four complementary data streams—natural language text, speech audio, static facial imagery, and continuous video recordings—through modality-specific encoders followed by a cross-modal attention fusion module. Textual features are extracted via a fine-tuned RoBERTa encoder, acoustic representations are obtained from a WavLM backbone, facial affect descriptors are generated by a ResNet-50 model pre-trained on the AffectNet corpus, and spatiotemporal video embeddings are produced by a Video Swin Transformer. An adaptive gated fusion layer dynamically weights modality contributions based on signal quality and relevance, while Gradient-weighted Class Activation Mapping (Grad-CAM) and SHAP- based explanations provide clinician-interpretable decision rationale. Preliminary experimental design on the DAIC-WOZ and E- DAIC benchmarks targets improvements in F1-score over prevailing bimodal baselines. The proposed system addresses critical shortcomings identified in the recent literature, including limited modality coverage, absence of visual biomarkers, and the lack of explainability necessary for responsible clinical deployment.

Index Terms—Multimodal Depression Detection, Facial Affect Recognition, Video-based Behavioral Analysis, Cross-Modal Attention Fusion, Explainable Artificial Intelligence, Mental Health Informatics.

I. INTRODUCTION

Depression constitutes a debilitating psychiatric condition marked by sustained emotional despondency, diminished engagement with previously rewarding activities, and pervasive cognitive impairment. Statistical data from the World Health Organization indicate that roughly 280 million individuals across the globe are affected, with annual suicide mortality surpassing 700,000 cases [1]. Conventional diagnostic workflows-hinge on structured clinical interviews and standardized self-report instruments, notably the Beck Depression Inventory (BDI-II) [2] and the Patient Health Questionnaire-9 (PHQ-9) [3]. While these instruments remain the clinical gold standard, they are inherently constrained by the subjective nature of self-disclosure, the pervasive stigma surrounding mental illness, and a growing global shortage of qualified psychiatric professionals [4].

The convergence of natural language processing, acoustic signal analysis, and computer vision with modern deep learning paradigms has spurred substantial interest in Automated Depression Detection (ADD) systems. These systems aspire to furnish objective, reproducible, and scalable screening mechanisms that complement clinician judgment. A significant body of work has concentrated on extracting depression indicators from textual and auditory channels. Li et al. [5] recently published a comprehensive systematic review covering 65 studies from 2018 through 2024, evaluating ADD methodologies that harness text and audio modalities .Their analysis catalogued advances in feature extraction, data augmentation, and multimodal fusion, yet explicitly acknowledged that the visual modality—a potent source of non-verbal behavioral cues—was excluded from their investigative scope.

This exclusion is significant from a clinical stand point. Psycho motor retardation, blunted facial expressivity, reduced eye-contact frequency, and altered head-movement patterns are well-documented diagnostic indicators in the DSM-5 criteria for MDD [6]. Neuro scientific investigations have established that depressed individuals exhibit measurably reduced Facial Action Unit (AU) activation, particularly in the upper face regions responsible for expressing positive affect [7]. Furthermore, temporal patterns observable in video—such as gaze aversion dynamics, postural rigidity, and gesture frequency—provide longitudinal behavioral signatures that static images alone cannot capture [8].

Motivated by these observations, this paper proposes QuadModal-DepNet, a unified deep learning frame work that extends the prevailing text-audio paradigm by incorporating two additional visual modalities: static facial imagery and continuous video recordings. The core contributions of this work are articulated as follows:

A systematic analysis of the limitations inherent in existing bimodal (text + audio) ADD approaches, with emphasis on the unrealized potential of visual biomarkers.

1) The architectural design of QuadModal-DepNet, with emphasis on the unrealized potential of visual biomarkers.

II RELATED WORK

A. Existing Text-based Approaches

Early text-based ADD systems employed classical machine learning classifiers such as Support Vector Machines and Random Forests in conjunction with hand-crafted linguistic features derived from tools like the Linguistic Inquiry and Word Count (LIWC) lexicon [9]. The advent of contextual embeddings transformed this landscape. Pre-trained transformer models like BERT [10] and RoBERTa [11] achieved substantial performance gains by encoding bidirectional semantic context. Ji et al. developed Mental RoBERTa [12], a domain-adapted variant that attained an F1 measure of 0.933 on the eRisk benchmark. More recently, large language models including GPT-3.5, GPT-4, and LLaMA have been fine-tuned for depression classification tasks, with Hadzic et al. [13] demonstrating that GPT-4 outperforms earlier models even without task- specific fine-tuning. Despite these advances, text-only systems remain vulnerable to contextual confounding, where interviewer phrasing rather than participant language drives classification decisions [14].

B. Existing Audio-based Approaches

A coustic analysis for depression detection exploits prosodic, spectral, and cepstral features that correlate with psychomotor and affective disruptions characteristic of depressive states. Ma et al. introduced DepAudioNet [15], a CNN-LSTM architecture processing Mel-Frequency Cepstral Coefficients (MFCCs) that established an early benchmark on DAIC-WOZ with an F1 of 0.61. Subsequent adoption of self-supervised learning models—wav2vec 2.0 [16], HuBERT [17], and WavLM [18]—has substantially improved representation quality. Wuetal.[19] demonstrated WavLM superiority over alternative SSL backbones and introduced Sub-Dialogue Shuffling for data augmentation, yield inga 78% relative performance increase. While audio-

only approaches capture vocal manifestations of depression, they remain insensitive to non-verbal visual behaviors that constitute a clinically significant portion of the diagnostic evidence base.

C. Existing Bimodal Approaches

Combining text and audio has yielded improvements over uni modal baselines. Ye et al. [20] achieved F1 of 0.906 through Deep Spectrum and word vector fusion via a Temporal Convolutional Network-Transformer architecture. Wu et al. [19] combined RoBERTa text embeddings with WavLM audio embeddings using attention-guided fusion to reach F1 of 0.83 on DAIC-WOZ. Lam et al. [21] employed topic

Basedaugmentationinapre-trainedTransformer-CNN pipeline achieving F1 of 0.87. Despite these results, all bimodal studies in the systematic review by Li et al. [5] are constrained to the text-audio dyad, foregoing visual modalities entirely.

D. Visual and Video-based Depression Detection

A parallel research stream has explored visual cues for depression assessment. Pampouchidou et al. [22] surveyed vision-based methods, documenting the diagnostic value of facial expressivity analysis. Studies have demonstrated that Action Unit (AU) intensity distributions differ significantly between depressed and non-depressed cohorts, with reduced AU6 (cheek raiser) and AU12 (lip corner puller) activation being particularly discriminative [7]. Video-based approaches extend static facial analysis by capturing temporal dynamics. Uddin et al. [23] employed spatio temporal deep features with multilayer BiLSTM networks, while Niu et al. [24]

Proposed multimodal spatio temporal representations achieving state-of-the-art depression level detection. Jan et al. [25] demonstrated that gaze direction, head pose trajectories, and facial landmark displacement sequences provide complementary temporal information not available from single-frame analysis.

III PROPOSED SYSTEM ARCHITECTURE

QuadModal-DepNetcomprisesfourmodality-specific encoder branches, across-modal attention fusion module,aclassification head, and an explainability layer. The architecture can be summarized by the following component pipeline.

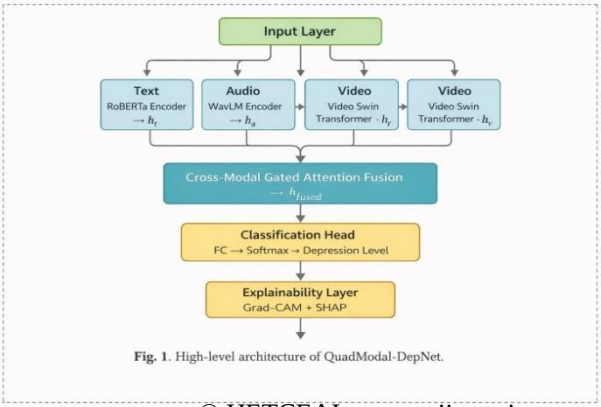

Fig. 1. High-level architecture of QuadModal-DepNet.

A. Text Encoder Module

The textual encoding branch receives either interview transcripts or self-reported text. Raw text is tokenized using the RoBERTa byte-pair encoding tokenizer with a maximum sequence length of 512 tokens. The pre-trained RoBERTa- base model [11] is employed as the backbone encoder, producing a 768-dimensional [CLS] token embedding $h_t \in \Re^{768}$. Domain adaptation is achieved through continued pre- training on a corpus

B. Audio Encoder Module

The acoustic processing pipeline begins with resampling all audio inputs to16kHz mono.The WavLM-Base+model[18] serves as the feature extractor, generating frame-level embeddings from raw wave form input. Following Wuetal.
[19], features are extracted from multiple transformer layers and combined via learned weighted averaging. The resulting sequence of frame embeddings is aggregated into a fixed- dimensional utterance-level representation $h_a \in \Re^{768}$ through an attentive statistics pooling layer that computes weighted mean and standard deviation across temporal
frames. This approach preserves both local acoustic detail and global prosodic contour information that correlates with depressive speech patterns including reduced pitch variability, slower articulation rate, and increased pause duration [27].

C. Facial Image Encoder Module

Static facial analysis targets the extraction of affective and morphological indicators from individual face images. The preprocessing pipeline employs MTCNN [28] for face detection and alignment, producing 224×224 pixel cropped face regions.AResNet-50 backbone [29] pre-trained on the Affect Net dataset [30]—which contains over one million facial images annotated with seven discrete emotion categories and continuous valence-arousal values—is utilized as the feature extractor. The penultimate fully connected layer output yields a 2048-dimensional facial representation. A projection layer maps this to the common embedding dimensionality: $h_f \in \Re^{768}$. For interview-based datasets, multiple frames are sampled at uniform intervals throughout the session, and their embeddings are averaged to produce a session-level facial descriptor. Key facial features captured include AU intensity distributions, emotional valence estimates, and micro-expression patterns that clinical research has associated with affective flattening in depression [7].

D. Video Encoder Module

This gated architecture ensures robust performance even when individual modalities are degraded or absent, addressing a practical concern in clinical deployment where not all four data streams may be simultaneously available.

The video encoding branch captures spatiotemporal behavioral dynamics that evolve over the course of an interaction. Video input is preprocessed to extract participant face-body regions at 2 frames per second, with each clip segmented into non-overlapping 16-frame segments. The Video SwinTransformer[31] pre-trained on Kinetics-400 [32] is employed as the s patio temporal backbone. This architecture applies shifted window multi-head self-attention in three-dimensional (spatial + temporal) patches, enabling efficient modeling of long-range dependencies in facial movement patterns, head pose trajectories, and gestural dynamics. The [CLS] token from the final transformer layer produces the video embedding $h_v \in \Re^{768}$. For sessions exceeding 16 frames, segment-level embeddings are aggregated via temporal attention pooling. The video modality captures behavioral signatures including psycho motor retardation (reduced gestural frequency), gaze aversion patterns, postural rigidity, and reduced facial animation—all of which are clinically associated with depressive episodes [8].

E. Cross-Modal Gated Attention Fusion

Rather than naively concatenating modality-specific embeddings, the proposed framework employs a Cross- Modal Gated Attention Fusion (CMGAF) mechanism that dynamically modulates the contribution of each modality based on signal quality and relevance. The fusion process operates in two stages.

Stage 1 — Pair wise Cross-Attention: Each modality embedding serves alternately as query and key-value pair with every other modality, producing cross-attended representations that capture inter-modal correlations. For modalities *i*and *j*:

$$c_{ij}=\text{softmax}(W_Q h_i \cdot (W_K h_j)^T/\sqrt{d}) \cdot W_V h_j$$

Stage 2 — Gated Aggregation: A gating network learns modality-specific confidence scores $g_m \in [0,1]$ for each modality $m \in \{t, a, f, v\}$, enabling the system to suppress noisy or unreliable modalities gracefully:

$$g_m = \sigma(W_g[h_m;\&bar;c_m]+b_g)$$

$$h_{fused} = \sum_m g_m \odot \&bar;c_m$$

where$\&bar;c_m$is the mean of all cross-attended

Representations involving modality *m*,σdenotes the sigmoid
function, [;] represents concatenation, and ⊙ is element- wise multiplicatio

F: Classification Head and ExplainabilityLayer

The fused representation h_{fused} is passed through a two- layer feed-forward network with GELU activation and dropout regularization, terminating in a softmax output-layer. The system supports both binary classification (depressed vs. non-depressed) and multi-level severity grading (normal, mild, moderate, severe) depending on the target data set labeling scheme.

The explainability layer operates post-hoc and generates modality-specific interpretations. For the visual branches, Gradient-weighted Class Activation Mapping(Grad-CAM)

[33] produces spatial heat maps highlighting facial regions most influential to the prediction. For text and audio, SHAP (SHapley Additive ex Planations) [34] values quantify the contribution of individual tokens and acoustic segments.The gating coefficients g_m themselves provide a global modality importance indicator per sample, which clinicians can inspect to understand which behavioral channel most strongly influenced a given screening outcome.

IV IMPLEMENTATION METHODOLOGY

A. Dataset Selection

The experiments are designed around two publicly available benchmark datasets that DAIC-WOZ [35]: Contains 189 clinical interview sessions conducted by the virtual interviewer Ellie, with audio recordings, manually annotated transcripts, extracted video features, and PHQ-8-based depression labels. This dataset provides the primary evaluation benchmark due to its widespread adoption in the ADD research community.

E-DAIC [36]: An extended version containing 219 sessions with audio, video, and ASR-generated transcripts alongside PHQ-8 labels. The use of automatic transcripts rather than manual annotations makes this dataset representative of realistic deployment conditions.

Additionally, the D-vlog dataset [37]containing YouTube video logs with depression annotations serves as a supplementary evaluation resource for the facial and video modalities.

TABLEI

COMPARISON OF DATASETS USED IN PROPOSED FRAMEWORK

Dataset	ModalitiesAvailable	Subjects	Depressed	Label Scale	Language
DAIC-WOZ[35]	Audio,Video,Text	189	56	PHQ-8	English
E-DAIC[36]	Audio,Video,Text	219	49	PHQ-8	English
D-vlog[37]	Audio,Video	961	-	Binary	English

B. Preprocessing Pipelines

Text Preprocessing: Transcripts undergo sentence tokenization, speaker turn segmentation (retaining only participant utterances to avoid interviewer-induced confounds[14]),lower casing, and special character removal. Tokenization employs the RoBERTa byte-pair encoding tokenizer. Sequences exceeding 512 tokens are processed via a sliding window approach with overlapping context.

Audio Preprocessing: Audio signals are resampled to 16 kHz, normalized to zero mean and unit variance, and segmented into non-overlapping 10-second chunks with 2- second overlap. Voice Activity Detection (VAD) using the SileroVAD model filters non-speech segments. Participant- only audio is isolated through speaker diarization when interviewer audio is present.

Facial Image Pre processing: Video frames are sampled at 1 fps. MTCNN detects and aligns facial regions, producing 224×224 cropped face images. Failed detections (profile views, occlusions) are discarded. Data augmentation includes random horizontal flipping, slight rotation (±15°), and brightness jittering to improve generalization.

Video Preprocessing: Raw video is resized to 224×224 spatial resolution and sampled at 2 fps. Sequences are divided into 16-frame clips (8 seconds each). Spatial augmentation includes random resized cropping and horizontal flipping. Temporal augmentation includes random temporal jittering within clip boundaries.

C. Training Protocol

The training strategy employs a two-phase approach:

Phase1—Modality- Specific Pre-training: Each encoder branch is individually fine-tuned on the target data set with a binary cross-entropy classification objective. This phase establishes strong uni modal baselines and ensures each encoder produces discriminate representations before fusion.

Phase 2 — Joint Multi-modal Fine-tuning:The four pre- trained encoders are connected to the CMGAF module and the classification head. The entire network is trained end-to- end with a combined loss function:

$$L=L_{cls}+\lambda_1 L_{dist}+\lambda_2 L_{reg}$$

where $L_{cls}$is the focal loss [38] to address class imbalance, L_{dist} is a cross-modal distillation loss that encourages alignment between modality embeddings, andL_{reg} is an L2 regularization term on gating parameters to prevent degenerate solutions where a single modality dominates.

Training employs the AdamW optimizer [39] with a cosine annealing learning rate schedule, initial learning rate of 2×10^{-5}, batch size of 8, and gradient accumulation over 4 steps. Mixed-precision training (FP16) is utilized to manage memory constraints. Early stopping with patience of 10 epochs based on

validation F1-score prevents over fitting.

TABLE II

TECHNOLOGY STACK FOR QUAD MODAL-DEPNET IMPLEMENTATION

Method	Modalities	Architecture	ReportedF1	Reference
DepAudioNet	Audio	CNN+LSTM	0.61	[15]
MentalRoBERTa	Text	RoBERTa(fine-tuned)	0.93*	[12]
Yeetal.	Text+Audio	TCN-Transformer	0.906	[20]
Wuetal.	Text+Audio	RoBERTa+WavLM	0.83	[19]
Lametal.	Text+Audio	Transformer+CNN	0.87	[21]
Parketal.	Text+Audio	BERT-CNN+BiLSTM	Acc:0.90	[40]
QuadModal-DepNet(Proposed)	Text+Audio+Image+Video	CMGAFFusion	Target: ≥0.92	This work

V EXPERIMENTAL DESIGN AND EVALUATION

A. Evaluation Metrics

Following standard practice in ADD research, the proposed system is evaluated using the following metrics for binary depression classification: Accuracy, Precision, Recall, F1- score,and Area Under the Receiver Operating Characteristic Curve (AUC- ROC). For multi -levels e verity estimation, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are additionally reported relative to PHQ-8 scores. Cross-validation follows the official DAIC-WOZ train/development/ test partition to ensure comparability with published results.

B. AblationStudyDesign

To quantify the contribution of each modality and architectural component, the following ablation experiments are planned:

- Uni-modal baselines: Text-only, Audio-only, Image- only, Video-only.
- Bimodal combinations: Text+Audio, Text+Image,Text+Video,Audio+Image,Audio+Video, Image+Video.

Trimodal combinations: Text+Audio+Image, Text+Audio+Video,Text+Image+Video,Audio+Image +Video.

- Fullquadmodal:Text+Audio+Image+Video.
- Fusion ablation: CMGAF vs. simple concatenation vs. late fusion averaging.

C.BaselineComparisons

Performance is benchmarked against the following established methods from the literature:

TABLEIII

BASELINESYSTEMSFORCOMPARATIVE EVALUATION

Component	Technology	Purpose
Deep Learning Framework	PyTorch2.x+HuggingFaceTransformers	Model development and training
TextEncoder	RoBERTa-base(HuggingFace)	Contextual text embeddings
AudioEncoder	WavLM-Base+(Microsoft)	Self-supervisedspeechrepresentations
FaceEncoder	ResNet-50(torchvision, Affect Net weights)	Facial affect feature extraction
VideoEncoder	VideoSwinTransformer (MMAction2)	Spatio temporal video embeddings
Face Detection	MTCNN(facenet-pytorch)	Face detectionandalignment
Audio Processing	Librosa,torchaudio,SileroVAD	AudiopreprocessingandVAD
VideoProcessing	OpenCV,decord	Videoframeextraction
Explainability	SHAP, pytorch -grad-cam	Modelinterpretation
Experiment Tracking	Weights&Biases(wandb)	Hyperparameterandmetriclogging
Hardware	NVIDIAA100/RTX4090 GPU(40–24GB)	Acceleratedtraining
DeploymentPrototype	Flask/Streamlit	Web-basedclinicalinterfacedemo

C. ExpectedOutcomes

Based on the theoretical grounding that visual biomarkers provide complementary diagnostic information to linguistic and acoustic cues, the quadmodal configuration is

Hypothesizedtosurpassbimodalbaselinesbyamarginof3– 8 percentage points in F1-score. The gated attention fusion mechanism is expected to outperform naive concatenationby 2–5 percentage points due to its capacity for dynamic modality weighting. The explainability layer is anticipatedto produce clinically coherent saliency maps that highlight diagnostically relevant facial regions (periorbital area,mouth corners) and temporally significant behavioral episodes.

V CONCLUSION

This paper introduced QuadModal-DepNet, acomprehensivedeeplearningframeworkthatextendsthecurrent bimodal (text + audio) paradigm for Automated Depression Detection by integrating static facial imageryd continuous video analysis. The architectural design employs state-of-the-art modality-specific encoders— RoBERTa for text, WavLM for audio,AffectNet-pretrained ResNet-50 for facial images, and Video Swin Transformer for temporal behavioral dynamics—unified through a novel Cross-Modal Gated Attention Fusion mechanism. The framework addresses five critical gaps identified through systematic analysis of the existing literature, most notably the underutilization of visual biomarkers and the absence of clinician-interpreterexplanationsincurrentADDsystems.

The proposed system represents an advancement toward replicating the holistic observational assessment that skilled psychiatrists perform, capturing not only what patients say and how they sound, but also their facial expressions and behavioral dynamics.The experimental protocol is designed to rigorously evaluate the marginal contribution of each visual modality and to demonstrate the superiority of adaptive gated fusion over conventional combination strategies.

REFERENCES

[1] World Health Organization, "Depressive disorder (depression)," WHOFact Sheets, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[2] A. T. Beck, R.A. Steer, and G. K. Brown, "Beck depression inventory-II," Psychological Assessment, 1996.

[3] K.Kroenke,R.L.Spitzer,andJ.B.W.Williams,"ThePHQ-9:Validityof a brief depression severity measure," *Journal of General InternalMedicine*, vol. 16, no. 9, pp. 606–613, 2001.

[4] A. Satiani, J. Niedermier, B. Satiani, and D. P. Svendsen, "Projectedworkforce of psychiatrists in the United States:Apopulationanalysis,"*Psychiatric Services*, vol. 69, no. 6, pp. 710–713, 2018.

[5] Y. Li, S. Kumbale, Y. Chen, T. Surana, E. S. Chng, and C. Guan,"Automated depression detection from text and audio: A systematicreview," *IEEE Journal of Biomedical and Health Informatics*, 2025,DOI:10.1109/JBHI.2025.3570900.

[6] AmericanPsychiatricAssociation,*DiagnosticandStatisticalManualofMental Disorders*, 5th ed. Arlington, VA: APA, 2013.

[7] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T.Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facialactions and vocal prosody," in *Proc. 3rd Int. Conf.AffectiveComputingand Intelligent Interaction*, IEEE, 2009, pp. 1–7.

[8] A. Pampouchidou, P. G. Simos, K. Marias, F. Meriaudeau, F.Yang, M.Pediaditis, and M. Tsiknakis, "Automatic assessment of depressionbased on visual cues: A systematic review," *IEEE Trans. AffectiveComputing*, vol. 10, no. 4, pp. 445–470, 2017.

[9] M. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing neural networksand linguistic metadata for early detection of depression indications intext sequences," *IEEE Trans. Knowledge and Data Eng.*, vol. 32, no. 3, pp.588–601,2018.

[10]J.Devlin,M.-W.Chang,K.Lee,andK.Toutanova,"BERT:Pre-trainingof deep bidirectional transformers for language understanding," *arXivpreprint arXiv:1810.04805*, 2018.

[11]Y.Liu,M.Ott,N.Goyal,J.Du,M.Joshi,D.Chen,O.Levy,M.Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimizedBERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[12]S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria,"MentalBERT: Publicly available pretrained language models formental healthcare," *arXiv preprint arXiv:2110.15621*, 2021.

[13]B.Hadzic,P.Mohammed,M.Danner,J.Ohse,Y.Zhang,Y.Shiban,and M. Rätsch, "Enhancing early depression detection with AI: Acomparative use of NLP models," *SICE Journal of Control,Measurement, and System Integration*, vol. 17, no. 1, pp. 135–143,2024.

[14]S.Burdisso,E.Reyes-Ramírez,E.Villatoro-Tello,F.Sánchez-Vega,P.López-Monroy,andP.Motlicek,"DAIC-WOZ:Onthevalidityofusingthe therapist's prompts in automatic depression detection from clinicalinterviews," *arXiv preprint arXiv:2404.14463*, 2024.

[15]X.Ma,H.Yang,Q.Chen,D.Huang,andY.Wang,"DepAudioNet:Aneffic ientdeepmodelforaudiobaseddepressionclassification,"in*Proc.6th Int. Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 35–42.

[16]A.Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: Aframework for self-supervised learning of speech representations,"*Advances in Neural Information Processing Systems*, vol. 33, pp.12449–12460,2020.

[17]W.-N.Hsu, B. Bolte,Y.-H. H.Tsai, K. Lakhotia, R. Salakhutdinov, and A.Mohamed,"HuBERT:Self-supervisedspeechrepresentationlearningby masked prediction of hidden units," *IEEE/ACM Trans. Audio,Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[18]S.Chen,C.Wang,Z.Chen,Y.Wu,S.Liu,Z.Chen,J.Li,N.Kanda,T.Yoshioka, X. Xiao, *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Selected Topics inSignal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[19]W.Wu,C.Zhang,andP.C.Woodland,"Self-supervisedrepresentationsin speech-based depression detection," in *Proc. ICASSP 2023*, IEEE,2023, pp. 1–5.

[20]J. Ye, Y. Yu, Q. Wang, W. Li, H. Liang, Y. Zheng, and G. Fu, "Multi-modal depression detection based on emotional audio and evaluationtext," *Journal of Affective Disorders*, vol. 295, pp. 904–913, 2021.

[21]G. Lam, H. Dongyan, and W. Lin, "Context-aware deep learning formulti-modaldepressiondetection,"in*Proc.ICASSP2019*,IEEE,2019, pp.3946–3950.

[22]A.Pampouchidou,P.G.Simos,K.Marias,F.Meriaudeau,F.Yang,M.Pediaditis, and M. Tsiknakis, "Automatic assessment of depressionbased on visual cues: A systematic review," *IEEE Trans. AffectiveComputing*, vol. 10, no. 4, pp. 445–470, 2017.

[23]M.A. Uddin, J. B. Joolee, andY.-K. Lee, "Depression level predictionusing deep spatiotemporal features and multilayer Bi-LSTM," *IEEETrans. Affective Computing*, vol. 13, no. 2, pp. 864–870, 2020.

[24]M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodalspatiotemporal representation for automatic depression level detection,"*IEEE Trans. Affective Computing*, vol. 14, no. 1, pp. 294–307, 2020.

[25]A. Jan, H. Meng, Y. F. A. Gaus, and F. Zhang, "Artificial intelligentsystem for automatic depression level analysis through visual and vocalexpressions," *IEEE Trans. Cognitive and Developmental Systems*, vol.10, no. 3, pp. 668–680, 2018.

[26]K.FengandT.Chaspari,"Apilotstudyonclinician-AIcollaborationindiagnosing depression from speech," *arXiv preprint arXiv:2410.18297*,2024.

[27]N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F.Quatieri, "A review of depression and suicide risk assessment usingspeech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[28]K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection andalignment using multitask cascaded convolutional networks," *IEEESignal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[29]K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for imagerecognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[30]A.Mollahosseini,B.Hasani,andM.H.Mahoor,"AffectNet:Adatabase forfacialexpression,valence,andarousalcomputinginthewild,"*IEEETrans. Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019.