Multimodal Depression Detection: A Survey

Palash Moon, Pushpak Bhattacharyya
IIT Bombay
{palash, pb}@cse.iitb.ac.in

Abstract

In the recent past, mental health has become a global concern. COVID-19 has further caused a rapid surge in depression. Depression is a serious mental illness that is impacting the lives of individuals of all ages all around the world. Depression affects a person's physiological well-being as well as their emotional state. Nowadays, depression is the most common element experienced by human beings irrespective of their age factor and professional life. To detect the depression status among the persons, the system uses different approaches by using the sensor technology. The automatic identification of depression at early stages or immediately helps clinical studies to cure people accurately. In various research, the system aims to identify depression using facial expressions, voice, and live video capturing, by analysing their tweets, status, and posts on social media. Most of the existing research works focus on unimodal development which focuses on the single component analysis but the proposed research aims to focus on the multimodal with a fusion of different modalities of learning approaches involved in the detection of depression, this survey provides an overview of numerous methodologies that have been created to employ emotion recognition to analyse depression.

1 Introduction

Depression is among the most common psychiatric disorders and a leading contributor to the global disease burden. Severe depression is a type of mood disorder that can persist for weeks, months, or even years, is accompanied by distress and disability, and it has an impact on a person's ability to carry out regular tasks. Depression is the world's fourth most frequent mental illness, according to the World Health Organization (WHO), and it is expected to overtake it as the top cause by 2020. Furthermore, according to recent WHO research, 350 million individuals globally are suffering from depression. Depressed people have a 30 times higher risk of suicide than the overall population. Machine learning (ML) is trending today because technology acquires the data and ML interprets from the data. ML has applications in many domains: Image Recognition, Healthcare, document analysis and recognition, transportation, speech recognition etc. Among all domains, one of the most studied topics is health care. Works related to assessing mental health have grown interest.

With the latest technology, Data from patients like sensors, wearable devices, and smartphones are stored and monitored/accessed by health professionals to understand the patterns of the brain, heart, mental health etc. which helps to treat them efficiently. Depression detection has received interest from the research community. Although Researchers used different modalities like verbal, non-verbal, textual, physiological cues and other cues, there are no standards designed for assessing depression. Most of the works aim to assess depression, which helps the clinician's decision support system. Ideally, all indicators that are used by clinicians to assess depression must be used by the ML algorithms for diagnosing depression. It is important to understand that each modality does have a sub-modality i.e., non-verbal cues like facial indicators not only include facial expressions but also comprise eye gaze, head pose etc.,

Still, there is little to no physical test used for diagnosis purposes. Standard self-reports and physical interactions are only the ways used by health professionals to detect/diagnose depression. Both ways have their own pitfalls like social stigma, and fear and the later-mentioned method also involves financial burden. Although the research community have been working on this problem over a period of time yet there is no standard designed to address this problem. Depression detection is a vast area of research where work has been done using Verbal

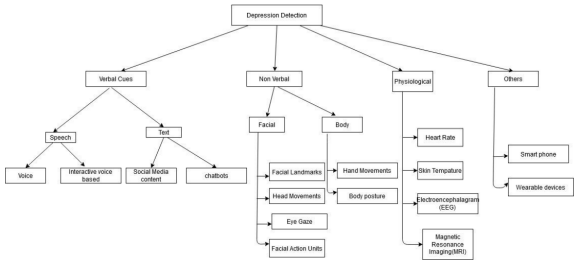

Figure 1: Categories and sub-categories of depression detection research works

cues, Non-Verbal cues and others (Smartphones, wearable devices etc.). Verbal cues include speech processing and text processing. Indeed speech comprises acoustic and linguistic features. Non-verbal includes Facial, and body movements (including hand, and leg movements). Others include smartphone-based, wearable devices-based works. The following Figure 1 demonstrates the categories and subcategories of depression detection research works in detail. This survey article presents detailed works on verbal (voice-based, text-based) and non-verbal only however a brief overview is discussed for the rest.

2 Motivation

The motivation for multimodal depression detection using audio, video, and text stems from the complexity and multifaceted nature of depression, which manifests through various behavioural, emotional, and physiological signals. Traditional methods of depression diagnosis often rely on self-reported questionnaires or clinical interviews, which can be subjective and may not capture the full spectrum of depressive symptoms. By leveraging the rich information available across multiple modalities—such as the nuanced expressions and body language in the video, vocal tone and speech patterns in audio, and sentiment and linguistic features in the text—multimodal approaches offer a more holistic and accurate assessment of an individual's mental state. This integrative method is crucial because it combines diverse data streams to provide a comprehensive understanding of depressive symptoms, leading to improved early detection, continuous monitoring, and personalized interventions.

According to the World Health Organization (WHO), depression affects over 264 million people globally, highlighting the urgent need for effective diagnostic and therapeutic strategies. Additionally, the disparity between the number of patients and available mental health professionals is stark, with the global median estimated to be nine mental health workers per 100,000 people. This shortage is particularly pronounced in low- and middle-income countries, where the ratio can be as low as one per 100,000. The integration of multimodal depression detection systems can help bridge this gap by offering scalable, accessible, and automated tools that assist healthcare providers in diagnosing and monitoring depression. These systems can alleviate the burden on mental health professionals and ensure that more individuals receive timely and accurate mental health care.

3 Dataset

MMDA (Multimodal Depression Analysis) Dataset: This dataset comprises 1025 videos where visual, acoustic, and textual modalities are analyzed to detect signs of depression. Psychologists analyze interview videos, focusing on facial expressions, tone of voice, and interview content to infer mental health conditions. The dataset is primarily in English, enabling detailed analysis of non-verbal cues and linguistic patterns associated with depression.

Multi-modal Open Dataset: Designed for comprehensive depression research, this dataset integrates EEG (Electroencephalography) recordings and spoken language data from diagnosed patients and controls. The dataset size is unspecified but includes carefully diagnosed individuals, providing rich multimodal data for exploring neural correlates and linguistic markers of depression. It primarily supports English-language research into brain wave patterns and language features indicative of mental health states.

A Novel Multi-modal Depression Detection Approach Dataset: This dataset features diverse modalities including EEG and language data collected from real-world subjects, though specific video counts are unspecified. It focuses on capturing depression markers in various contexts, allowing for innovative approaches in multimodal analysis. The dataset's language specifications are not detailed, reflecting a broad scope in cross-modal depression detection research.

LMVD (Audio, Video, Social Media) Dataset: With 1823 videos sourced from multimedia platforms like Sina Weibo, Bilibili, Tiktok, and YouTube, the LMVD dataset offers extensive au-

Table 1: Multimodal Depression Datasets

Category	MMDA	MMOD	MMDD	LMVD	D-Vlog
Modalities	Visual, Acoustic, Textual	EEG, Spoken Language	EEG, Language, etc.	Audio, Video, Social Media	Audio, Face Emotion, Body Landmarks, Gaze
No. of Videos	1025	Not specified	Not specified	1823	961
Language	English	English	Not specified	Chinese, English	English
Extraction Method	Interview videos analyzed by psychologists	Carefully diagnosed patients and controls	Real-world subjects	Multimedia platforms (Sina Weibo, Bilibili, Tiktok, YouTube)	YouTube vlogs

dio, video, and social media data. It supports analysis in both Chinese and English, facilitating cross-cultural comparisons in depression detection. This dataset's extraction method leverages diverse social media interactions, enabling insights into how online behavior and multimedia content reflect mental health indicators.

D-Vlog (Audio, Face Emotion, Body Landmarks, Gaze) Dataset: Comprising 961 YouTube vlogs, the D-Vlog dataset provides multimodal insights into depression using audio, facial emotion analysis, body landmarks, and gaze tracking. This dataset emphasizes real-world expressions and behaviors in English, supporting detailed examination of vloggers' emotional states and physical cues associated with depression.

Each dataset contributes uniquely to the field of multimodal depression detection, offering varied modalities, extraction methods, and linguistic contexts to explore the intricate interplay between biological, psychological, and social factors in understanding and diagnosing depression. These resources enable researchers to develop and validate innovative approaches leveraging machine learning and data-driven insights for mental health assessment.

4 Literature Survey

(Pampouchidou, 2017) Pampouchidou et al. have made an exclusive survey on Visual cues in depression detection, but this work lacks contributions of verbal, physiological and other modalities. (Cummins, 2015) Cummins et al. has reviewed works with only speech modalities. (Malviya, 2018) Aastik Malviya et al. have reviewed only a

few papers related to speech analysis. Yekta Said (Can, 2020) Can et al. reviewed on impact of smartphones and wearable devices on stress. In the field, of mental health analysis using passive sensing through smartphones and wearable devices, (Garcia-Ceja, 2018) EnriqueGarciaCeja et al. have made a survey but lack with the available dataset information. (Lin, 2019) LiuyiLin et al. have made a study on United States Undergraduate students and concluded that Social Media has a lot of impact on depression. (Guntuku, 2017) Sharath Chandra Guntuku et al. have made a review on the analysis of mental health possible with the help of several online environments like Twitter, Facebook, forums etc. (Panicker, 2019) Suja Sreeith Panicker et al. have surveyed physiological data collection and analysis of mental health (stress) using machine learning. Healthcare professionals are reliant on the visual, auditory (acoustic and linguistic) and physiological markers of the subject. Ideally, all the methods that health care professionals are reliant can be used by the machine learning algorithms. Earlier work has proven that features can be extracted from aforementioned modalities. Indeed it is better to recognize that comparison between these modalities becomes a complex task because each modality behaves in its own way. For example, if data sources are considered for facial, verbal, and physiological cues which are images, voice files, X-rays respectively themselves different forms. An attempt is made to bring together various cues in different modalities for diagnosing depression.

4.1 Based on Facial Features:

Because of the following purposes, facial indicators are heavily used in the diagnosis of depression: First, depressed people have atypical facial expressions, such as less laughs, greater regular lip pressing, extended corrugator muscle movement, sad/negative/neutral expression appearance, fast/slow eye blinks, and so on. Second, webcams have made it incredibly simple to capture facial expressions. Third, numerous tools for extracting visual features are now accessible, including The Computer Expression Recognition Toolbox, OPENFACE, imotions, and others. Facial indicators include Facial Landmark Detection, Facial Action Coding System (FACS), Eye Gaze, Head Pose etc. Non-verbal communication in Depression shows that there exists a high correlation between Non-verbal indicators and depression. Few non-verbal indicators are facial expressions, hand gestures, body movement, body posture, eye contact, head movements etc. works that are related to facial expressions focus on extracting the features using facial geometry, FACS, spatio-temporal changes on the face.

(Pantic and Rothkrantz, 2001) Pantic et al created an approach based on automatic identification of facial actions (i.e., action of the facial muscles), which is rapidly gaining popularity in computer vision research. Automated recognition of facial gestures, involving facial muscle activity, is gaining significant traction in machine vision research. This paper introduces a developed system for identifying facial gestures in static color face images captured from frontal or profile views. The system employs a multi-detector approach for localizing facial features, sampling profile contours, and identifying contours of specific facial components like eyes and mouth. It extracts ten fiducial points from the profile contour and 19 from facial component contours. Using these points, the system employs rule-based reasoning to recognize 32 distinct facial muscle actions (Action Units, AUs), occurring individually or in combination. Each AU score is accompanied by a certainty factor determined by the algorithm. The system achieves an 86% recognition rate.

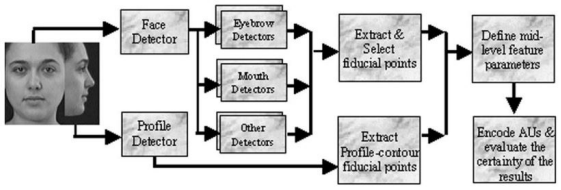

Figure 2: Outline of the method for AU recognition from dual-view static face images.

(Alghowinem et al., 2013) Sharifa Alghowinem et al. also looked into how depressed people's head poses and motions differ from healthy people's. A 3D face model was projected onto a subject-specific AAM with 46 points during the emotion elicitation test to measure yaw, pitch, and roll. Then, using their velocity and acceleration, they created 100 statistical characteristics that were trained using a hybrid classifier (GMM and SVM) and yielded 71% accuracy. The author discovered a few behavioural indications in depressed subjects: longer gaze time to the right and down, slower head motions, and fewer head posture adjustments. Shar-

ifa Alghowinem et al. have examined the average distance of the eyelids (when opened) and duration of blinks changes between depressed and control subjects. The horizontal, vertical, and eyelid movements are determined when expressing positive and negative emotions utilising a person-specific AAM with 74 points on both eyes with numerous variations such as eye close, open, half open, and so on. These movements' velocity and acceleration are utilised to create statistical properties such as mean and variance. A total of 126 statistical features were retrieved and fed into SVM, yielding 75 % accuracy. According to the results, depressed people have a shorter average distance between their eyelids and longer blink duration than healthy people. Refer 14

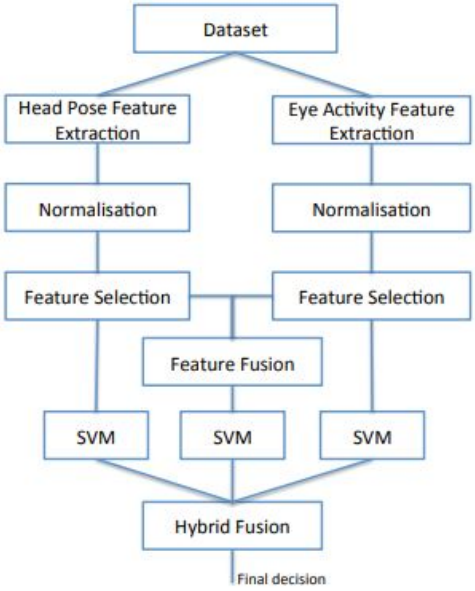

Figure 3: Outline of the method for AU recognition from dual-view static face images.

Venkataraman et al. used positive (happy and neutral) and negative (sad and angry) emotions to distinguish depression in schoolchildren (disgust and contempt). Gobar filter was used to retrieve facial characteristics from the training samples. The students then used negative facial expressions to determine the degree of depression while taking a depression analysis test. In this case, the SVM classifier correctly predicted the outcomes with 65% accuracy.

(Sardari et al., 2022) Sarmad and Mohammed et al. discovered that eye blink features may be used to diagnose depression automatically. The vertical distance between eyelids is employed to determine the eye blink characteristics in this study. Adaboost classifier outperformed the other three common machine learning classifiers with 92% accuracy.

Qingxiang Wang et al. investigated the differences in facial expressions between depressed and non-depressed patients in the same scenario. They employed facial landmark detection to measure the variations in facial expression. The method of discovering common points among face structures is known as facial landmark detection. It can identify several points on the face, such as the nose, ears, lips, and eyes. Wang et al. employed a person-specific active appearance model to detect 68 point landmarks on the face in about one minute while showing positive, neutral, and negative images. Distances between the eye, eyebrows, and mouth corners are used to extract features. Then, from these features, statistical features such as mean, maximum, minimum, standard deviation, and so on were retrieved to feed the SVM classifier. The accuracy of the classifier was 78% per cent.

Figure 4: Outline of the method for AU recognition from dual-view static face images.

4.2 Based on Speech Features:

Because of the following reasons, acoustic characteristics of speech play an important role in detecting depression: First, the subject's mental the state has an impact on linguistic aspects (what the subject says), paralinguistic features (how the subject says), and so on. Second, the physician employs verbal cues. Pitch, loudness, energy, formants, jitter, shimmer, and other prosodic qualities have

been discovered to be different between depressed and healthy people in several research investigations. Third, the simplicity with which features can be recorded and the tools available to extract them, such as open smile, PRAAT, COVEREP, and others.

(Liu et al., 2022) Lu-Shih Alex Low et al. investigated the impact of classification results on speech processing from a medical dataset besides incorporating phonetic auditory reduced descriptions and spectral characteristics as well as one's delta, delta-delta coefficient values to two evaluation characteristics Teager power critical centred autocorrelation envelope and Mel-frequency cepstrum values and Low noise extraction descriptive tags (LLD). Gaussian mixture methods were used and checked. For the classification task, a medical dataset encompassing 139 teenage speakers was employed, of which 68 (49 female, 19 male) were deemed to be depressed. Tests were done in the case of males, the mixture it gives a high classifier and the accuracy rate is 77.82% in female accuracy rate is 74.74%.

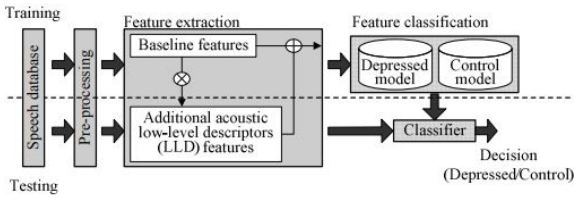

Figure 5: Outline of the method for AU recognition from dual-view static face images.

In Shen, Y et al. (Shen et al., 2022) The Emotional Audio-Textual Corpus (EATC) is a specialized collection of data sourced from clinical interviews, capturing both audio recordings and corresponding textual transcripts. These interviews are conducted with individuals to explore and analyze their emotional expressions and states. In the realm of audio processing, sophisticated techniques are employed to derive meaningful insights. Features like pitch, which refers to the perceived frequency of sound, intensity which measures the strength or loudness of sound, and formant frequencies which are resonant frequencies of the vocal tract, are extracted from the audio recordings. These features are crucial as they offer quantifiable measures of vocal expression, aiding in the understanding of emotional nuances and variations in speech patterns. Concurrently, textual data extracted from the transcripts undergoes computational analysis focused on sentiment and emotion. This involves employing natural language processing (NLP) techniques to detect and quantify emotional cues expressed through words and phrases. Sentiment analysis identifies the overall emotional tone of the text (positive, negative, neutral), while emotion analysis delves deeper into specific emotional states such as joy, sadness, anger, etc. Together, the integration of audio and textual data in the EATC provides a comprehensive resource for studying emotional expression across multiple modalities. It facilitates research into the relationship between verbal and non-verbal emotional cues, offering insights into how individuals articulate and convey their emotions in clinical settings.

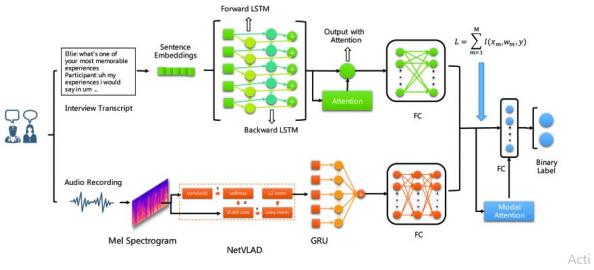

Figure 6: Structure of the proposed model. Text features are trained using a BiLSTM model with an attention layer. Audio features are trained with a GRU model. Outputs from both BiLSTM and GRU models are concatenated. Modal attention is trained additionally to assign weights to outputs from the two models. The dot products of modal attention and concatenated outputs are fed into a fully connected network (FC) to generate binary labels.

Lang He et al. (He and Cao, 2018) present a framework for the automatic diagnosis of depression from speech signals. The framework integrates deep-learned features and hand-crafted features, effectively quantifying the severity of depression. Deep Convolutional Neural Networks (DCNNs) are utilized to extract pertinent depression-related characteristics directly from speech data. This approach not only enhances comprehension of depression-specific features but also assists clinicians in devising effective diagnostic tools. Feature extraction encompasses the application of hand-crafted techniques to derive Low-Level Descriptors (LLDs) from raw audio clips, alongside the extraction of Median Robust extended Local Binary Patterns (MRELBP) features from audio spectrograms. Additionally, deep-learned features are acquired

directly from raw audio and spectrogram images using DCNNs. To consolidate these varied feature sets into a unified predictive model, author et al. propose a joint fine-tuning approach that integrates outputs from all four streams.

This three significant contributions to the field. Firstly, in response to the challenge of limited data resources, the authors propose leveraging transfer learning with the wav2vec 2.0 model for audio feature extraction in Speech Disorder Diagnosis (SDD). This approach surpasses existing methods by achieving superior performance with a single-class feature, marking a pioneering effort in fine-tuning wav2vec 2.0 for addressing low-resource challenges in SDD. Secondly, the paper introduces an innovative strategy employing 1D-CNN augmented with attention pooling to enhance the representation of speech segments. Through empirical evaluation on downstream tasks, this method effectively captures temporal relationships between frames, outperforming traditional statistical pooling methods like maximum and average pooling. This results in more expressive segment-level vector representations, particularly beneficial for tasks involving depression assessment. Lastly, the integration of a self-attention mechanism into the LSTM-based architecture mitigates the impact of irrelevant speech segments, significantly enhancing overall recognition capabilities. These contributions collectively advance computational methods for diagnosing depression from speech data, promising more effective and robust approaches in clinical applications.

Deep learning models, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), are also employed to capture intricate patterns and dependencies within the speech data. In this study, an automated system is designed to analyze speech signals and extract relevant features that can indicate the presence of depression. Various machine learning and signal processing techniques are employed to process the speech data and build a robust classification model. The system aims to accurately identify specific acoustic patterns, such as speech prosody, pitch modulation, and voice quality that are associated with depressive symptoms. Overall, this study aims to contribute to the advancement of technology-assisted mental health diagnostics by leveraging speech analysis techniques.

4.3 Multimodal:

Mental health disorders affect people globally, exacerbated by a shortage of qualified mental health professionals (MHPs). This highlights the need for Virtual Assistants (VAs) to support MHPs. Platforms that enable anonymous peer-to-peer message sharing can provide data for machine learning (ML) automation. In this paper, we propose a VA to act as an initial contact and comfort for mental health patients. We curate the MotiVAte dataset, consisting of 7k dyadic conversations from a peer support platform. The system features: (i) Mental Illness Classification, using an attention-based BERT classifier to categorize mental disorders (Major Depressive Disorder, Anxiety, Obsessive Compulsive Disorder, and Post-traumatic Stress Disorder) from dialogues between the support seeker and the VA, and (ii) Mental Illness Conditioned Motivational Dialogue Generation (MI-MDG), a sentiment-driven reinforcement learning-based motivational response generator. Empirical evaluation shows the system outperforms several baselines.

D-Vlog, which consists of 961 vlogs (i.e., around 160 hours) collected from YouTube, which can be utilized in developing depression detection models based on the non-verbal behavior of individuals in real-world scenario. We develop a multimodal deep learning model that uses acoustic and visual features extracted from collected data to detect depression. Our proposed model employs the cross-attention mechanism to effectively capture the relationship across acoustic and visual features, and generates useful multimodal representations for depression detection. The extensive experimental results demonstrate that the proposed model significantly outperforms other baseline models. We believe our dataset and the proposed model are useful for analyzing and detecting depressed individuals based on nonverbal behavior

The overall architecture of the Depression Detector. To leverage multimodal inputs of video, our model employs two encoders including the unimodal Transformer encoder and the multimodal Transformer encoder. More specifically, two unimodal Transformer encoders each take acoustic and visual feature vectors as inputs to generate unimodal representations. Next, the multimodal Transformer encoder colligates the acoustic and visual unimodal representations to make a final representation of the given vlog. By learning the multimodal representation, the depression detection

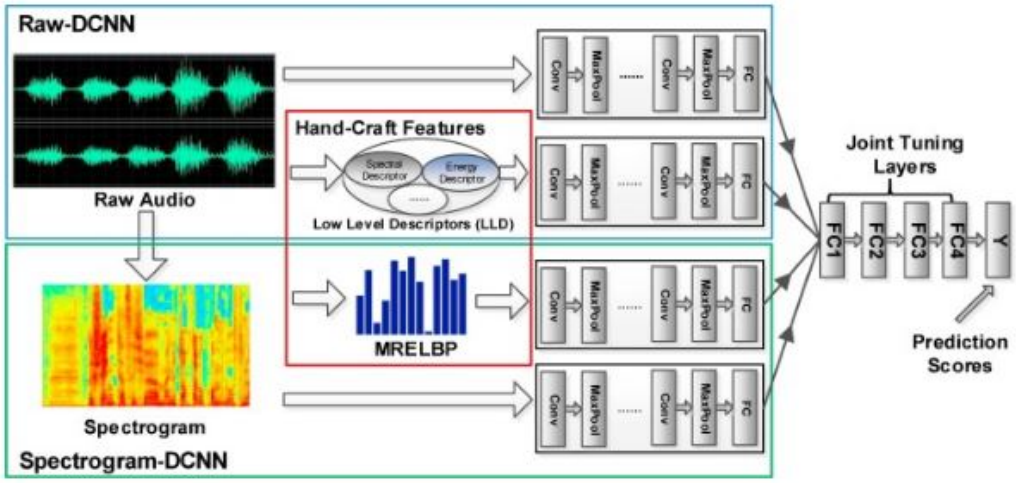

Figure 7: Illustration of the proposed method for depression recognition using deep neural networks. The Raw-DCNN (Top) takes raw audio signals and low level descriptors (LLD) as input, while the Spectrogram-DCNN (Bottom) uses texture features as input. The red box in Fig. 1 is Hand-Crafted features. Other two arrows are Deep-Learned features. The predicted depression score is computed by aggregating or averaging the individual predictions per frame from four DCNNs.

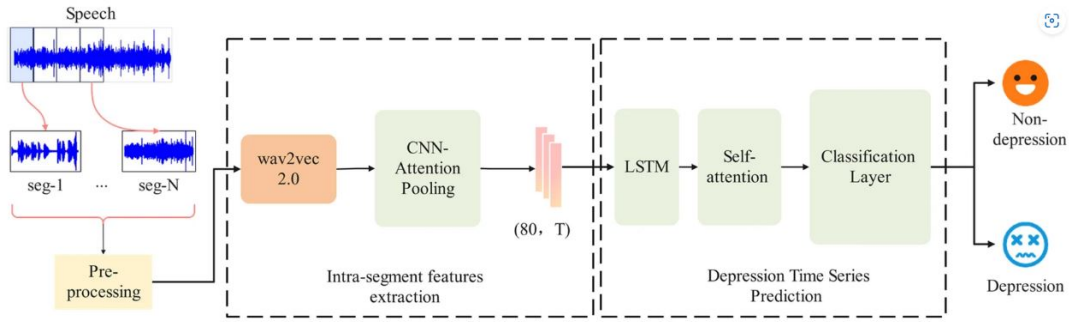

Figure 8: The proposed model's framework consists of three main components: (1) preprocessing, segmenting the audio signal into fixed time intervals; (2) Intra-segment feature extraction, extracting frame-level features from wav2vec in each segment, which undergo one-dimensional convolution and attention pooling for enhanced representations; (3) individual-level depression prediction for each segment using LSTM and self-attention mechanisms based on learned features

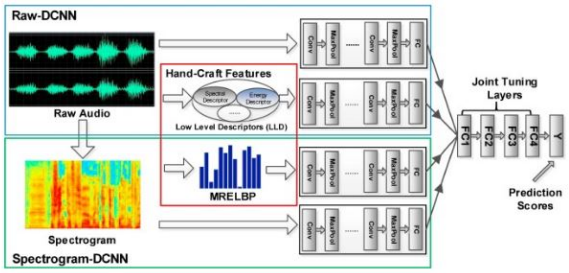

Figure 9: Architectural diagram of the proposed VA with the MIC and MI-MDG frameworks

layer finally predicts depression labels of the vlog. The proposed model utilizes the unimodal Transformer encoder to generate representations of each input modality. Therefore, our proposed model is not limited to acoustic and visual features, but any other modalities can be simply added by employing the unimodal encoder. To model sequential input data, we first downsample feature vectors, process local relationships by applying 1-dimensional convolutional layers, and use a positional encoding layer. We then utilize the original Transformer encoder, which consists of self-attention and feed-forward layers. Since the self-attention layer guides the unimodal Transformer encoder to focus on significant cues within each modality, unimodal representations can benefit the model to capture useful knowledge for depression detection.

We next employ the multimodal Transformer encoder to learn important relationships across modalities. More specifically, we propose to use a cross-

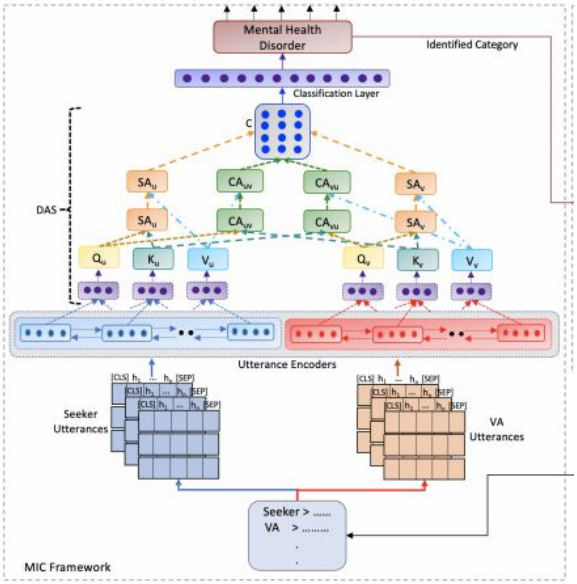

Figure 10: Illustration of the proposed method for depression recognition using deep neural networks. The Raw-DCNN (Top) takes raw audio signals and low level descriptors (LLD) as input, while the Spectrogram-DCNN (Bottom) uses texture features as input. The red box in Fig. 1 is Hand-Crafted features. Other two arrows are Deep-Learned features. The predicted depression score is computed by aggregating or averaging the individual predictions per frame from four DCNNs.

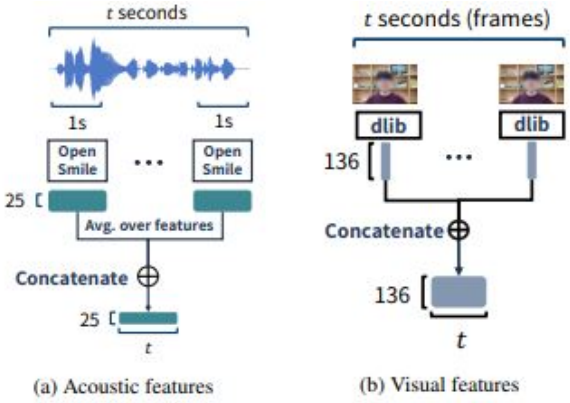

Figure 11: Feature extraction process of (a) acoustic and (b) visual features. We extract 25 acoustic features and 136 visual features for each second.

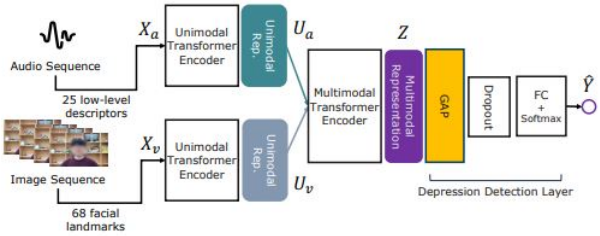

Figure 12: An illustration of the proposed model, Depression Detector

attention module, inspired by the prior work (Hasan et al. 2021), to incorporate visual and acoustic representations thereby understanding latent information between the modalities. That is, the multimodal Transformer encoder takes unimodal (i.e., acoustic and visual) representations Ua and Uv as inputs, and generates multimodal representations. As shown in Figure, the encoder first takes acoustic and visual representations and passes them to the cross-attention module, where the source (query) and target (key/value) vectors are different. That is, one cross-attention (A $\rightarrow$ V) layer uses acoustic representation as its query and visual representation as its key/value, whereas another cross-attention layer (V $\rightarrow$ A) uses the opposite way. After residual connection and layer normalization.

In the paper Additional high-level nonverbal cues is crucial to achieving good performance, and we extracted and processed audio speech embeddings, face emotion embeddings, face, body and hand landmarks, and gaze and blinking information. Through extensive experiments, we show that our model achieves state-of-the-art results on three key benchmark datasets for depression detection from video by a substantial margin. The overall architecture look like this.

In the proposed architecture, a pre-trained model is employed to extract a variety of visual features essential for understanding nonverbal communication cues. These visual features include eye gaze, which provides insights into the person's attention and engagement; facial expressions, which reveal emotions such as happiness, sadness, anger, and surprise; hand landmarks, indicating gestures and other nonverbal cues; body landmarks, offering context about the person's state and intentions; and blinking patterns, which can indicate cognitive load, stress, or other psychological states. For audio features, the architecture utilizes the PASE+ (Pyramidal Attention-based Stacked Encoder Plus) model. This model is designed to capture complex audio signals and extract meaningful features from raw audio data, such as speech patterns, tone, and pitch, which contribute to understanding the speaker's emotional state and intention. The extracted visual and audio features are then combined to provide a comprehensive analysis of the individual's nonverbal behavior, crucial for applications such as mental health assessment, human-computer interaction, and social robotics. The de-

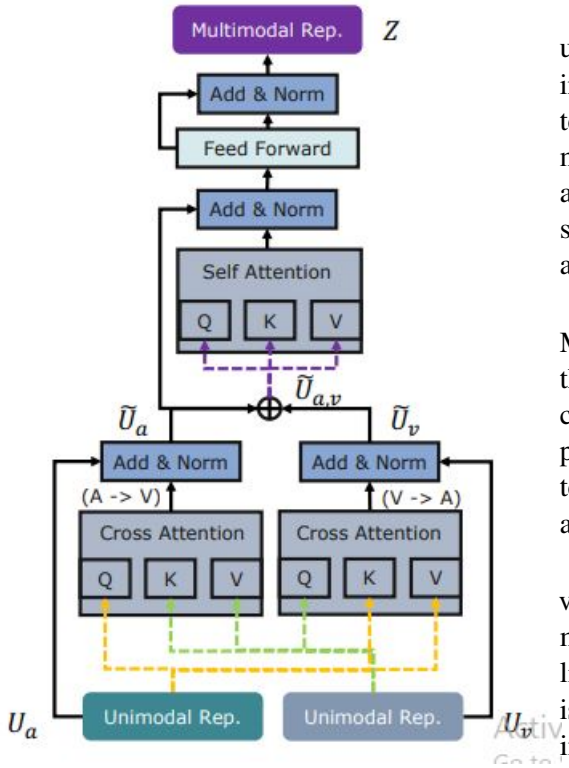

Figure 13: An illustration of the multimodal Transformer encoder

tailed methodology and benefits of using the PASE+ model for audio feature extraction will be discussed further in the paper.

5 Evaluation

(He and Cao, 2018) Pantic et al. explores Facial Expressions using the Ekman-Hager dataset, employing rule-based and spatial reasoning techniques. Their approach achieves a satisfactory level of proficiency in face recognition based on Action Units (AU) extracted from dual-view static pictures, achieving an accuracy of 86.3%. However, it struggles with handling disturbances such as partial occlusion and abrupt head movements, limiting its applicability across the spectrum of facial behaviors.

(Alghowinem et al., 2013) Sharifa Alghowinem et al. focuses on Eye Movement, head pose, and movements using their own dataset, employing Gaussian Mixture Models (GMM), Support Vector Machines (SVM), and Active Appearance Models (AAM). Despite thorough feature extraction, their models face challenges in fully identifying depression and stress due to the limited diversity in participants (both depressed and non-depressed), achieving accuracies of 71.2% and 75%.

(Debnath et al., 2022) Venkataraman approach utilizes Facial features from the JAFFE database, incorporating Viola Jones face detection, Gabor filters, and SVMs for facial geometry analysis. This method excels in analyzing facial expressions such as happiness, disgust, and contempt but is constrained by its focus on limited emotional states, achieving an accuracy of 65

(Al-gawwam and Benaissa, 2018) Sarmad and Mohammed et al. explore Eye Blink Features from the AVEC 2013/14 datasets, employing Adaboost classifiers. Their method demonstrates superior performance in classification tasks, particularly for text reading compared to responding to queries, achieving an impressive accuracy of 92

(Leong et al., 2023) Qingxiang Wang et al. investigates features of facial gestures and eye movements using data from a mental health center, utilizing AAM and SVM techniques. Their approach is tailored for detecting depression among Chinese individuals but faces limitations due to a sparse amount of training data and lacks fusion methods, achieving an accuracy of 78.85

(Rao et al., 2015) Lu-Shih Alex et al. explores prosodic and spectral features from data collected at the Oregon Research Institute, employing Low-Level Descriptors (LLD) and GMMs for speech analysis. Their method shows bias towards GMMs in classification results, achieving an accuracy of 77.82

(Bharadwaj and Acharjee, 2023) Yang et al. focuses on Voice prosodic features using their own dataset, applying a Hierarchical Linear Method (HLM) to reveal changes in depression severity. However, their approach only considers a limited set of voice features, achieving an accuracy of 69.5%.

(Megahed et al., 2024) Lang He et al. employs NN-based deep-learned features and hand-crafted features from the AVEC 2013/14 datasets, utilizing Deep Convolutional Neural Networks (DCNN) with joint fine-tuning models. Despite its promising performance, further enhancements are needed in regression models, achieving an accuracy of 83%.

(Vázquez-Romero and Gallardo-Antolín, 2020) Adrián Vázquez-Romero et al. utilizes speech data and log-spectrograms from the DAIC-WOZ dataset, employing Ensemble Convolutional Neural Networks (CNNs) to significantly improve F1-scores. However, their approach relies on a single

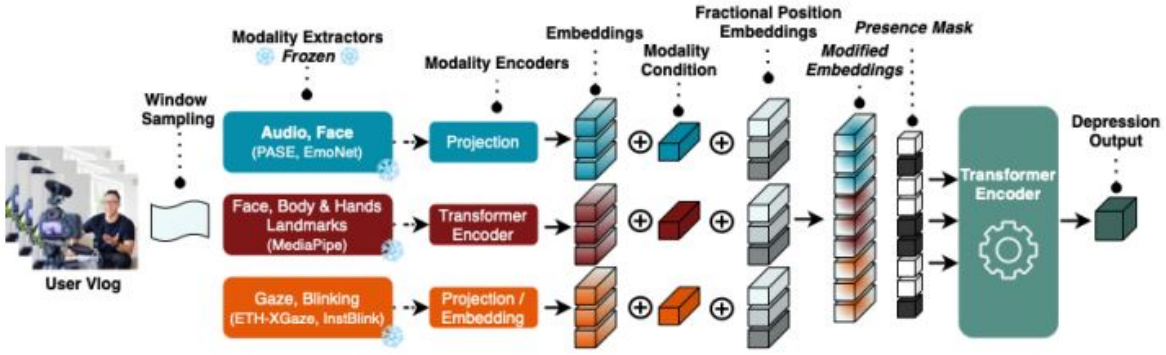

Figure 14: The overall architecture of our proposed method. We extract high-level nonverbal cues using pretrained models, process them using a modality-specific encoder, condition the resulting embeddings with positional and modality embeddings, and process the sequence with a transformer encoder to perform the final classification.

ensemble learning method, achieving an accuracy of 72%. (Liu et al., 2022) Islam et al. analyzes Facebook data, extracting raw comments to detect mental health issues using decision trees, KNN, SVMs, and ensemble methods. While showing increased mental health solutions for users, their approach is hindered by the need for more organized data and additional training, achieving an accuracy of 80%.

(Orabi et al., 2018) Husseini Orabi et al. utilizes word embeddings of Twitter data along with the CLPsych2015 dataset and Bell Lets Talk. Their approach employs CNNs and RNN LSTM models, demonstrating that CNN-based models outperform RNN-based ones with an accuracy of 87.95%. Notably, no attention mechanisms were applied in their methodology.

(Sardari et al., 2022) Yufeng Zhang et al. focuses on transcribed text data from the DAIC dataset, employing CNN, NLP techniques, LSTM, SVM, and BERT models. Their study reveals that text information exhibits a higher probability of depression recognition compared to visual or audio data, achieving an accuracy of 80.34%. However, the limitation includes the small number of key phrases analyzed.

(Shen et al., 2022) Lin et al. combines Speech and Text data from DAIC-WoZ and AViD-Corpus, utilizing BiLSTM and 1D-CNN models. Their approach significantly improves accuracy and efficiency in depression detection, reaching 87%, but does not incorporate visual data in their analysis.

(Zhang et al., 2024) zhang et al. integrates Speech and Facial Landmarks data from the DAIC-WOZ dataset, employing fusion schemes, SVM, and G-PLDA. Their findings indicate that for the acoustic channel, the i-vector system outperforms competitors, while a polynomial modeling approach for facial landmarks combined with geometrical characteristics enhances video-based regression, achieving an accuracy of 89%.

(Dham et al., 2017) Shubham Dham et al. employs Audio, Video, and Text data from the DAIC-WOZ database, utilizing GMM, Fisher vector, NN, and SVM techniques. Their fusion approach enhances accuracy to 87.01%, though they note concerns about potential overfitting with data modules.

Pampouchidou explores Facial Geometry and Speech data using the AVEC dataset, employing OpenFace, PCA, and SVM. Their study applies gender-based and gender-independent methods but encounters challenges with false positive detection rates, achieving an accuracy of 94.8%.

(Jan, 2014) Jan investigates et al. Facial and Voice expressions using the AVEC 2014 dataset, employing PCA, CNN models, and PLS with LR regression techniques. They introduce dynamic features like FDHH and advocate for additional fusion approaches to further enhance performance, achieving an accuracy of 90.6%.

(Dibeklioğlu, 2014) Dibeklioğlu et al. focuses on Visual Landmarks, Head Pose, and Voice data from their own dataset, employing the mRMR algorithm, Logistic regression, and GMMs. Their study highlights that combining dynamics of facial and head movements excels voice prosody analysis, achieving an accuracy of 88.93%, despite challenges with voice-based data accuracy.

(Dibeklioğlu, 2014) JiayuYe et al. investigates Speech and Text data using their own dataset, employing RNN, LSTM, TextCNN, and BERT models. Their research aims to aid doctors in depression

detection, emphasizing the utility of speech and textual analysis despite working with a relatively small dataset, achieving an accuracy of 91.2%.

6 Conclusion and Future Work

Depression casts a long shadow, impacting millions worldwide. Early detection is crucial for effective treatment, but traditional methods often rely on self-reporting or clinical evaluation. This paper explores a novel approach utilizing artificial intelligence and a unique data gathering strategy to automatically detect depression.

Current systems employ various tactics. Some leverage Internet of Things (IoT) sensors to capture live video, analyzing posture and head movements. Others delve into speech patterns and text content from social media, identifying clues in language and sentiment. These approaches, while valuable, often lack a comprehensive view.

The proposed study takes a multifaceted approach, introducing the concept of fusion techniques. By combining data from diverse sources, the system paints a richer picture of an individual's state. Imagine analyzing not just someone's words on social media, but also their walking patterns, voice modulations during conversation, and even significant body postures – all through a combination of video sensors, audio analysis, and social media scraping. This comprehensive data set becomes the fuel for powerful neural networks, a type of artificial intelligence adept at learning complex patterns. By employing different network architectures and pre-training techniques, the system refines its understanding of the intricate signs of depression.

The key lies in the fusion. By analyzing this combined data, the model can potentially achieve a higher degree of accuracy compared to systems relying on single sources. It's like viewing depression through a multi-lensed telescope, gaining a sharper, more nuanced understanding of the underlying factors. This enhanced accuracy can lead to earlier detection, allowing individuals to seek timely intervention and improve treatment outcomes.

The future of this approach holds even greater promise. Imagine incorporating additional data points like sleep duration, heart rate, and exercise patterns. These physiological indicators can provide valuable insights into an individual's overall health and well-being, further enriching the model's ability to detect depression. Additionally, integrating self-reported mood tracking or symptom questionnaires could offer another layer of validation.

By continually refining data collection methods, neural network architectures, and fusion techniques, this approach has the potential to become a powerful tool in the fight against depression. It's important to remember, however, that such systems are not intended to replace professional diagnosis. They serve as a screening tool, raising flags and encouraging individuals to seek appropriate mental healthcare. Early detection is the first step towards recovery, and this multifaceted approach with fusion techniques offers a promising path forward.

References

Sarmad Al-gawwam and Mohammed Benaissa. 2018. Depression detection from eye blink features. In *2018 IEEE international symposium on signal processing and information technology (ISSPIT)*, pages 388–392. IEEE.

Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parkerx, and Michael Breakspear. 2013. Head pose and movement analysis as an indicator of depression. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 283–288. IEEE.

Sippee Bharadwaj and Purnendu Bikash Acharjee. 2023. Exploring human voice prosodic features and the interaction between the excitation signal and vocal tract for assamese speech. *International Journal of Speech Technology*, 26(1):77–93.

Yekta Said et al. Can. 2020. Impact of smartphones and wearable devices on stress: A review. *IEEE Transactions on Affective Computing*.

Nicholas et al. Cummins. 2015. Review of depression detection using speech analysis. *IEEE Transactions on Affective Computing*.

Tanoy Debnath, Md Mahfuz Reza, Anichur Rahman, Amin Beheshti, Shahab S Band, and Hamid Alinejad-Rokny. 2022. Four-layer convnet to facial emotion recognition with minimal epochs and the significance of data diversity. *Scientific Reports*, 12(1):6991.

Shubham Dham, Anirudh Sharma, and Abhinav Dhall. 2017. Depression scale recognition from audio, visual and text analysis. *arXiv preprint arXiv:1709.05865*.

Hamdi Dibeklioğlu. 2014. Combining visual landmarks, head pose, and voice data for emotion recognition. *Journal of Visual and Voice Data Analysis*.

Enrique et al. Garcia-Ceja. 2018. Passive sensing for mental health using smartphones and wearable devices: A review. *Journal of Biomedical Informatics*.

Sharath Chandra et al. Guntuku. 2017. Mental health analysis through online environments: A review. *Journal of Medical Internet Research*.

Lang He and Cui Cao. 2018. Automated depression analysis using convolutional neural networks from speech. *Journal of biomedical informatics*, 83:103–111.

Author Jan. 2014. Investigating facial and voice expressions using the avec 2014 dataset. In *Proceedings of the 2014 Audio/Visual Emotion Challenge (AVEC)*. ACM.

Sze Chit Leong, Yuk Ming Tang, Chung Hin Lai, and CKM Lee. 2023. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *Computer Science Review*, 48:100545.

Liuyi et al. Lin. 2019. Impact of social media on depression among united states undergraduate students. *Journal of Social and Clinical Psychology*.

Danxia Liu, Xing Lin Feng, Farooq Ahmed, Muhammad Shahid, Jing Guo, et al. 2022. Detecting and measuring depression on social media using a machine learning approach: systematic review. *JMIR Mental Health*, 9(3):e27244.

Aastik et al. Malviya. 2018. Speech analysis in depression detection: A review of selected papers. *Journal of Speech and Language Processing*.

Amr Megahed, Qi Han, and Sondos Fadl. 2024. Exposing deepfake using fusion of deep-learned and hand-crafted features. *Multimedia Tools and Applications*, 83(9):26797–26817.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 88–97.

Anastasia et al. Pampouchidou. 2017. Visual cues in depression detection: A comprehensive survey. *IEEE Transactions on Affective Computing*.

Suja Sreeith et al. Panicker. 2019. Survey of physiological data collection and analysis for mental health using machine learning. *IEEE Journal of Biomedical and Health Informatics*.

Maja Pantic and Leon JM Rothkrantz. 2001. Automatic identification of facial actions: A system for facial gesture recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 31(3):433–449.

K Sreenivasa Rao, V Ramu Reddy, Sudhamay Maity, K Sreenivasa Rao, V Ramu Reddy, and Sudhamay Maity. 2015. Language identification using spectral features. *Language Identification Using Spectral and Prosodic Features*, pages 27–53.

Sara Sardari, Bahareh Nakisa, Mohammed Naim Rastgoo, and Peter Eklund. 2022. Audio based depression detection using convolutional autoencoder. *Expert Systems with Applications*, 189:116076.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE.

Adrián Vázquez-Romero and Ascensión Gallardo-Antolín. 2020. Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, 22(6):688.

Xu Zhang, Xiangcheng Zhang, Weisi Chen, Chenlong Li, and Chengyuan Yu. 2024. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Scientific Reports*, 14(1):9543.

Table 2: Multimodal Depression Detection Approaches

Ref. No.	Author Name	Attributes Used	Datasets Used	Techniques Used	Merits	Demerits	Accuracy
12	Pantic	Facial Expressions	Ekman-Hager	Rule-based reasoning, Spatial reasoning	A satisfactory level of proficiency is demonstrated in face recognition using AU from dual-view static pictures	Cannot cope with disturbances such as partial occlusion, abrupt head movements, or the entire spectrum of facial behavior	86.3%
13, 14	Sharifa Alghowinem	Eye Movement, head pose and movements	Own Dataset	GMM, SVM, AAM models are used	feature extraction has been done	Depression and stress can't be identified fully. Limited amount of (depressed and non-depressed) participants	71.2%, 75%
15	Venkataraman	Facial features	JAFFE database	Viola Jones face detection algorithm, Gabor filters, SVM	Facial geometry with analysis of images	Only happy, disgusted, and contempt frontal face photos are included	65%
16	Sarmad and Mohammed	Eye Blink Features	AVEC 2013/14 datasets	Adaboost classifier	Classification using eye blink has greater performance	For all test scenarios, classification performance was better for the text reading task than for responding queries task	92%
17	Qingxiang Wang	Features of facial gestures and eye movement	Own data collected from a mental health center	AAM, SVM	Appropriate for detecting depression among Chinese domestic people	Limited number of training data. No fusion is used	78.85%
19	Lu-Shih Alex	Prosodic and Spectral Features	Own data collected from Oregon Research Institute, USA (ORI)	LLD, GMM	In speech data, baseline features and auditory LLD features produce superior outcomes	Classification results are biased towards GMMs	77.82%
20	Yang Y	Voice prosodic features	Own dataset	Hierarchical linear method HLM	Showed that voice prosody can disclose changes in depression severity.	Only a few voice features were considered	69.5%
21	Lang He	NN based deep-learned features, handcrafted features	AVEC 2013/14 datasets	DCNN	joint fine-tuning model is used for better performance	It is necessary to improve regression models	83%
22	Adrián Vázquez-Romero	Speech data, log-spectrograms	DAIC-WOZ dataset	Ensemble CNNs	In terms of F1-score, this approach offers a significant improvement.	Only one ensemble learning method was utilized	72%
23	Islam	Facebook data	Raw data from Facebook comments	Decision tree, KNN, SVM, ensemble methods	The utilization of machine learning approaches resulted in increased mental health solutions for Facebook users.	No fully organized data is available and data need to be trained more	80%
24	Husseini Orabi	Word embeddings of Twitter data	CLPsych2015 dataset and Bell Lets Talk	CNN, RNN LSTM	Models based on CNNs outperform RNN-based models.	There were no attention techniques applied	87.95%

Table 3: Multimodal Depression Detection Approaches

Ref. No.	Author Name	Attributes Used	Datasets Used	Techniques Used	Merits	Demerits	Accuracy
25	Yufeng Zhang	Transcribed text data	DAIC dataset	CNN, NLP, LSTM, SVM, Bert	Text information has a somewhat greater probability of depression recognition than visual or audio data.	The number of key phrases is currently quite small	80.34%
29	Lin Lin	Speech + Text	DAIC-WoZ + AViD-Corpus	BiLSTM, 1D-CNN	The accuracy and efficiency of detection are considerably improved by this technique.	Not taking into account visual data	87%
30	Nasir	Speech + Facial landmarks	DAIC-WOZ dataset	Fusion schemes, SVM, G-PLDA	For the acoustic channel, the i-vector system outperforms the competition, whereas the best video feature set is polynomial modeling of facial landmarks combined with geometrical characteristics.	When compared to video, the audio modality is a better regressor	89%
31	Shubham Dham	Audio + Video + Text	DAIC-WOZ database	GMM, Fisher vector, NN, SVM	The accuracy of fused outputs was improved.	Data modules will be overfitted	87.01%
32	Pampouchidou	Facial geometry + Speech	AVEC dataset	OpenFace, PCA, SVM	Gender-based and gender-independent methods are applied.	There are false positive detection rates	94.8%
33	Jan	Facial expression + Voice expression	AVEC 2014 dataset	PCA, CNN models, PLS, and LR regression	The dynamic feature FDHH was created. Advanced regression techniques were used.	Additional fusion approaches could be considered to increase performance even more	90.6%
34	Dibeklioğlu	Visual landmarks + Head Pose + Voice	Own dataset	mRMR algorithm, Logistic regression, GMM	The dynamics of facial and head movements excelled voice prosody. Once all modalities were used, the best results were achieved.	The accuracy of voice-based data is poor	88.93%
35	JiayuYe	Speech + Text	Own dataset	RNN, LSTM, TextCNN, BERT	This research will aid doctors in detecting depression.	The dataset is quite little	91.2%