Deep Learning for Emotion Recognition and Depression Detection from Speech: A Survey

www.surveyx.cn

Abstract

This survey paper explores the intersection of deep learning, speech analysis, and emotion recognition in the context of mental health assessment, with a particular focus on depression detection. The integration of deep learning techniques with speech analysis offers transformative approaches to diagnosing mental health conditions, enhancing accuracy and scalability compared to traditional methods. Advanced architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are employed to extract nuanced features from speech signals, providing objective markers for depression diagnosis. The incorporation of multimodal data, including audio, text, and visual inputs, enhances diagnostic robustness and reliability, allowing for a comprehensive understanding of emotional and psychological states. Attention mechanisms and feature fusion further refine model performance by dynamically weighting relevant features across modalities, improving both accuracy and interpretability. The paper underscores the importance of diverse datasets and standardized evaluation metrics in advancing research, ensuring model generalizability across diverse populations. Addressing challenges such as data scarcity and privacy concerns through innovative techniques will further enhance system effectiveness. The integration of deep learning and speech analysis holds significant potential for revolutionizing mental health assessment and intervention, promising early intervention and improved management of mental health conditions, ultimately enhancing individual well-being and reducing societal impacts.

1 Introduction

1.1 Significance of Mental Health Assessment

Mental health assessment is crucial due to its profound implications for individuals and society. Depression, a widespread mental disorder, significantly impairs individuals' daily functioning and contributes to substantial social and economic burdens, affecting over 300 million people globally and ranking as a leading cause of disability and suicide [1]. In the United States, major depressive disorder (MDD) impacts nearly 7% of adults, necessitating effective detection methods [2]. The World Health Organization identifies depression as a common yet often underdiagnosed condition, complicating timely identification and treatment [3].

The societal impact of mental health disorders extends to considerable economic costs, with depression and anxiety leading to trillions in lost productivity annually [4]. Despite its prevalence, depression remains frequently underdiagnosed, underscoring the need for effective mental health assessment strategies [5]. The complexity of diagnosing depression, characterized by multifaceted symptoms, further necessitates advanced detection technologies [6]. Improved classification of depression severity is essential for resource prioritization [7].

Traditional treatments for depression can be time-consuming and ineffective, with patients often concealing their mental states during clinical interviews. This highlights the urgent need for automatic

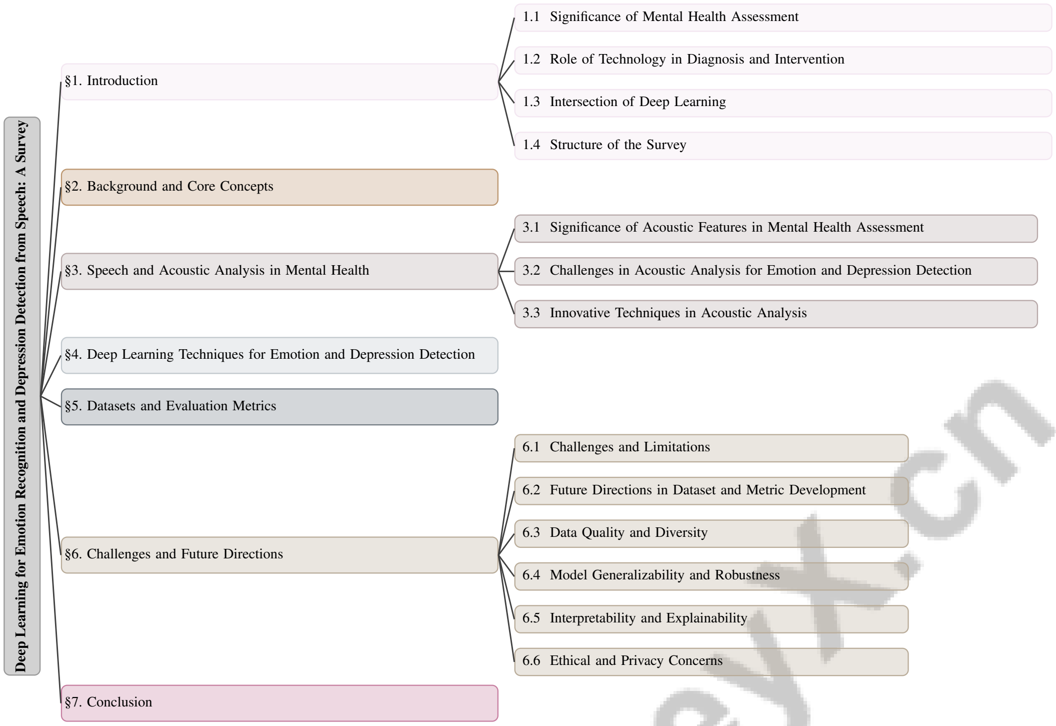

Deep Learning for Emotion Recognition and Depression Detection from Speech: A Survey
§1. Introduction
1.1 Significance of Mental Health Assessment
1.2 Role of Technology in Diagnosis and Intervention
1.3 Intersection of Deep Learning
1.4 Structure of the Survey
§2. Background and Core Concepts
§3. Speech and Acoustic Analysis in Mental Health
3.1 Significance of Acoustic Features in Mental Health Assessment
3.2 Challenges in Acoustic Analysis for Emotion and Depression Detection
3.3 Innovative Techniques in Acoustic Analysis
§4. Deep Learning Techniques for Emotion and Depression Detection
§5. Datasets and Evaluation Metrics
§6. Challenges and Future Directions
6.1 Challenges and Limitations
6.2 Future Directions in Dataset and Metric Development
6.3 Data Quality and Diversity
6.4 Model Generalizability and Robustness
6.5 Interpretability and Explainability
6.6 Ethical and Privacy Concerns
§7. Conclusion

Figure 1: chapter structure

depression detection systems that facilitate self-assessment and enhance diagnostic accuracy [8]. Innovative approaches, including the analysis of social media data, are being explored to enhance mental health assessment, addressing the shortage of trained clinicians and resources [9]. Early identification of depressive symptoms is vital for assessment, intervention, and relapse prevention, emphasizing the importance of advancing these methodologies to improve individual well-being and reduce societal impacts.

1.2 Role of Technology in Diagnosis and Intervention

The integration of technology in mental health diagnostics and interventions marks a significant shift from traditional methods, which often rely on subjective self-reports and standardized questionnaires. These conventional approaches are limited by biases and accessibility issues, necessitating more objective and scalable solutions [1]. The rise of smart devices enables passive monitoring of depression through behavioral signals, such as speech, offering innovative avenues for objective mental health assessments [10].

Deep learning models, particularly those utilizing speech-based features, enhance diagnostic precision by providing objective markers for depression prediction, thereby improving both accuracy and scalability in real-time assessments. The application of machine learning to social media data presents novel pathways for addressing mental health concerns, as these platforms allow users to express feelings anonymously, providing valuable insights for practitioners [9]. This underscores the critical role of technology in refining diagnostic and intervention strategies.

Innovative tools such as AI chatbots and self-supervised learning models expand the diagnostic toolkit by enabling the extraction of clinical features from electronic health records, fostering the development of adaptive systems responsive to evolving data sources. The use of multimodal cues—including speech, language, and visual data—facilitates the automatic prediction of depression severity scores, allowing for earlier and more accurate detection of depressive symptoms [11].

Non-invasive technologies, such as remote monitoring systems, enhance the diagnosis of depression through more accessible means, eliminating the need for contact-based sensors [12]. The ongoing development and integration of these technologies hold significant promise for improving mental health outcomes and addressing the growing demand for effective mental health care solutions. As the field advances, there is an increasing emphasis on developing accurate and privacy-preserving methods to assess depression using speech signals, ensuring that these systems do not compromise individual privacy [13].

1.3 Intersection of Deep Learning, Speech Analysis, and Emotion Recognition

The intersection of deep learning, speech analysis, and emotion recognition is foundational for advancing methodologies in depression detection. Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are pivotal in extracting depression-relevant features from speech signals. These models effectively capture intricate patterns that serve as objective markers for mental health assessments, leveraging alterations in speech patterns as reliable indicators of depression [10]. The integration of sentiment analysis with depression detection, exemplified by the DeSK method, enhances diagnostic accuracy [14].

Speech analysis offers a non-invasive approach to mental health evaluation by examining acoustic features, such as prosody and timing cues, indicative of altered emotional states in individuals with depression [15]. The incorporation of non-verbal signals, highlighted by the D-Vlog benchmark, differentiates between depressed and non-depressed individuals in real-world situations, enhancing assessment robustness [16]. Furthermore, Bayesian Networks (BNs) model complex relationships between depression, its symptoms, and various data streams, providing a transparent framework for understanding these interactions [17].

Emotion recognition, facilitated by affective computing, leverages data from diverse modalities—including video, audio, and text—to assess sentiment and emotional states. This multimodal integration enhances diagnostic accuracy compared to traditional methods [18]. Advanced methodologies utilizing deep neural networks and unsupervised label correction classify suicidal versus depressed individuals based on online content, showcasing the potential of these technologies in mental health assessment [19]. The exploration of non-verbal behavioral cues enriches the diagnostic process by adding layers of information relevant to depression detection [20].

Incorporating social media text analysis and physiological data from wearable devices offers a holistic approach to monitoring emotional states and early detection of depressive symptoms [21]. These advancements underscore the transformative potential of combining deep learning, speech analysis, and emotion recognition in mental health assessment, paving the way for early intervention and improved outcomes in depression management. Automated systems for detecting depression in speech ensure timely mental health care, highlighting the significance of these technologies in contemporary mental health strategies [22].

The proposed SAD-TIME method, utilizing a spatiotemporal-fused network, exemplifies the innovative integration of multiple neural network components to automate feature extraction and improve classification accuracy [6]. Moreover, the ambient intelligence-based approach for detecting depression relapse through automatic speech recognition (ASR) technology analyzes speech patterns indicative of depression [23]. The multi-stage deep learning model proposed by Seneviratne et al. integrates segment-level and session-level predictions to classify depression severity, highlighting the intersection of deep learning and speech analysis [7]. Additionally, the integration of audio and text features in a multimodal approach enhances depression detection without relying on specific questions [8].

By exploiting interactions across various modalities, these methodologies enhance both performance and interpretability, offering a comprehensive framework for depression diagnosis [3]. The innovative combination of topic modeling with multimodal analysis enables a context-aware approach that retains critical temporal features, further illustrating the potential of these technologies in mental health assessment [2]. Furthermore, the notion that multiple acoustic features can provide a more comprehensive understanding of depression than single-feature models highlights the potential of feature fusion in this domain [1].

1.4 Structure of the Survey

This survey is meticulously structured to provide a comprehensive overview of the intersection of deep learning, speech analysis, and emotion recognition in the context of mental health assessment, particularly focusing on depression detection. The paper begins with an *Introduction* that underscores the significance of mental health assessment and elucidates the transformative role of technology in enhancing diagnostic and intervention strategies. This section also highlights the convergence of deep learning, speech analysis, and emotion recognition, setting the stage for subsequent discussions.

Following the introduction, the *Background and Core Concepts* section delves into the foundational elements of the study, providing definitions and explanations of key concepts such as depression, emotion recognition, deep learning, and acoustic analysis, establishing a solid theoretical framework for readers.

The survey transitions into *Speech and Acoustic Analysis in Mental Health*, exploring the utilization of speech and acoustic features for assessing mental health. This section emphasizes the significance of acoustic features in detecting emotional and psychological states, identifies challenges in acoustic analysis, and discusses innovative techniques that enhance emotion and depression detection.

In *Deep Learning Techniques for Emotion and Depression Detection*, the focus shifts to the application of various deep learning models in the field. This section examines different architectures, such as CNNs and RNNs, and explores the use of attention mechanisms and feature fusion to improve model performance and detection accuracy.

The *Datasets and Evaluation Metrics* section reviews the datasets and evaluation metrics commonly employed in this research area, discussing their characteristics, limitations, and importance in advancing the field, providing insights into the resources available for model training and evaluation.

The survey concludes with the section on *Challenges and Future Directions*, highlighting existing obstacles and limitations in employing deep learning techniques for emotion recognition and depression detection from speech. This includes the need for robust models that accurately interpret emotional nuances in vocal expressions, challenges posed by diverse speech patterns across demographics, and the integration of multimodal data for improved diagnostic accuracy. It underscores the importance of ongoing research to enhance the effectiveness of deep learning frameworks like EmoAudioNet, which have shown promise in the early detection of MDD but still face hurdles in real-world applications [24, 21]. Potential future directions emphasize the need for improved datasets, evaluation metrics, data quality, model generalizability, interpretability, and ethical considerations.

The survey synthesizes essential findings, emphasizing the transformative potential of integrating deep learning techniques with speech analysis for mental health assessment and intervention. This integration enhances the accuracy of detecting mental health conditions such as depression and anxiety, paving the way for more personalized and effective mental health care solutions. By leveraging advanced models that analyze speech features and text representations, the research underscores the importance of multimodal approaches in improving diagnostic capabilities and fostering timely interventions in mental health care [25, 26, 27, 28].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Definitions and Explanations of Depression

Major Depressive Disorder (MDD) is a complex mental health condition marked by persistent sadness, diminished interest in activities, and significant functional impairment, complicating diagnosis and treatment due to its diverse emotional, cognitive, and physical symptoms [1]. Traditional diagnostic tools, reliant on subjective assessments, often discourage help-seeking due to stigma [8]. The societal impact of depression is severe, with decreased productivity and increased healthcare costs, underscoring the need for objective assessment methods [4]. Depression often coexists with other conditions like anxiety, necessitating comprehensive diagnostic tools to differentiate overlapping symptoms.

Technological advancements have introduced multimodal approaches for depression assessment, utilizing audio, video, and text data to predict emotional expressions, though real-world classification remains challenging due to limitations in feature extraction from pre-trained models [8]. Integrating audio and text features within a multimodal framework enhances detection accuracy, providing a more comprehensive assessment strategy [1]. Addressing these challenges requires innovative approaches that incorporate multimodal data and advanced analytical techniques to improve assessment accuracy, particularly in detecting depression from speech signals and social media language [4].

AI-generated, for reference only.

2.2 Emotion Recognition and Its Relevance

Emotion recognition is crucial in mental health assessment, enabling the analysis of emotional expressions across modalities like speech, text, and social media, which are vital for identifying disorders such as depression characterized by emotional dysregulation. Integrating sentiment features is essential for accurate depression detection, emphasizing emotion recognition's significance in mental health contexts [14]. In speech analysis, emotion recognition utilizes linguistic features to capture emotional subtleties, enriching depression detection models [29]. Comparing traditional acoustic features with deep representation features enhances predictions of depression severity, highlighting effective feature extraction's necessity [21]. High-level dialogue features further illustrate emotion recognition's role in assessing mental health, improving automatic mood episode detection in clinical interviews [19].

Emotion recognition also extends to text analysis, where detecting emotional nuances in social media communications provides insights into users' mental health states. The dynamic nature of emotional expressions in informal texts poses challenges for depression detection, yet datasets of tweets filtered by hashtags offer a rich source of informal language reflective of users' mental states [9]. This approach, inspired by affective computing, enhances understanding and prediction of psychological states [29]. Emotion recognition's relevance is further underscored by its application in analyzing interactions within free-living audio environments, crucial for accurately assessing depression severity. Continuous monitoring of speech timing features and dyadic interaction frequency links to varying depression levels, with mild depression showing increased interaction frequency and moderate to severe depression showing a decrease, highlighting audio-based markers' potential to enhance diagnostic accuracy and inform treatment strategies [30, 31]. By leveraging emotional information from diverse data sources, emotion recognition systems contribute to a holistic understanding of psychological states, facilitating better patient outcomes and enhancing well-being.

2.3 Deep Learning in Mental Health

Deep learning has transformed mental health care, particularly in early detection and management of depression, through advanced algorithms like Long Short-Term Memory (LSTM) networks. These models analyze text data, including social media posts, to identify self-reported depressive symptoms with remarkable accuracy, reaching up to 99

Recurrent Neural Networks (RNNs), including LSTM and Gated Recurrent Units (GRU), capture temporal dependencies in data, crucial for understanding depressive symptoms' progression. These models effectively analyze social media content, providing insights into individuals' mental health states [8]. The integration of Natural Language Processing (NLP) with deep learning refines depressive symptoms detection, allowing nuanced mental states understanding [4]. Innovative architectures like ABAFnet enhance depression detection accuracy by integrating multiple acoustic features from speech data [1], offering diagnostic accuracy and clinical insights. Convolutional neural networks analyze spectrograms from voice samples, facilitating objective depression assessment [8].

Deep learning extends beyond conventional data sources, incorporating methodologies like articulatory coordination features and multimodal classifiers integrating audio, text, and visual data for improved depression detection. Benchmarks inspired by combining multiple data modalities, including EEG, yield richer insights into mental health conditions than single modalities, exemplifying deep learning's potential in capturing nuanced features indicative of mental health conditions [4]. Utilizing remote photoplethysmography (rPPG) signals from facial videos to compute physiological features for depression assessment highlights deep learning's versatility in mental health applications. Federated Learning, a decentralized training approach, allows models to train on local datasets without sharing data, ensuring privacy and data security, crucial in mental health contexts. Data augmentation techniques like FrAUG enhance depression detection from speech by modifying frame-rate parameters during feature extraction, improving model robustness [8].

3 Speech and Acoustic Analysis in Mental Health

The integration of speech and acoustic analysis into mental health research offers crucial insights into emotional and psychological states. This section delves into the importance of acoustic features for mental health assessments, highlighting their role in enhancing diagnostic precision and providing non-invasive evaluation methods. As illustrated in Figure 2, the hierarchical structure of speech and acoustic analysis in mental health emphasizes the significance of these acoustic features, while also addressing the challenges faced in the field and the innovative techniques employed. The figure categorizes the main ideas into diagnostic precision enhancement, multimodal data analysis, data scarcity and privacy concerns, technical challenges, advanced models, and data augmentation. By examining diverse acoustic parameters, we gain a comprehensive understanding of their contributions to mental health diagnostics.

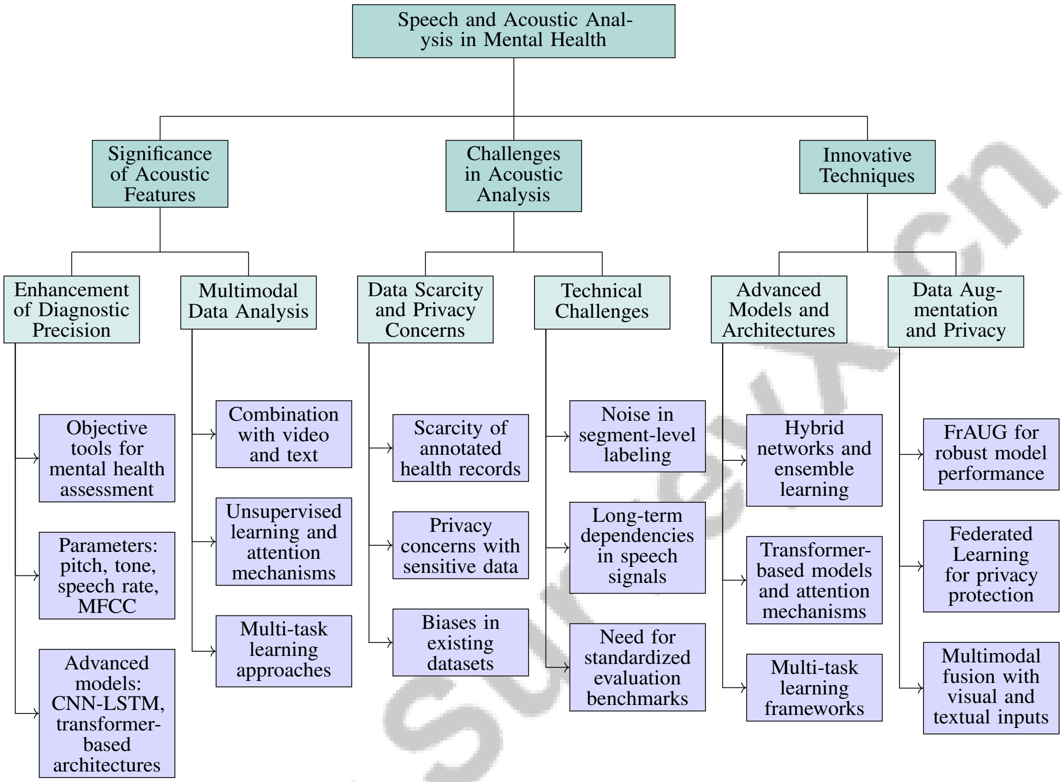

Speech and Acoustic Analysis in Mental Health
Significance of Acoustic Features
Challenges in Acoustic Analysis
Innovative Techniques
Enhancement of Diagnostic Precision
Multimodal Data Analysis
Data Scarcity and Privacy Concerns
Technical Challenges
Advanced Models and Architectures
Data Augmentation and Privacy
Objective tools for mental health assessment
Parameters: pitch, tone, speech rate, MFCC
Advanced models: CNN-LSTM, transformer-based architectures
Combination with video and text
Unsupervised learning and attention mechanisms
Multi-task learning approaches
Scarcity of annotated health records
Privacy concerns with sensitive data
Biases in existing datasets
Noise in segment-level labeling
Long-term dependencies in speech signals
Need for standardized evaluation benchmarks
Hybrid networks and ensemble learning
Transformer-based models and attention mechanisms
Multi-task learning frameworks
FrAUG for robust model performance
Federated Learning for privacy protection
Multimodal fusion with visual and textual inputs

Figure 2: This figure illustrates the hierarchical structure of speech and acoustic analysis in mental health, emphasizing the significance of acoustic features, challenges faced in the field, and innovative techniques employed. It categorizes the main ideas into diagnostic precision enhancement, multimodal data analysis, data scarcity and privacy concerns, technical challenges, advanced models, and data augmentation.

3.1 Significance of Acoustic Features in Mental Health Assessment

Acoustic features serve as objective tools in mental health assessment, enhancing traditional diagnostic approaches by evaluating emotional and psychological states through parameters like pitch, tone, speech rate, and Mel-frequency cepstral coefficients (MFCC) [32]. Advanced deep learning models, such as CNN-LSTM and transformer-based architectures, process audio data to detect depression, utilizing complex vocal patterns to increase diagnostic accuracy [24]. Ensemble learning combined with feature extraction techniques, including word embeddings and transformer models, further refines detection by capturing a wide range of acoustic features [33]. The fusion of audio and visual cues is critical for estimating depression severity accurately [34].

As illustrated in Figure 3, the hierarchical structure of acoustic features in mental health assessment highlights key techniques, multimodal analysis approaches, and considerations for privacy and accessibility. Acoustic features are integral to multimodal data analysis, where they are combined with video and text for comprehensive mental health assessments [35]. Unsupervised learning and attention mechanisms enhance depression severity assessments based on speech characteristics [15]. Multi-task learning (MTL) approaches, incorporating dialog structure and emotional information, further improve depression detection performance [36].

Timing-related speech features, such as speech rate and pause time, are crucial in assessing mental health, reflecting psychomotor slowing, a key feature of Major Depressive Disorder (MDD) [37]. Advanced neural network architectures capture speech motor coordination, underscoring acoustic features' transformative potential in mental health evaluations [38]. These methodologies highlight the importance of acoustic analysis in providing nuanced insights into emotional and psychological states, ultimately enhancing diagnostic accuracy and patient outcomes.

Federated Learning techniques enhance privacy protection and facilitate real-time assessments on mobile devices, making acoustic analysis more accessible and secure [13]. Data augmentation techniques like FrAUG improve model performance without distorting crucial acoustic information, creating a robust framework for enhancing depression detection [5]. Methods leveraging both audio and textual cues provide richer and more accurate representations of depressive states [8].

Integrating acoustic, behavioral, linguistic, and visual features leads to comprehensive models for depression detection, as demonstrated in studies combining these modalities to predict depression severity. Non-invasive approaches utilizing visual information to capture physiological changes associated with depression present promising avenues for remote mental health assessment [12]. Innovative methodologies, such as one-shot learning comparing audio and textual encodings, exemplify advancements in depression relapse detection [23]. These developments underscore acoustic features' critical role in advancing mental health diagnostics and improving patient outcomes.

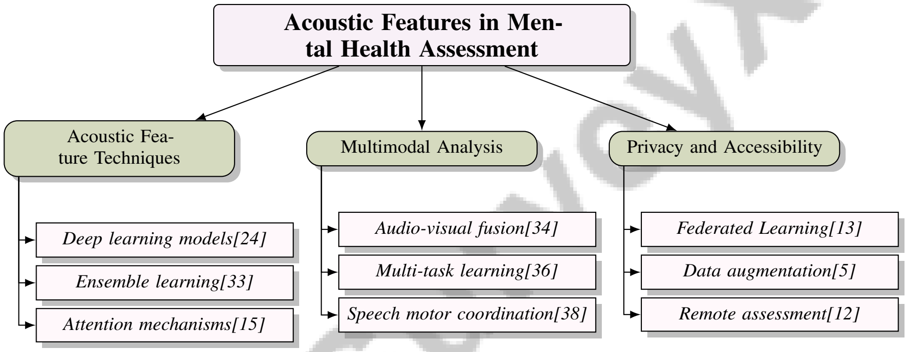


Figure 3: This figure illustrates the hierarchical structure of acoustic features in mental health assessment, highlighting key techniques, multimodal analysis approaches, and considerations for privacy and accessibility.

3.2 Challenges in Acoustic Analysis for Emotion and Depression Detection

Acoustic analysis for emotion and depression detection faces significant challenges, including the scarcity of annotated health records and privacy concerns with sensitive mental health data. These limitations restrict comprehensive training datasets crucial for robust model development [39]. The reliance on labeled data for model training presents a substantial hurdle, as acquiring such data is costly and time-consuming, complicating model generalization to unseen data [40].

Data availability and privacy concerns are compounded by biases in existing datasets, which can skew model predictions and limit applicability across diverse populations. The lack of standardized evaluation benchmarks complicates model performance assessment, impeding field advancements [41]. Existing data augmentation techniques, such as Vocal Tract Length Perturbation (VTLP), risk distorting acoustic data, potentially losing vital information related to the speaker's mental state [42].

Another challenge is the noise introduced by segment-level labeling, where segments may lack relevant information for depression detection, leading to inaccurate predictions. Capturing long-term dependencies in speech signals, especially with lengthy recordings, poses an additional obstacle, impacting the model's ability to analyze extended speech data accurately [43].

Addressing these challenges requires innovative data collection approaches, privacy-preserving techniques, and sophisticated models capable of handling acoustic data complexities. Implementing standardized benchmarks and exploring innovative data augmentation techniques that preserve acoustic features' integrity is crucial. This approach enhances depression-relevant characteristic extraction from speech signals while leveraging advanced hybrid networks and frame rate-based data augmentation methods, which improve predictive accuracy across datasets. Combining self-attention networks with deep convolutional neural networks and refining feature extraction processes can develop more robust systems for assessing mental health conditions, facilitating earlier interventions and improving diagnostic outcomes [43, 25, 5].

3.3 Innovative Techniques in Acoustic Analysis

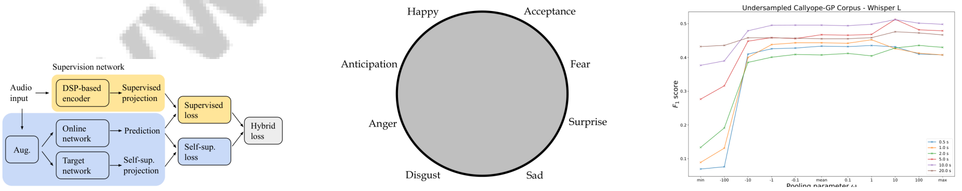
Recent advancements in acoustic analysis have led to innovative techniques for improving emotion and depression detection, including hybrid networks extracting depression-relevant features from speech signals, the Emotional Audio-Textual Depression Corpus for multimodal analysis, and advanced models predicting individual symptoms from acoustic features, enhancing diagnostic accuracy and facilitating early intervention [43, 44, 45, 8]. These methods employ deep learning architectures and novel data processing strategies to enhance mental health assessments' accuracy and efficiency. Advanced feature extraction methods capture intricate acoustic patterns indicative of emotional states, with MFCC, spectral centroid, and zero-crossing rate integration forming a robust framework for evaluating speech characteristics associated with depression.

Transformer-based models and attention mechanisms significantly contribute to the field by enhancing the model's ability to focus on relevant speech data segments. Combined with ensemble learning approaches, these models improve performance in identifying depression-related cues from vocal inputs [33]. Multi-task learning (MTL) frameworks incorporating dialog structure and emotional information show promise in refining depressive symptom detection by capturing a broader range of acoustic features [36].

Innovative data augmentation techniques, such as FrAUG, address limited training data challenges and enhance model robustness without compromising acoustic features' integrity. These techniques modify frame-rate parameters during feature extraction, creating diverse training samples that improve model generalization [5]. Federated Learning approaches ensure data privacy while allowing accurate model development for real-time assessment on mobile devices [13].

Integrating acoustic analysis with multimodal data, including visual and textual inputs, offers a comprehensive mental health assessment approach. This multimodal fusion enhances depression severity detection by extracting complementary information from various sources [35]. Timing-related speech features, such as pause duration and speech rate, provide valuable insights into psychomotor retardation, a key depression symptom [37].

These innovative acoustic analysis techniques represent significant advancements in emotion and depression detection. Utilizing vocal acoustic features as objective markers for diagnosing mental health disorders like major depressive disorder, bipolar disorder, schizophrenia, and generalized anxiety disorder, researchers have achieved high classification performance, with specificity reaching up to 93.80



(a) Audio classification with hybrid supervised/self-supervised learning[46]

(b) The image represents a Venn diagram with five overlapping circles, each representing a different emotion.[47]

(c) Undersampled Callyope-GP Corpus - Whisper L.[27]

Figure 4: Examples of Innovative Techniques in Acoustic Analysis

As shown in Figure 4, innovative techniques in speech and acoustic analysis are invaluable for understanding emotional and psychological states. This example highlights three cutting-edge methodologies showcasing the diversity and potential of acoustic analysis in mental health research. The first technique involves a hybrid supervised/self-supervised learning approach for audio classification, depicted through a detailed flowchart. This method leverages both supervised and online networks to process audio inputs, transforming them into representations suitable for further analysis.

The second technique is illustrated by a Venn diagram with five overlapping circles, each representing different emotions such as happiness, acceptance, fear, surprise, and disgust, helping to understand the complex interplay of emotions, with acceptance being the most dominant. The final example focuses on the undersampled Callyope-GP Corpus using the Whisper L model, demonstrating the relationship between the F1 score and the pooling parameter across various undersampling rates. These innovative techniques underscore the importance of acoustic analysis in advancing mental health research, offering new avenues for emotion detection and cognitive state assessment [46, 47, 27].

4 Deep Learning Techniques for Emotion and Depression Detection

Category	Feature	Method
Deep Learning Architectures	Emotion-Based Representation Attention and Temporal Models Hybrid CNN-RNN Architectures	DeepBoSE[48] MDLF[28], DN[20], MMD[29], ABAFnet[1] AVCN[49], MS-DCNN-LSTM[7]
Attention Mechanisms and Feature Fusion	Attention and Fusion Integration	MT-CombAtt[50], MMDDS[51], ND[52], AST[53], SDCNL[19], WMFG[22]
Innovative Approaches and Hybrid Models	Cost-Effective Strategies Hybrid Neural Architectures Contextual Analysis Techniques	CDDS[32] LSTM-RNN[21] TMMFV[2]

Table 1: The table presents a comprehensive overview of various deep learning methodologies employed in emotion and depression detection. It categorizes these methodologies into three main areas: deep learning architectures, attention mechanisms and feature fusion, and innovative approaches and hybrid models. Each category highlights specific features and methods, illustrating the diversity and advancement in the field.

Deep learning techniques have revolutionized emotion and depression detection, offering more nuanced assessments of mental health. Table 4 offers a detailed classification of deep learning methods in emotion and depression detection, underscoring the architectures and techniques that contribute to improved diagnostic accuracy and efficiency. This section examines key deep learning architectures and their applications in improving emotion recognition and depression detection accuracy.

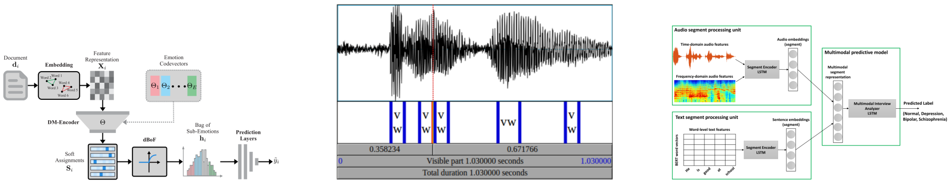
4.1 Deep Learning Architectures

Method Name	**Architectural Components**	**Data Integration**	**Feature Extraction**
DN[20]	Cnn And Gru	Multimodal Data Integration	Summarization Techniques
MMD[29]	Svm, Lstm	Multi-modal Analysis	Low-level Acoustic
LSTM-RNN[21]	Lstm, Rnn	Multiple Data Types	One-Hot Encoding
WMFG[22]	Recurrent Neural Network	Audio, Visual, Textual	Gating Mechanisms
MS-DCNN-LSTM[7]	Dilated Cnns	Multimodal Data	Extract Acfs
ABAFnet[1]	Lstm-Attention Pipeline	Multiple Acoustic Features	Acoustic Feature Fusion
MDDM[8]	Gru, Bilstm	Audio, Text	Mel Spectrograms
DeepBoSE[48]	Deep Learning Model	-	Clustering
AVCN[49]	Deep Cnn	Audio, Visual	Phoneme-level Features
MDLF[28]	Lstm With Attention	Audio And Text	Advanced Embedding Techniques

Table 2: Comparison of various deep learning architectures for emotion recognition and depression detection, detailing their architectural components, data integration methods, and feature extraction techniques. The table highlights the diversity of approaches, ranging from CNNs and RNNs to attention mechanisms and multimodal data integration, showcasing their applications in mental health diagnostics.

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are pivotal in advancing emotion recognition and depression detection. CNNs effectively extract spatial features from audio signals and spectrogram images, which are crucial for accurate classification [20]. Their integration into multimodal frameworks enhances depression detection by utilizing diverse data sources [29]. RNNs, including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), capture temporal dependencies in speech, aiding emotion detection systems [21]. Frameworks like WMFG, which combine audio, visual, and lexical data, demonstrate RNNs' effectiveness in emotion recognition and depression detection [22]. Hybrid models, such as the Multi-Stage Dilated CNN-LSTM Model (MS-DCNN-LSTM), leverage CNNs and RNNs to enhance depression analysis [7]. Attention-enhanced GRUs within the DepressionNet framework exemplify innovative architecture integration for depression classification [20]. Models like ABAFnet, which use acoustic features and attention mechanisms, improve depression detection [1], while GRUs and Bidirectional LSTM (BiLSTM) with attention layers enhance audio and text feature representation [8].

Emerging architectures, including large language models (LLMs) and encoder-based Transformers, leverage linguistic features, audio analysis, and emotion-specific embeddings to improve mental health diagnostics. These models address biases and methodological challenges, offering scalable solutions for early intervention in mental health care [26, 54, 28].


Document
$\mathbf{d}_i$
Embedding
Word 1
Word 2
Word 3
Word 4
Word 5
Word 6
Feature Representation
$\mathbf{X}_i$
Emotion Codevectors
Θ_1 Θ_2 • • • Θ_E
DM-Encoder
Θ
Soft Assignments
$\mathbf{S}_i$
dBoF
Bag of Sub-Emotions
$\mathbf{h}_i$
Prediction Layers
$\tilde{y}_i$
V
W
VW
0.358234
0.671766
0
Visible part 1.030000 seconds
1.030000
Total duration 1.030000 seconds
Audio segment processing unit
Time-domain audio features
Frequency-domain audio features
Segment Encoder LSTM
Audio embeddings (segment)
Text segment processing unit
Word-level text features
BERT word vectors
He
is
good
at
school
Sentence embeddings (segment)
Multimodal predictive model
Multimodal segment representation
Multimodal Interview Analyzer LSTM
Predicted Label
(Normal, Depression, Bipolar, Schizophrenia)

(a) Document Emotion Recognition System[48]

(b) Seismic Waveform with Visible and Total Duration Indicated[49]

(c) Audio and Text Multimodal Predictive Model[28]

Figure 5: Examples of Deep Learning Architectures

As shown in Figure 5, deep learning techniques in emotion and depression detection leverage advanced architectures for complex data analysis. The Document Emotion Recognition System transforms input documents into feature representations using a DM-Encoder, generating a bag of sub-emotions for prediction. The Seismic Waveform image represents seismic wave amplitudes, visually differentiating waveform sections for analysis. The Audio and Text Multimodal Predictive Model exemplifies the integration of audio and text data, employing separate processing units for accurate label prediction [48, 49, 28]. Additionally, Table 2 provides a comprehensive comparison of deep learning models used in emotion recognition and depression detection, illustrating the methodologies and components employed in these advanced architectures.

4.2 Attention Mechanisms and Feature Fusion

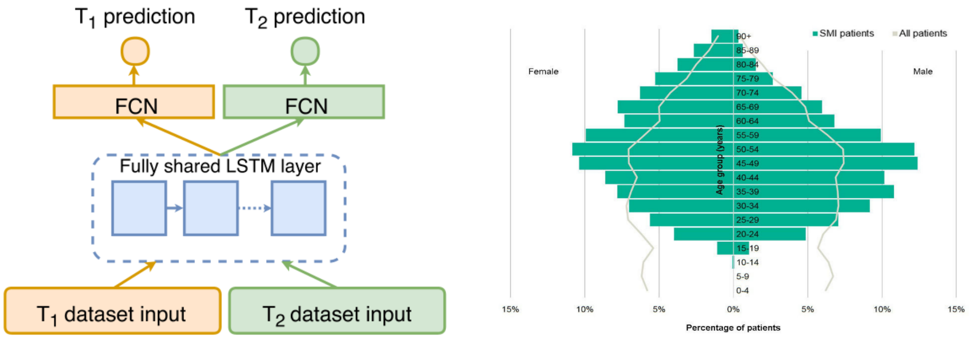
Attention mechanisms and feature fusion significantly enhance model performance in emotion recognition and depression detection by emphasizing relevant features across modalities. Attention mechanisms focus on significant input data segments, improving understanding of emotional and psychological states [53]. The hierarchical attention-based clustering network (HACN) models tweets and user clusters, improving depression detection [52]. Combining attention mechanisms with linguistic embeddings enhances classification by capturing temporal dynamics and the interplay between verbal and non-verbal cues [55, 56]. Haque et al. employ unsupervised learning to handle noisy data, enhancing depression detection robustness [19].

Feature fusion integrates audio, text, and visual data, creating comprehensive input representations. Rohanian et al.'s method illustrates learning interdependencies and temporal dynamics between modalities, refining detection [22]. Multi-task learning frameworks use feature fusion for dynamic modality weighting, enhancing model adaptability and performance.

In Figure 6, the integration of deep learning techniques, particularly attention mechanisms and feature fusion, shows promise in emotion and depression detection. The first image presents a model for time series prediction, highlighting its architecture's reliance on fully connected layers and a shared LSTM layer. The second image visualizes age distribution among Sickle Cell Malaria (SMI) patients compared to a broader patient population, using a bar chart to convey demographic insights. These examples illustrate deep learning techniques' potential in advancing mental health diagnostics through sophisticated data analysis and interpretation [50, 51].

4.3 Innovative Approaches and Hybrid Models

Innovative approaches and hybrid models have advanced emotion recognition and depression detection, enhancing accuracy and addressing challenges like data scarcity and computational efficiency. Random Forest classifiers analyze part-of-speech features' discriminatory power, capturing linguistic



(a) Deep Learning Model for Time Series Prediction[50]

(b) Age Distribution of SMI Patients and All Patients[51]

Figure 6: Examples of Attention Mechanisms and Feature Fusion

Method Name	Methodological Approaches	Feature Utilization	Application Context
CDDS[32]	Machine Learning Models	Acoustic Features	Speech Detection
LSTM-RNN[21]	Lstm-RNN Model	One-Hot Encoding	Social Media Posts
TMMFV[2]	Topic Modeling	Audio, Video	Long Interviews

Table 3: Overview of various methodological approaches and their applications in emotion recognition and depression detection, highlighting the integration of machine learning models, feature utilization, and application contexts. The table compares different methods such as CDDS, LSTM-RNN, and TMMFV, emphasizing their unique contributions to mental health diagnostics.

nuances relevant to mental health assessment. SHAP values enhance transparency, providing insights into features contributing to depression detection [57]. Tasnim et al. show that conventional acoustic features can perform comparably or better than deep representation features with lower computational costs, highlighting traditional acoustic analysis techniques' importance [32]. Amanat et al. demonstrate deep learning frameworks' transformative potential in improving detection accuracy compared to traditional frequency-based models [21].

Hybrid models combining CNNs and RNNs offer comprehensive solutions for mental health diagnostics, capturing spatial and temporal features to improve depression detection accuracy. Gong et al.'s integration of topic modeling with multimodal analysis outperforms context-unaware methods, emphasizing contextual information's effectiveness in mental health assessments [2].

These innovative approaches and hybrid models signify advancements in mental health technology, delivering precise, efficient, and scalable solutions for emotion recognition and depression detection. For example, a hybrid method combining Collgram and sentiment analysis based on the BERT architecture achieves 71

Feature	Deep Learning Architectures	Attention Mechanisms and Feature Fusion	Innovative Approaches and Hybrid Models
Data Type	Audio, Visual, Text	Audio, Text, Visual	Linguistic, Acoustic
Model Integration	Cnns, Rnns	Attention Networks	Hybrid Models
Feature Enhancement	Attention Mechanisms	Feature Fusion	Shap Values

Table 4: This table provides a comprehensive comparison of various deep learning methodologies applied in emotion and depression detection. It categorizes the methods based on data types utilized, model integration strategies, and feature enhancement techniques, highlighting the diversity and innovation in current research approaches.

5 Datasets and Evaluation Metrics

5.1 Commonly Used Datasets

In the realm of emotion recognition and depression detection, diverse datasets comprising audio, text, and multimodal data have been instrumental in advancing model development and evaluation. These datasets facilitate the application of deep learning models like XLNet, BERT, and WaveNet, enabling the extraction of significant features from speech and text for accurate mental health condition predictions [58, 28].

The Distress Analysis Interview Corpus - Wizard-of-Oz (DAIC-WOZ) is pivotal in depression detection research, offering audio, video, and transcripts from clinical interviews with 107 subjects. Despite its small sample size and uneven depression case distribution, it remains widely used, with studies ensuring balanced representation [2, 8]. The Extended DAIC-WOZ (E-DAIC-WOZ) further expands the dataset for comprehensive analysis [3].

The AVEC series, particularly AVEC2013 and AVEC2014, provides benchmarks with speech files and human-agent interaction recordings labeled with PHQ-8 scores, crucial for evaluating speech-based depression detection techniques. Additionally, smartphone-collected speech data have identified linguistic topics related to depression severity, enhancing automated mental health monitoring [31, 43, 25, 59, 60].

Social media platforms, notably Twitter, serve as significant sources for depression detection datasets. Amanat et al. utilized over 4000 tweets from Kaggle to explore linguistic features associated with depression [21], while large-scale datasets have been employed to compare advanced models like DepressionNet against baseline methods [20]. Rajput et al.'s dataset from tweets with depression-related hashtags and essays further underscores social media data's utility [9].

The CNRAC and CS-NRAC databases offer clinical speech recordings for acoustic analysis in depression detection [1]. The EATD-Corpus, with audio and text responses from 162 volunteers, is valuable for evaluating automatic emotion recognition and depression detection methods [8]. Other notable datasets include the Chinese Multimodal Compression Corpus (CMDC) and the PRED+CT and MODMA datasets, which enhance depression detection model evaluation [3, 6]. Speech data from the MD-1 and MD-2 databases, with clinician-rated scales, are crucial for defining depression severity in research [7].

These datasets collectively advance emotion recognition and depression detection research, supporting the development of sophisticated models utilizing techniques like LSTM neural networks and semi-supervised learning. By integrating audio and text features from interviews and analyzing emotional recall in texts, these datasets foster innovative methodologies that improve diagnostic accuracy and accessibility, addressing challenges faced by individuals seeking mental health support [61, 31].

5.2 Dataset Characteristics and Limitations

Datasets used in emotion recognition and depression detection exhibit distinct characteristics and limitations impacting their effectiveness. A key feature is their multimodal nature, incorporating audio, text, and sometimes visual data to comprehensively represent emotional and psychological states. For instance, the DAIC-WOZ dataset includes audio, video, and transcripts from clinical interviews, providing a rich data source for diagnosing mental disorders [2]. Similarly, the AVEC series features speech files and interaction recordings with labeled PHQ-8 scores, essential for evaluating speech-based depression detection techniques.

However, these datasets face significant limitations. A major challenge is the limited sample size and uneven distribution of depression cases, affecting generalizability [8]. The DAIC-WOZ dataset, with only 107 subjects, may not adequately represent diverse manifestations of depression [3]. Social media-derived datasets, like those from Twitter, often contain noisy data lacking clinical validation for accurate depression detection.

Another limitation is potential bias in existing datasets. Reliance on specific languages or cultural contexts can restrict the applicability of models across diverse demographic groups, limiting emotion recognition and depression detection systems' scope [20]. Moreover, the absence of standardized evaluation benchmarks complicates performance assessment, making cross-study and cross-dataset comparisons challenging [41].

The scarcity of annotated health records and privacy concerns surrounding sensitive mental health data further constrain comprehensive training dataset availability, crucial for developing robust models [39]. Innovative data augmentation techniques, like those in the FrAUG framework, attempt to mitigate limited data issues by modifying frame-rate parameters during feature extraction; however, these methods can distort acoustic data, risking the loss of vital information [42].

AI-generated, for reference only.

5.3 Evaluation Metrics for Model Performance

Benchmark	Size	Domain	Task Format	Metric
MultiEEG-GPT[4]	3,000	Mental Health	Multi-class Classification	Accuracy, F1-score
CogLoad[46]	1,000	Speech Analysis	Stress Detection	UAR
DEM[62]	365	Neuropsychology	3-class Classification	UAR
MMPsy[63]	12,024	Psychology	Mental Health Assessment	Accuracy, F1
D2D[64]	2,448	Speech Analysis	Clustering	Silhouette Score, Elbow Method
MASON-NLP[65]	3,807,115	Mental Health	Sentence Ranking	Average Precision, R-Precision
AD-CLT[66]	3,503	Psychology	Multi-label Classification	Accuracy, F1-score
MH-SSB[27]	10,120	Mental Health	Speech Analysis	F1 Score

Table 5: This table presents a comprehensive overview of various benchmarks utilized in the evaluation of model performance for emotion recognition and depression detection. It details the benchmark name, dataset size, domain, task format, and the evaluation metrics employed, highlighting the diversity and scope of datasets used in this field.

Evaluation metrics are critical for assessing model performance in emotion recognition and depression detection, quantifying their effectiveness in accurately identifying emotional states and depression levels. Table 5 provides a detailed summary of representative benchmarks used in evaluating model performance within the domains of emotion recognition and depression detection. Various studies employ diverse approaches, including text analysis, audio modeling, and hybrid methodologies, to capture human communication nuances and enhance diagnostic accuracy. Integrating clinical markers and improving explainability through these metrics fosters trust among healthcare professionals and supports effective monitoring and intervention for individuals at risk of depression [67, 68, 69, 31]. These metrics provide comprehensive assessments of model accuracy, robustness, and generalizability across different datasets.

The F1-score, widely used in classification tasks, balances precision and recall, particularly beneficial in contexts with class imbalances, as it accounts for false positives and negatives [3]. Precision measures the proportion of true positives among all positive predictions, while recall assesses true positives among all actual positives, both critical for evaluating model effectiveness in distinguishing depressed from healthy individuals [4].

Accuracy remains a standard metric, reflecting the proportion of correctly classified instances. It is often used alongside the F1-score for holistic model effectiveness evaluation [23]. Confusion matrices detail the distribution of true positives, true negatives, false positives, and false negatives, facilitating overall accuracy and F1 score calculations [3].

For regression models predicting continuous outcomes like depression severity scores, metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are vital. RMSE emphasizes larger errors, providing an average magnitude of errors, while MAE offers a straightforward average of absolute errors, both crucial for assessing model precision in capturing nuanced mental health states [3]. The Concordance Correlation Coefficient (CCC) evaluates the agreement between predicted and actual values, ensuring models accurately reflect depression severity [3].

Cross-validation techniques, such as grouped 10-fold cross-validation, are commonly employed for robust and unbiased model evaluations. This approach involves splitting data into multiple folds, ensuring that samples from the same individual do not appear in both training and test sets, thus providing reliable performance assessments [23]. Additionally, metrics like the Pearson Correlation Coefficient (CC) evaluate the correlation between predicted and actual outcomes, offering insights into model predictive capabilities [23].

The meticulous selection and implementation of evaluation metrics are crucial for advancing research in emotion recognition and depression detection. These metrics facilitate the development of models that achieve high accuracy and reliability across varied datasets and conditions while addressing the complexities of depression assessment, such as interpreting language nuances, the sequential nature of conversations, and integrating diverse emotional dimensions. By leveraging advanced techniques like machine learning and hybrid models, researchers can better capture the intricacies of emotional expression, ultimately leading to more effective monitoring and intervention strategies for individuals experiencing depression [67, 69, 61, 31].

AI-generated, for reference only.

6 Challenges and Future Directions

In the realm of emotion recognition and depression detection, numerous challenges arise, encompassing technical, data-related, and ethical dimensions associated with advanced machine learning models. This section delves into these challenges, highlighting areas necessitating innovation.

6.1 Challenges and Limitations

Emotion recognition and depression detection through speech analysis face significant hurdles, including the complexity of feature extraction from diverse audio and textual data, limited datasets due to privacy issues, and the need for robust model architectures that generalize across varied contexts. These challenges obstruct the development of reliable models, despite advancements in multimodal data systems for enhanced diagnostic accuracy [31, 26, 70, 43, 8]. Individual variability in depression expression complicates the creation of universally applicable models, leading to inconsistent diagnostic tools. Additionally, the scarcity and quality of labeled data are substantial challenges, as deep learning models require extensive datasets often limited in size and quality.

Data from platforms like social media introduces biases and challenges due to their informal, unstructured nature, complicating the extraction of meaningful insights [9]. The focus on English language tweets limits generalizability across linguistic and cultural contexts, while web-scraped data may harbor biases that affect model performance despite label correction efforts [22].

Privacy concerns are paramount, especially regarding continuous audio recording for real-time emotion recognition. Ensuring the privacy and security of sensitive mental health data is challenging due to the nature of data collection methods [13]. Models need extensive validation across diverse populations to ensure generalizability in real-world settings [23].

Integrating diverse features from multiple modalities, including audio, video, and text, poses further challenges. Current methods often struggle with high computational costs and inefficiencies in blending multimodal features necessary for comprehensive assessments of emotional and psychological states [6]. Data quality and completeness, along with potential biases in training datasets, further limit the effectiveness of emotion recognition systems [20].

Methodological limitations also impact model performance. For instance, the MMFF approach may face challenges with limited data or highly variable patient responses [3]. Increased complexity and training time associated with processing multiple features pose challenges for real-time applications [1]. Additionally, reliance on high-quality speech data, which may not always be accessible, can affect model generalizability [7]. Methods using small sample sizes may face challenges in feature selection and risk overfitting [2], while variations in audio quality and the subjective nature of depression symptoms can further influence model performance [8].

Addressing these challenges requires innovative approaches in data collection, model development, and training dataset enhancement. Employing advanced analytical techniques that effectively integrate multimodal information—such as lexical content, audio features, and gestural behaviors—can enhance the reliability and applicability of emotion recognition and depression detection systems. Utilizing models like recurrent neural networks and transformer-based architectures can improve diagnostic accuracy and provide deeper insights into cognitive and motor states, ultimately leading to better mental health diagnostics and improved patient outcomes through automated, passive monitoring of communication patterns [70, 22, 3, 31].

6.2 Future Directions in Dataset and Metric Development

Future research in emotion recognition and depression detection should prioritize expanding and diversifying datasets to enhance model generalizability and applicability across various populations. This includes broadening datasets to encompass a wider range of languages, emotions, and demographic groups, addressing current limitations in representativeness and applicability [9]. Validating expanded corpora against clinical standards will ensure relevance and accuracy in real-world applications, while exploring applicability across different social media platforms and languages will further enhance utility [23].

Incorporating additional modalities, such as visual and textual features, will provide a more comprehensive understanding of mental health indicators and improve model performance [23]. Integrating multimodal data fusion techniques is essential for refining feature extraction processes and enhancing depression detection accuracy [29]. Future research should focus on refining gating mechanisms and exploring sophisticated approaches for capturing visual data applicable in real-time clinical settings [22].

Improving evaluation metrics is critical for future research. Refining existing metrics and exploring additional features that may enhance model performance will contribute to more accurate assessments. Research should prioritize optimizing input lengths and speaking rates tailored to specific datasets, as these factors significantly influence machine learning model performance, particularly in speaker-independent depression classification and automatic speech assessment for voice disorders. Studies indicate that NLP system effectiveness varies based on natural and elapsed speech input lengths and their order within a session. Additionally, large language models have demonstrated superior performance in detecting mental disorders, especially in noisy and diverse datasets, highlighting the importance of refining input parameters to enhance data elicitation and model accuracy across clinical applications [53, 26, 71]. Furthermore, incorporating phase information to improve spectrogram intelligibility and developing novel blocks for enhanced feature representations will be explored, extending to multimodal approaches.

6.3 Data Quality and Diversity

The importance of data quality and diversity in model training and evaluation for emotion recognition and depression detection is paramount. High-quality, diverse datasets are essential for developing robust models capable of generalizing across different populations and conditions. The variability in emotional expression and depressive symptoms necessitates datasets that capture a wide range of demographic and cultural contexts, ensuring models are not biased towards specific groups or environments [9].

Data quality directly impacts model prediction accuracy and reliability. High-quality data should be devoid of noise and inconsistencies, providing clear representations of emotional and psychological states. This is particularly crucial in speech and acoustic analysis, where audio quality variations can significantly affect model performance [8]. Efforts to enhance data quality include employing advanced preprocessing techniques and validating datasets against clinical standards to ensure relevance and accuracy in real-world applications.

Diversity in datasets is equally important, enabling models to learn from a wide array of emotional expressions and depressive symptoms. Diverse datasets should encompass various languages, cultures, and demographic groups, thereby improving model generalizability and reducing biases arising from homogeneous training data [23]. The integration of multimodal data sources, including audio, text, and visual inputs, further enriches datasets, providing a comprehensive representation of mental health indicators and enhancing model performance [29].

6.4 Model Generalizability and Robustness

Enhancing model generalizability and robustness is a critical challenge in emotion recognition and depression detection. Implementing robust data processing techniques combined with adaptive machine learning algorithms significantly improves models' ability to generalize across diverse datasets and conditions [72]. This approach ensures models remain effective despite variations in input data, such as those encountered in different demographic groups or linguistic contexts.

Dynamic systems integration (DSI) offers additional flexibility and real-time adaptability, allowing models to better integrate and process data from dynamic environments. This adaptability is crucial for maintaining accuracy and reliability in real-world applications, where data conditions frequently change [73]. The test-time training approach, such as TTT-MAE, enhances performance by adapting to distribution shifts at test time, outperforming traditional methods that do not account for such variations [74].

Reducing labeling noise and improving model interpretability are essential for enhancing generalizability. Implementing frame-based attention interpretation methods can effectively minimize labeling noise, providing clearer insights into acoustic features relevant for depression detection [44]. This improves model interpretability and robustness by ensuring decisions are based on accurate information.

AI-generated, for reference only.

Future research should focus on enhancing multimodal feature extraction processes and exploring the effectiveness of visualizations and text summaries to improve user understanding and engagement [56]. Advancements in user interaction and processing long contextual information are critical for addressing model generalizability, ensuring models effectively handle complex and varied inputs [75].

By implementing these strategies, researchers can develop computational models that enhance robustness and generalizability, delivering reliable mental health assessments across diverse conditions and populations, particularly through language use analysis in social media. This approach not only deepens understanding of the relationship between offensive language and mental health but also facilitates creating a representative corpus for detecting mental illness, evidenced by high correlations between social media data and established depression indicators [9, 76].

6.5 Interpretability and Explainability

Interpretability and explainability are crucial for deploying deep learning models in mental health assessment, ensuring these models are effective, transparent, and trustworthy. The complexity of deep learning architectures often leads to a "black box" perception, where decision-making processes remain obscure to clinicians or end-users. This lack of transparency can impede integration into clinical environments, where understanding diagnostic reasoning is vital for informed decision-making. Recent advancements in explainable artificial intelligence (XAI) underscore the necessity of incorporating clinical markers and interpretability into these models. For instance, methodologies utilizing Large Language Models (LLMs) provide diagnostic classifications alongside understandable explanations based on established criteria, fostering trust among healthcare professionals. Enhancing diagnostic tool interpretability facilitates better engagement and tailored recommendations, ultimately improving patient care and outcomes [75, 68, 77, 78].

Attention mechanisms within deep learning models have emerged as critical tools for enhancing interpretability. By enabling models to focus on relevant input features, attention mechanisms offer insights into which data aspects are indicative of mental health conditions like depression. This transparency is essential for fostering trust in automated systems and ensuring they complement, rather than replace, human expertise [8].

Integrating multimodal data—encompassing audio, text, and visual inputs—significantly enhances model interpretability by providing a comprehensive understanding of psychological states and mental health disorders. Recent studies demonstrate improved classification accuracy in mental health assessments when combining diverse data sources such as EEG signals, speech samples, and facial expressions [41, 4, 28]. By offering a holistic view of an individual's mental state, these models can provide comprehensive insights into factors influencing mental health, supporting clinicians in making informed decisions. Techniques like Local Interpretable Model-agnostic Explanations (LIME) allow for deconstructing model predictions, providing clear explanations that can be communicated to patients and healthcare providers.

Future research should focus on developing applications enabling users to self-assess their depressive states using interpretable models, empowering individuals to actively manage their mental health [8]. Additionally, exploring privacy-preserving techniques in federated learning settings can enhance user data security while maintaining model transparency and accuracy. By prioritizing interpretability and explainability, researchers can ensure deep learning models for mental health are effective and trustworthy, ultimately improving mental health care quality and patient outcomes.

6.6 Ethical and Privacy Concerns

Integrating deep learning in mental health assessment presents significant ethical and privacy concerns that require careful consideration for responsible technology use. A primary concern is the handling of personal data, crucial for developing accurate models but introducing substantial privacy risks. Anonymization and encryption of collected data, as emphasized by Levinson et al., are essential to safeguard participant privacy and prevent unauthorized access [79]. Similarly, Zogan et al. highlight the importance of ensuring user privacy and data protection when analyzing social media data related to mental health [20].

The ethical implications of utilizing large language models (LLMs) in mental health applications are underscored by Hu et al., who emphasize the necessity of ethical considerations in deploying these technologies [4]. This includes ensuring AI systems support rather than replace human healthcare providers, maintaining human oversight in mental health diagnostics to mitigate potential misuse or harm. The potential for false classifications in automated systems necessitates careful domain adaptation and supervision by qualified therapists to minimize risks and ensure accurate assessments.

Ilias et al. explore label smoothing for model calibration in stress and depression detection, highlighting the importance of model reliability and accuracy in sensitive applications [55]. Well-calibrated models reduce the likelihood of erroneous predictions that could adversely affect individuals' mental health.

Furthermore, the energy consumption associated with advanced models, such as transformer models, raises ethical considerations regarding sustainability and resource usage. Researchers must navigate the delicate balance between advancing technological capabilities and upholding environmental stewardship, as highlighted by Laguna et al. This is particularly crucial to ensure that implementing innovative models for automatic depression detection using artificial intelligence and natural language processing does not incur excessive environmental costs, given the significant energy demands associated with these technologies [33, 69, 80].

7 Conclusion

This survey explores the intersection of deep learning, speech analysis, and emotion recognition, highlighting their collective impact on mental health assessment, particularly in diagnosing depression. The utilization of deep learning models in speech analysis represents a paradigm shift in mental health diagnostics, offering enhanced precision and scalability over traditional methods. By employing sophisticated architectures like CNNs and RNNs, these models can extract complex features from speech, providing reliable markers for depression diagnosis.

Incorporating multimodal data—encompassing audio, text, and visual elements—strengthens diagnostic systems' comprehensiveness, enabling a nuanced understanding of emotional and psychological states. The strategic use of attention mechanisms and feature fusion enhances model performance by effectively prioritizing relevant features across different modalities, thus improving diagnostic accuracy and offering valuable insights for clinical use.

The survey also emphasizes the critical role of datasets and evaluation metrics in advancing this research area. Developing diverse, representative datasets and standardized evaluation metrics is essential to ensure models' generalizability and effectiveness across various populations and settings. Addressing challenges such as data scarcity and privacy through novel data collection and processing techniques will further improve these systems' impact and utility.

AI-generated, for reference only.

References

[1] Xiao Xu, Yang Wang, Xinru Wei, Fei Wang, and Xizhe Zhang. Attention-based acoustic feature fusion network for depression detection, 2023.

[2] Yuan Gong and Christian Poellabauer. Topic modeling based multi-modal depression detection, 2018.

[3] Chengbo Yuan, Qianhui Xu, and Yong Luo. Depression diagnosis and analysis via multimodal multi-order factor fusion, 2022.

[4] Yongquan Hu, Shuning Zhang, Ting Dang, Hong Jia, Flora D. Salim, Wen Hu, and Aaron J. Quigley. Exploring large-scale language models to evaluate eeg-based multimodal data for mental health, 2024.

[5] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. Fraug: A frame rate based data augmentation method for depression detection from speech signals. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6267–6271. IEEE, 2022.

[6] Han-Guang Wang, Hui-Rang Hou, Li-Cheng Jin, Chen-Yang Xu, Zhong-Yi Zhang, and Qing-Hao Meng. Sad-time: a spatiotemporal-fused network for depression detection with automated multi-scale depth-wise and time-interval-related common feature extractor, 2024.

[7] Nadee Seneviratne and Carol Espy-Wilson. Speech based depression severity level classification using a multi-stage dilated cnn-lstm model, 2021.

[8] Ying Shen, Huiyu Yang, and Lin Lin. 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, 2022.

[9] Adil Rajput and Samara Ahmed. Making a case for social media corpus for detecting depression, 2019.

[10] Sandeep Kumar Pandey, Hanumant Singh Shekhawat, SRM Prasanna, Shalendar Bhasin, and Ravi Jasuja. A deep tensor-based approach for automatic depression recognition from speech utterances. *Plos one*, 17(8):e0272659, 2022.

[11] Evgeny Stepanov, Stephane Lathuiliere, Shammur Absar Chowdhury, Arindam Ghosh, Radu-Laurentiu Vieriu, Nicu Sebe, and Giuseppe Riccardi. Depression severity estimation from multiple modalities, 2017.

[12] Constantino Álvarez Casado, Manuel Lage Cañellas, and Miguel Bordallo López. Depression recognition using remote photoplethysmography from facial videos, 2022.

[13] Suhas BN and Saeed Abdullah. Privacy sensitive speech analysis using federated learning to assess depression, 2022.

[14] Yan Shi, Yao Tian, Chengwei Tong, Chunyan Zhu, Qianqian Li, Mengzhu Zhang, Wei Zhao, Yong Liao, and Pengyuan Zhou. Detect depression from social networks with sentiment knowledge sharing, 2023.

[15] Ziping Zhao, Zhongtian Bao, Zixing Zhang, Jun Deng, Nicholas Cummins, Haishuai Wang, Jianhua Tao, and Björn Schuller. 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, 2019.

[16] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234, 2022.

[17] Salvatore Fara, Orlaith Hickey, Alexandra Georgescu, Stefano Goria, Emilia Molimpakis, and Nicholas Cummins. Bayesian networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data, 2023.

[18] Katharina Schultebraucks, Vijay Yadav, Arieh Y Shalev, George A Bonanno, and Isaac R Galatzer-Levy. Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. *Psychological Medicine*, 52(5):957–967, 2022.

[19] Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. Deep learning for suicide and depression identification with unsupervised label correction, 2021.

[20] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. Depressionnet: A novel summarization boosted deep framework for depression detection on social media, 2021.

[21] Amna Amanat, Muhammad Rizwan, Abdul Rehman Javed, Maha Abdelhaq, Raed Alsaqour, Sharnil Pandya, and Mueen Uddin. Deep learning for depression detection from textual data. *Electronics*, 11(5):676, 2022.

[22] Morteza Rohanian, Julian Hough, Matthew Purver, et al. Detecting depression with word-level multimodal fusion. In *Interspeech*, pages 1443–1447, 2019.

[23] Alice Othmani and Muhammad Muzammel. An ambient intelligence-based approach for longitudinal monitoring of verbal and vocal depression symptoms, 2023.

[24] Alice Othmani, Daoud Kadoch, Kamil Bentounes, Emna Rejaibi, Romain Alfred, and Abdenour Hadid. Towards robust deep neural networks for affect and depression recognition from speech. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 5–19. Springer, 2021.

[25] Nasser Ghadiri, Rasoul Samani, and Fahime Shahrokh. Integration of text and graph-based features for detecting mental health disorders from voice, 2022.

[26] Gleb Kuzmin, Petr Strepetov, Maksim Stankevich, Artem Shelmanov, and Ivan Smirnov. Mental disorders detection in the era of large language models, 2024.

[27] Marc de Gennes, Adrien Lesage, Martin Denais, Xuan-Nga Cao, Simon Chang, Pierre Van Remoortere, Cyrille Dakhlia, and Rachid Riad. Probing mental health information in speech foundation models, 2024.

[28] Habibeh Naderi, Behrouz Haji Soleimani, and Stan Matwin. Multimodal deep learning for mental disorders prediction from audio speech samples, 2020.

[29] Evgeny A Stepanov, Stephane Lathuiliere, Shammur Absar Chowdhury, Arindam Ghosh, Radu-Laurenţiu Vieriu, Nicu Sebe, and Giuseppe Riccardi. Depression severity estimation from multiple modalities. In *2018 ieee 20th international conference on e-health networking, applications and services (healthcom)*, pages 1–6. IEEE, 2018.

[30] Bishal Lamichhane, Nidal Moukaddam, Ankit B. Patel, and Ashutosh Sabharwal. Dyadic interaction assessment from free-living audio for depression severity assessment, 2022.

[31] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720, 2018.

[32] Mashrura Tasnim and Jekaterina Novikova. Cost-effective models for detecting depression from speech. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1687–1694. IEEE, 2022.

[33] Andrea Laguna and Oscar Araque. A cost-aware study of depression language on social media using topic and affect contextualization, 2023.

[34] Syed Arbaaz Qureshi, Mohammed Hasanuzzaman, Sriparna Saha, and Gaël Dias. The verbal and non verbal signals of depression–combining acoustics, text and visuals for estimating depression level. *arXiv preprint arXiv:1904.07656*, 2019.

[35] Hira Dhamyal, Shahan Ali Memon, Bhiksha Raj, and Rita Singh. The phonetic bases of vocal expressed emotion: natural versus acted, 2020.

[36] Junjie Yin, Zixun Chen, Kelai Zhou, and Chongyuan Yu. A deep learning based chatbot for campus psychological therapy, 2019.

[37] Salvatore Fara, Stefano Goria, Emilia Molimpakis, and Nicholas Cummins. Speech and the n-back task as a lens into depression. how combining both may allow us to isolate different core symptoms of depression, 2022.

[38] Sumit Dalal, Sarika Jain, and Mayank Dave. Deep knowledge-infusion for explainable depression detection, 2024.

[39] Andrea Kang, Jun Yu Chen, Zoe Lee-Youngzie, and Shuhao Fu. Synthetic data generation with llm for improved depression prediction, 2024.

[40] Srinivas Parthasarathy and Carlos Busso. Semi-supervised speech emotion recognition with ladder networks, 2019.

[41] Zahraa Al Sahili, Ioannis Patras, and Matthew Purver. Multimodal machine learning in mental health: A survey of data, algorithms, and challenges, 2024.

[42] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. Fraug: A frame rate based data augmentation method for depression detection from speech signals, 2022.

[43] Ziping Zhao, Qifei Li, Nicholas Cummins, Bin Liu, Haishuai Wang, Jianhua Tao, and Björn Schuller. Hybrid network feature extraction for depression assessment from speech. 2020.

[44] Qingkun Deng, Saturnino Luz, and Sofia de la Fuente Garcia. A frame-based attention interpretation method for relevant acoustic feature extraction in long speech depression detection, 2024.

[45] Sebastian Rodriguez, Sri Harsha Dumpala, Katerina Dikaios, Sheri Rempel, Rudolf Uher, and Sageev Oore. Predicting individual depression symptoms from acoustic features during speech, 2024.

[46] Gasser Elbanna, Alice Biryukov, Neil Scheidwasser-Clow, Lara Orlandic, Pablo Mainar, Mikolaj Kegler, Pierre Beckmann, and Milos Cernak. Hybrid handcrafted and learnable audio representation for analysis of speech under cognitive and physical load, 2022.

[47] Arslan Shaukat and Ke Chen. Emotional state categorization from speech: Machine vs. human, 2010.

[48] Juan S. Lara, Mario Ezra Aragon, Fabio A. Gonzalez, and Manuel Montes y Gomez. Deep bag-of-sub-emotions for depression detection in social media, 2021.

[49] Muhammad Muzammel, Hanan Salam, Yann Hoffmann, Mohamed Chetouani, and Alice Othmani. Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis, 2020.

[50] Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52, 2019.

[51] Yichun Li, Shuanglin Li, and Syed Mohsen Naqvi. A novel audio-visual information fusion system for mental disorders detection, 2024.

[52] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. Narrationdep: Narratives on social media for automatic depression detection, 2024.

[53] Hok-Shing Lau, Mark Huntly, Nathon Morgan, Adesua Iyenoma, Biao Zeng, and Tim Bashford. Interpreting pretrained speech models for automatic speech assessment of voice disorders, 2024.

[54] Yuchen Cao, Jianglai Dai, Zhongyan Wang, Yeyubei Zhang, Xiaorui Shen, Yunchong Liu, and Yexin Tian. Machine learning approaches for mental illness detection on social media: A systematic review of biases and methodological challenges, 2025.

[55] Loukas Ilias, Spiros Mouzakitis, and Dimitris Askounis. Calibration of transformer-based models for identifying stress and depression in social media, 2023.

[56] Joshua Y. Kim, Greyson Y. Kim, and Kalina Yacef. Detecting depression in dyadic conversations with multimodal narratives and visualizations, 2020.

[57] Ana-Maria Bucur, Ioana R. Podină, and Liviu P. Dinu. A psychologically informed part-of-speech analysis of depression in social media, 2021.

[58] Dominik Schiller, Silvan Mertes, and Elisabeth André. Embedded emotions – a data driven approach to learn transferable feature representations from raw speech input for emotion recognition, 2020.

[59] Yuezhou Zhang, Amos A Folarin, Judith Dineley, Pauline Conde, Valeria de Angel, Shaoxiong Sun, Yatharth Ranjan, Zulqarnain Rashid, Callum Stewart, Petroula Laiou, Heet Sankesara, Linglong Qian, Faith Matcham, Katie M White, Carolin Oetzmann, Femke Lamers, Sara Siddi, Sara Simblett, Björn W. Schuller, Srinivasan Vairavan, Til Wykes, Josep Maria Haro, Brenda WJH Penninx, Vaibhav A Narayan, Matthew Hotopf, Richard JB Dobson, Nicholas Cummins, and RADAR-CNS consortium. Identifying depression-related topics in smartphone-collected free-response speech recordings using an automatic speech recognition system and a deep learning topic model, 2023.

[60] Pongpak Manoret, Punnatorn Chotipurk, Sompoom Sunpaweravong, Chanati Jantrachotechatchawan, and Kobchai Duangrattanalert. Automatic detection of depression from stratified samples of audio data, 2021.

[61] Asra Fatima, Li Ying, Thomas Hills, and Massimo Stella. Dasentimental: Detecting depression, anxiety and stress in texts via emotional recall, cognitive networks and machine learning, 2021.

[62] Franziska Braun, Sebastian P. Bayerl, Paula A. Pérez-Toro, Florian Hönig, Hartmut Lehfeld, Thomas Hillemacher, Elmar Nöth, Tobias Bocklet, and Korbinian Riedhammer. Classifying dementia in the presence of depression: A cross-corpus study, 2023.

[63] Jinghui Qin, Changsong Liu, Tianchi Tang, Dahuang Liu, Minghao Wang, Qianying Huang, and Rumin Zhang. Mental-perceiver: Audio-textual multi-modal learning for estimating mental disorders, 2025.

[64] Malikeh Ehghaghi, Frank Rudzicz, and Jekaterina Novikova. Data-driven approach to differentiating between depression and dementia from noisy speech and language data, 2022.

[65] Fardin Ahsan Sakib, Ahnaf Atef Choudhury, and Ozlem Uzuner. Mason-nlp at erisk 2023: Deep learning-based detection of depression symptoms from social media texts, 2023.

[66] Junwei Sun, Siqi Ma, Yiran Fan, and Peter Washington. Evaluating large language models for anxiety and depression classification using counseling and psychotherapy transcripts, 2024.

[67] Agnieszka Wołk, Karol Chlasta, and Paweł Holas. Hybrid approach to detecting symptoms of depression in social media entries, 2021.

[68] Eliseo Bao, Anxo Pérez, and Javier Parapar. Explainable depression symptom detection in social media, 2024.

[69] Ashwath Kumar Salimath, Robin K Thomas, Sethuram Ramalinga Reddy, and Yuhao Qiao. Detecting levels of depression in text based on metrics, 2018.

[70] Yanrong Guo, Chenyang Zhu, Shijie Hao, and Richang Hong. A topic-attentive transformer-based model for multimodal depression detection, 2022.

[71] Tomasz Rutowski, Amir Harati, Yang Lu, and Elizabeth Shriberg. Optimizing speech-input length for speaker-independent depression classification, 2024.

[72] Qiang Li, Yufeng Wu, Zhan Xu, and Hefeng Zhou. Exploration of adolescent depression risk prediction based on census surveys and general life issues, 2024.

[73] Juho Jung, Chaewon Kang, Jeewoo Yoon, Seungbae Kim, and Jinyoung Han. Hique: Hierarchical question embedding network for multimodal depression detection, 2024.

[74] Sri Harsha Dumpala, Chandramouli Shama Sastry, Rudolf Uher, and Sageev Oore. Test-time training for depression detection, 2024.

[75] Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media, 2023.

[76] Ana-Maria Bucur, Marcos Zampieri, and Liviu P. Dinu. An exploratory analysis of the relation between offensive language and mental health, 2021.

[77] Elysia Shi, Adithri Manda, London Chowdhury, Runeema Arun, Kevin Zhu, and Michael Lam. Enhancing depression diagnosis with chain-of-thought prompting, 2024.

[78] Loris Belcastro, Riccardo Cantini, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Detecting mental disorder on social media: a chatgpt-augmented explainable approach, 2024.

[79] Adam Valen Levinson, Abhay Goyal, Roger Ho Chun Man, Roy Ka-Wei Lee, Koustuv Saha, Nimay Parekh, Frederick L. Altice, Lam Yin Cheung, Munmun De Choudhury, and Navin Kumar. Using audio data to facilitate depression risk assessment in primary health care, 2023.

[80] Fuxiang Tao. *Speech-based automatic depression detection via biomarkers identification and artificial intelligence approaches*. PhD thesis, University of Glasgow, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.