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

www.surveyx.cn

Abstract

The interdisciplinary application of deep learning techniques in analyzing speech for emotion recognition and depression detection is pivotal in advancing mental health diagnostics. This survey explores the integration of multimodal data, including audio, text, and visual inputs, to enhance the accuracy and robustness of diagnostic models. Highlighted methodologies include hybrid networks combining self-attention and deep convolutional neural networks, which demonstrate superior performance in early-stage depression diagnosis. The potential of multimodal machine learning approaches is emphasized, addressing data organization and scarcity issues, as evidenced by advancements in large language models and clinically validated benchmarks. Ensemble methods and adversarial disentanglement techniques, such as the NUSD method, show significant improvements in detection accuracy while preserving privacy. Furthermore, Bayesian Networks offer robust support for clinical decision-making in diagnosing Major Depressive Disorder. The survey underscores the importance of addressing ethical concerns, developing standardized evaluation methods, and enhancing model generalizability. Future research directions focus on expanding datasets, refining feature extraction, and integrating AI strategies for real-time monitoring. These advancements promise to improve mental health care by providing more accurate, reliable, and accessible diagnostic tools.

1 Introduction

1.1 Significance of Depression Detection

Depression is a widespread mental health disorder affecting approximately 3.8% of the global population and contributing to over 700,000 suicides annually [1]. Its increasing prevalence across all age groups necessitates effective and timely detection, which remains a significant challenge [2]. Traditional diagnostic methods, including clinical interviews and self-reported questionnaires, are resource-intensive and subjective, often leading to misdiagnosis and inadequate treatment [3]. Consequently, there is a pressing need for automatic detection techniques to facilitate timely care [2].

Major Depressive Disorder (MDD) affects over 264 million individuals worldwide, highlighting the urgent requirement for innovative diagnostic approaches in mental health care [4]. Utilizing speech-based analysis with deep learning techniques provides a promising non-invasive method for early identification of depressive symptoms, thereby enabling timely assessment and intervention [5]. By analyzing emotional cues in speech, these systems enhance the consistency and sensitivity of clinical assessments for MDD [6]. The integration of such technologies is essential as mental health issues increasingly impact the global economy, with conditions like depression and anxiety resulting in trillions in lost productivity [7].

Moreover, the development of speech-based detection systems fosters personalized treatment strategies, potentially mitigating the stigma associated with seeking mental health support. By employing

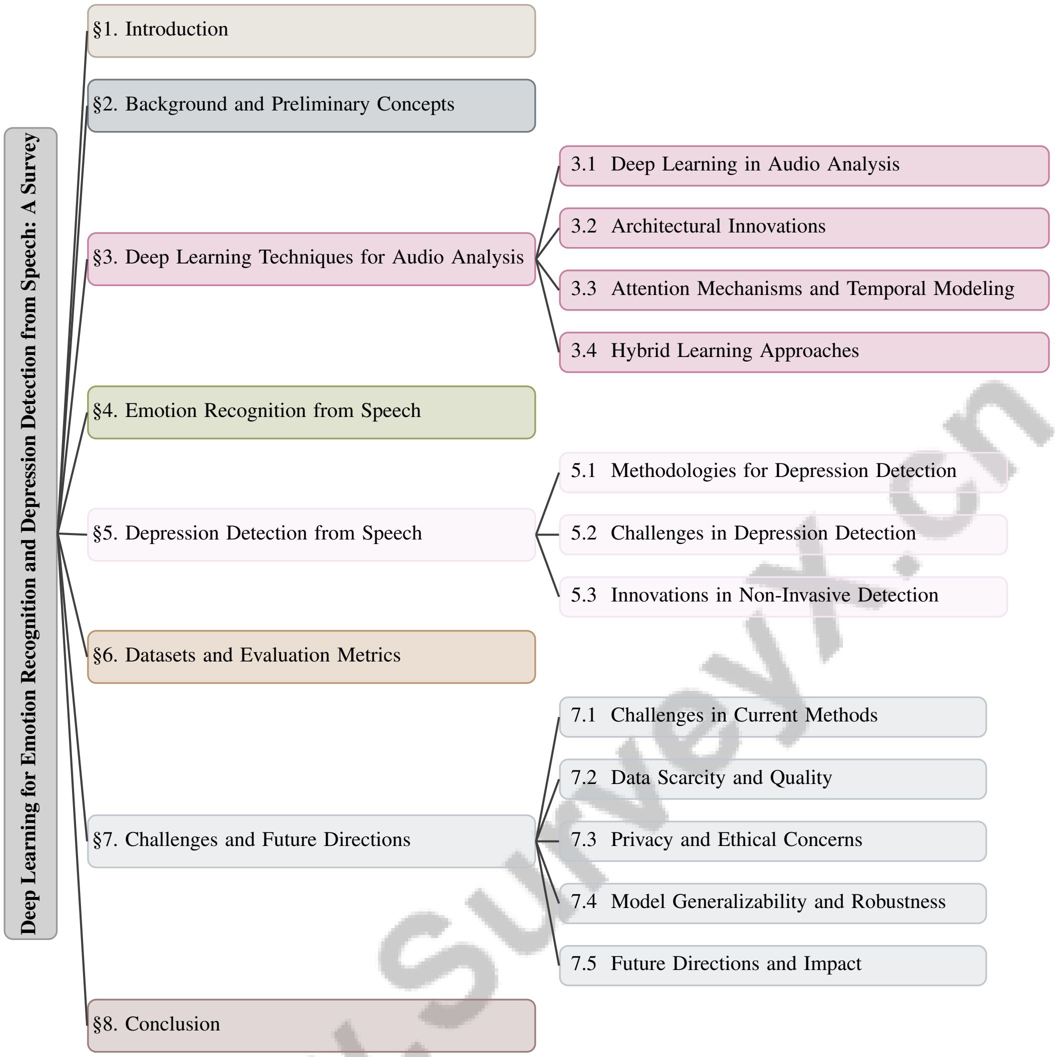

Deep Learning for Emotion Recognition and Depression Detection from Speech: A Survey
§1. Introduction
§2. Background and Preliminary Concepts
§3. Deep Learning Techniques for Audio Analysis
3.1 Deep Learning in Audio Analysis
3.2 Architectural Innovations
3.3 Attention Mechanisms and Temporal Modeling
3.4 Hybrid Learning Approaches
§4. Emotion Recognition from Speech
§5. Depression Detection from Speech
5.1 Methodologies for Depression Detection
5.2 Challenges in Depression Detection
5.3 Innovations in Non-Invasive Detection
§6. Datasets and Evaluation Metrics
§7. Challenges and Future Directions
7.1 Challenges in Current Methods
7.2 Data Scarcity and Quality
7.3 Privacy and Ethical Concerns
7.4 Model Generalizability and Robustness
7.5 Future Directions and Impact
§8. Conclusion

Figure 1: chapter structure

deep learning for speech analysis, healthcare providers can better address diverse populations' mental health needs, ultimately improving outcomes and quality of life [3].

1.2 Role of Deep Learning in Emotion Recognition

Deep learning has significantly advanced emotion recognition from speech, employing sophisticated methodologies to discern intricate patterns in audio data. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are crucial in this field, with CNNs excelling in processing high-dimensional data and extracting salient audio features. Their efficacy is enhanced when combined with time-frequency representations and visual spectrograms, vital for improving emotion and depression recognition. LSTM networks, noted for modeling temporal sequences, further augment emotion recognition by capturing inherent temporal dependencies in speech, thus increasing model accuracy [8].

The integration of multimodal approaches has enriched emotion recognition systems by combining audio, video, and textual data to enhance predictive accuracy. For instance, word-level multimodal fusion methods incrementally integrate audio and visual information alongside lexical content, thereby improving depression detection accuracy [9]. Attention-based neural networks, particularly in sub-attentional models, prioritize relevant features across modalities, enhancing both interpretability and performance [10]. This focus on interpretability is essential, as current methods often lack the transparency required for clinical applications [11].

AI-generated, for reference only.

Self-supervised learning (SSL)-based speech models have emerged as promising tools for identifying individual symptoms of depression, providing a nuanced understanding of the disorder [12]. The development of large-scale multimodal datasets is critical for advancing depression detection methods, enabling the training of robust models capable of distinguishing subtle emotional states and enhancing emotion recognition systems [13]. Furthermore, deep learning algorithms analyze facial expressions, voice prosody, and speech content to classify PTSD and MDD based on free speech responses from patients, demonstrating the versatility of these technologies in mental health diagnostics [14].

1.3 Structure of the Survey

This survey is meticulously structured to provide a comprehensive overview of the application of deep learning techniques in emotion recognition and depression detection from speech. The paper begins with an **Introduction**, discussing the significance of detecting depression through speech and the pivotal role of deep learning in emotion recognition. This is followed by a detailed **Background and Preliminary Concepts** section, elucidating fundamental concepts such as depression, emotion recognition, deep learning, and audio analysis, emphasizing the interdisciplinary nature of this research.

In the **Deep Learning Techniques for Audio Analysis** section, various deep learning methodologies, including CNNs, RNNs, and transformers, are explored for their effectiveness in processing speech data. This section also covers architectural innovations, attention mechanisms, temporal modeling, and hybrid learning approaches.

The subsequent section, **Emotion Recognition from Speech**, examines the process of recognizing emotions from speech, emphasizing audio features, associated challenges, and the concept of dimensional emotion recognition.

The survey then transitions to **Depression Detection from Speech**, providing an in-depth examination of advanced methodologies, including automated algorithms that leverage audio and text features to assess depression through sequential modeling of interviews. It highlights challenges in this domain, such as the need for large-scale validation and limitations of traditional diagnostic methods, while discussing recent innovations in non-invasive detection techniques, including the development of specialized corpora like the Emotional Audio-Textual Depression Corpus (EATD-Corpus) and the application of deep learning models for more accurate and accessible depression screening [15, 16, 17].

In the **Datasets and Evaluation Metrics** section, we provide an overview of datasets frequently utilized in depression detection research, such as the Distress Analysis Interview Corpus (DAIC) and a newly developed Social Media Corpus. We also delve into evaluation metrics crucial for assessing model performance, highlighting complexities in measuring depression through textual data and challenges posed by non-standard language in social media. Additionally, we discuss implications for practitioners and suggest areas for future research [18, 19].

The penultimate section, **Challenges and Future Directions**, identifies current challenges such as data scarcity, privacy concerns, and model generalizability, while proposing potential future research directions.

In the **Conclusion**, we synthesize principal findings from our analysis of the relationship between offensive language and mental health, the part-of-speech characteristics of discourse among social media users with depression, and advancements in explainable AI for detecting depressive symptoms. We emphasize the critical need for ongoing research to enhance mental health diagnostics and treatment, particularly through the development of robust computational models and social media corpora that can better capture user expression nuances and improve the interpretability of AI-driven assessments. This research underscores the significance of understanding language patterns associated with mental health and the potential for AI to support clinicians in identifying and addressing mental health issues more effectively [20, 21, 22, 18, 23].The following sections are organized as shown in Figure 1.

2 Background and Preliminary Concepts

2.1 Definitions and Explanations of Depression

Depression is a multifaceted mental disorder characterized by persistent sadness, disinterest, and various emotional and physical symptoms that significantly disrupt daily life. Major Depressive Disorder (MDD) is categorized into normal, moderate, and severe levels, highlighting the complexity of accurate diagnosis [4]. Traditional diagnostic methods, including clinician judgment and semi-structured interviews, often falter due to individual variability in speech and the limitations of singular diagnostic features.

The symptom overlap between depression and other mood disorders like bipolar disorder complicates accurate diagnosis [14]. Traditional questionnaires, being subjective, often miss objective behavioral indicators, leading to misdiagnoses [24]. This is particularly problematic for early detection, where subtle signs are crucial yet often hindered by data scarcity [25].

Advancements in deep learning and multimodal data analysis offer promising non-invasive diagnostic methods, utilizing vocal characteristics, facial expressions, and cognitive assessments to provide a comprehensive mental state view [11]. For example, changes in vocal fold behavior, linked to neurophysiological alterations in depression, such as slowed speech and monotonous pitch, serve as valuable diagnostic indicators [26]. Integrating audio, visual, and textual data enhances diagnostic accuracy by offering a holistic view of the individual's mental state.

Current methods' limitations in modeling social media narratives highlight the need for innovative approaches to improve detection accuracy. The lack of interpretability in deep learning models for Automatic Speech Assessment further restricts their effectiveness in identifying voice disorders [11]. Developing models with explanatory capabilities is crucial for user engagement and medical reference [27].

Depression's complexity and diagnostic challenges make it a focal point in mental health research. Integrating advanced technologies like deep learning and multimodal data analysis offers significant opportunities to enhance depression detection methods' accuracy and effectiveness. Automated systems leveraging deep learning algorithms to analyze audiovisual cues and textual data can identify subtle depression indicators overlooked in traditional assessments. Studies show that variations in speech patterns, facial expressions, and written content can serve as objective depression markers, facilitating early diagnosis and intervention. These approaches aim to address challenges posed by the doctor-patient ratio and improve mental health outcomes by enabling timely recognition and treatment [28, 27, 9].

2.2 Emotion Recognition and Its Relevance to Mental Health

Emotion recognition from speech is crucial in mental health research, offering a non-invasive and efficient means of assessing emotional states, which are critical indicators of psychological conditions [29]. Accurate emotion detection is particularly important in diagnosing and treating disorders like depression, where emotional dysregulation is prevalent. Traditional sentiment analysis often fails to capture emotional distress's complexity, highlighting the need for advanced techniques [30].

Integrating emotion recognition systems into mental health diagnostics can lead to more efficient assessments, utilizing speech data to derive insights into an individual's emotional state without intrusive measures [31]. Deep learning techniques have transformed emotion recognition, enabling more accurate detection of disorders like MDD through speech analysis. However, developing models that generalize across diverse datasets remains challenging, as current Speech Emotion Recognition (SER) models often perform poorly on new data due to a lack of diverse labeled datasets.

Social media platforms, especially Twitter, are valuable tools for monitoring mental health at a population level, reflecting users' mental states through their language. Emotion recognition from social media content is essential for mental health, facilitating depression identification through user-generated text [30]. Additionally, distinguishing depression from other cognitive disorders, such as dementia, via speech analysis enhances diagnosis and treatment, contributing to a nuanced understanding of mental health conditions.

Exploring hybrid models and ensemble approaches deepens the comprehension of mental health conditions as expressed through language, enabling multiclass classification of disorders such as

AI-generated, for reference only.

anxiety, depression, and PTSD. This advancement utilizes sophisticated natural language processing techniques, including transformer-based architectures like BERT and RoBERTa, and linguistic feature analysis to capture nuanced language pattern differences associated with specific mental health issues. Consequently, these methodologies enhance emotion recognition capabilities and support developing targeted clinical applications for early intervention and monitoring [32, 33, 34, 35, 36]. Structured evaluations of machine learning models emphasize the importance of detecting depression from speech, thereby improving diagnostic accuracy.

Emotion recognition from speech has emerged as a vital tool in mental health research, enabling valuable insights into individuals' emotional states, crucial for accurate diagnosis and effective treatment. This approach employs advanced techniques such as hybrid neural networks and machine learning models to analyze speech patterns, facilitating the identification of depression and other conditions through objective markers. Recent studies demonstrate that analyzing acoustic features and integrating dialogue patterns can enhance mood episode detection and disorder severity, promoting early intervention strategies that improve outcomes [37, 38, 39, 35, 40]. The ongoing development of generalizable models and diverse data integration continues to enhance emotion recognition systems' potential in mental health care.

2.3 Interdisciplinary Nature of the Research

The interdisciplinary nature of research in emotion recognition and depression detection from speech is evident in integrating mental health insights, audio analysis, and advanced deep learning methodologies. This convergence fosters a comprehensive understanding of mental health disorders and enhances diagnostic precision. Combining topic modeling with multimodal analysis illustrates how merging mental health, audio analysis, and deep learning can enrich assessments [6].

Deep learning architectures capable of processing time-sequential data have proven effective in detecting depressive symptoms, showcasing the potential of integrating textual and audio data for nuanced diagnostics [5]. Resources like the Emotional Audio-Textual Depression Corpus (EATD-Corpus) serve as valuable assets for future research [15].

A significant challenge in this field is developing comprehensive diagnostic methods that objectively assess depression's complexities. Current methods often underutilize the diverse information available in speech data, necessitating innovative approaches that can effectively integrate and analyze multimodal inputs [3]. The exploration of large-scale language models is motivated by the observation that multimodal data can yield richer insights into mental health than single modalities, highlighting the importance of interdisciplinary research in advancing diagnostics [7].

In recent years, deep learning has revolutionized the field of audio analysis, leading to significant advancements in various applications, including emotion recognition and depression detection. To better understand this evolution, it is essential to examine the hierarchical structure of deep learning techniques, which can be visualized in Figure 2. This figure illustrates the categorization of key concepts into four main areas: Deep Learning in Audio Analysis, Architectural Innovations, Attention Mechanisms and Temporal Modeling, and Hybrid Learning Approaches. Each of these areas is further dissected into specific techniques and applications, thereby providing a comprehensive overview of the innovations that have emerged within this domain. By analyzing these categories, we can appreciate the intricate relationships between different methodologies and their contributions to the advancement of audio analysis technologies.

3 Deep Learning Techniques for Audio Analysis

3.1 Deep Learning in Audio Analysis

Deep learning has revolutionized audio analysis, particularly in enhancing emotion recognition and depression detection from speech. Advanced architectures, such as EmoAudioNet, leverage time-frequency representations of audio signals to accurately identify emotional states and diagnose conditions like Major Depressive Disorder (MDD), outperforming traditional methods in precision and recall [35, 41, 42, 43]. This is crucial for improving mental health assessments, especially in regions with limited psychiatric resources, where deep learning aids early diagnosis and intervention. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, are pivotal, with CNNs adept at extracting features from

AI-generated, for reference only.

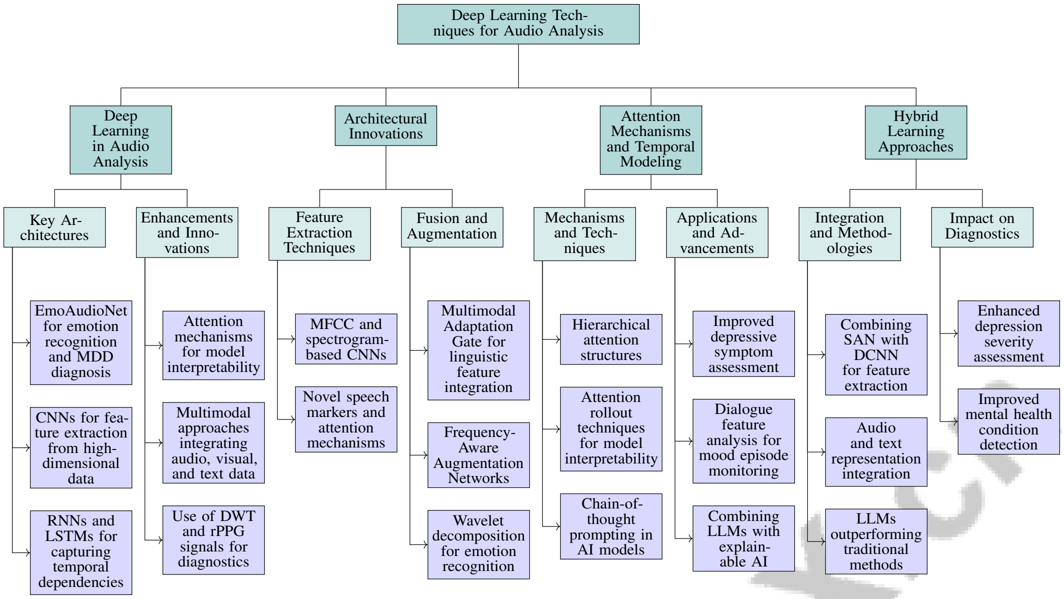

Deep Learning Techniques for Audio Analysis
Deep Learning in Audio Analysis
Architectural Innovations
Attention Mechanisms and Temporal Modeling
Hybrid Learning Approaches
Key Architectures
Enhancements and Innovations
Feature Extraction Techniques
Fusion and Augmentation
Mechanisms and Techniques
Applications and Advancements
Integration and Methodologies
Impact on Diagnostics
EmoAudioNet for emotion recognition and MDD diagnosis
CNNs for feature extraction from high-dimensional data
RNNs and LSTMs for capturing temporal dependencies
Attention mechanisms for model interpretability
Multimodal approaches integrating audio, visual, and text data
Use of DWT and rPPG signals for diagnostics
MFCC and spectrogram-based CNNs
Novel speech markers and attention mechanisms
Multimodal Adaptation Gate for linguistic feature integration
Frequency-Aware Augmentation Networks
Wavelet decomposition for emotion recognition
Hierarchical attention structures
Attention rollout techniques for model interpretability
Chain-of-thought prompting in AI models
Improved depressive symptom assessment
Dialogue feature analysis for mood episode monitoring
Combining LLMs with explainable AI
Combining SAN with DCNN for feature extraction
Audio and text representation integration
LLMs outperforming traditional methods
Enhanced depression severity assessment
Improved mental health condition detection

Figure 2: This figure illustrates the hierarchical structure of deep learning techniques for audio analysis, categorizing key concepts into four main areas: Deep Learning in Audio Analysis, Architectural Innovations, Attention Mechanisms and Temporal Modeling, and Hybrid Learning Approaches. Each area further breaks down into specific techniques and applications, highlighting the advancements in emotion recognition and depression detection through innovative architectures, attention mechanisms, and hybrid models.

high-dimensional audio data and RNNs capturing temporal dependencies essential for analyzing speech [3].

Attention mechanisms further enhance model interpretability and performance by focusing on informative data segments. Hierarchical attention, as used in ABAFnet, fuses multiple acoustic features for depression detection [6]. Multimodal approaches integrating audio, visual, and textual data via RNNs with gating mechanisms improve depression detection, exemplified by methods like MMFF, which utilize shared latent proxies for data fusion [1]. Discrete Wavelet Transform (DWT) and remote Photoplethysmography (rPPG) signal extraction from facial videos illustrate deep learning's potential in objective diagnostics [29, 31].

The ongoing development of novel architectures and hybrid approaches is transforming audio analysis, promising robust models that leverage speech data's rich information. These advancements aim to improve emotion and depression detection systems, achieving greater accuracy and reliability than traditional methods. The exploration of large-scale language models, such as GPT-4o, highlights multimodal capabilities' potential in advancing audio analysis [7].

3.2 Architectural Innovations

Recent deep learning innovations have significantly advanced audio analysis for emotion recognition and depression detection. Integrating Mel Frequency Cepstral Coefficients (MFCC)-based and spectrogram-based CNNs enhances feature extraction, improving emotional state classification [43, 42]. Novel speech markers and attention mechanisms increase depression detection accuracy by focusing on relevant speech segments [44]. Hierarchical attention-based clustering models tweets and their clusters for improved social media-based depression detection [45].

Multimodal fusion innovations, such as the Multimodal Adaptation Gate, effectively integrate linguistic features, complementing label smoothing for transformer-based models [46]. Frequency-Aware Augmentation Networks combining CNNs and Gated Recurrent Units (GRUs) capture local and global voice data patterns, enhancing detection capabilities [47]. Wavelet decomposition further advances emotion recognition, outperforming traditional feature methods [29].

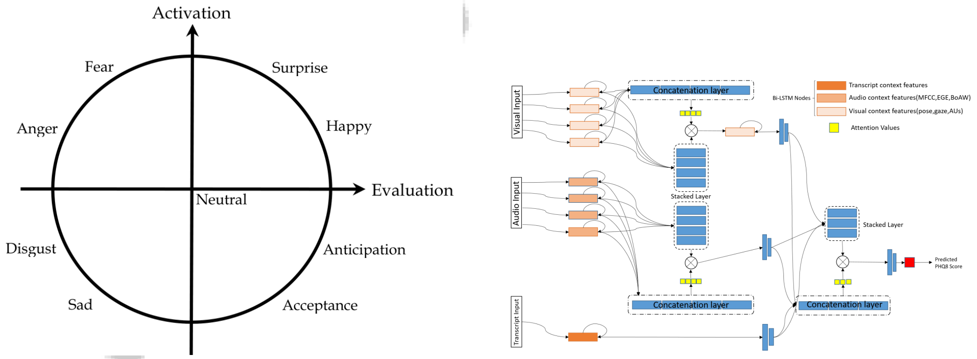
Late fusion strategies, like those in ABAFnet, dynamically adjust feature weights, improving depression pattern capture [3]. Processing speech spectrograms with deep learning enhances depression screening accuracy, offering a robust alternative to traditional methods [48]. These architectural innovations underscore deep learning research's dynamic nature in audio analysis, continually advancing emotion recognition and depression detection.

3.3 Attention Mechanisms and Temporal Modeling

Attention mechanisms and temporal modeling are pivotal in advancing speech data processing for emotion recognition and depression detection. They enhance prediction accuracy and interpretability by focusing on informative speech segments. Hierarchical attention structures, such as those by Zhao et al., mirror speech's sequential nature, using attention transfer to boost model performance [49]. LSTM networks with attention layers effectively classify depression-specific emotions, leveraging LSTM's temporal dependency modeling and attention's feature highlighting [50].

Attention rollout techniques visualize pretrained speech models' focus on phoneme regions, increasing interpretability and clinical applicability [51]. These mechanisms capture verbal and non-verbal communication patterns influenced by mood episodes, providing insights into emotional and mental states [39]. Focusing on relevant speech data features enhances model performance in resource-constrained scenarios [25].

The integration of attention mechanisms and temporal modeling in speech data analysis represents a significant advancement, enabling more accurate and interpretable models for emotion recognition and depression detection. These techniques are at the forefront of mental health diagnostics, creating sophisticated systems to tackle the complex challenges of identifying mental health conditions. For instance, chain-of-thought prompting in AI models improves depressive symptom assessment accuracy, and large language models outperform traditional methods in detecting mental health disorders in noisy datasets. Dialogue feature analysis during clinical interviews enhances mood episode monitoring in bipolar disorder. Combining LLMs with explainable AI paves the way for interpretable mental health assessments on social media, enhancing diagnostic precision and accessibility [38, 39, 22, 52, 36].



(a) The image represents a circular diagram with four quadrants, each labeled with a different emotion.[53]

(b) A neural network model for predicting PHQ8 scores[54]

Figure 3: Examples of Attention Mechanisms and Temporal Modeling

In Figure 3, attention mechanisms and temporal modeling play crucial roles in deep learning for audio analysis, enhancing emotion and psychological state understanding from audio data. The first image categorizes emotions into quadrants—Fear, Surprise, Happy, and Neutral—based on evaluation and activation axes, illustrating attention mechanisms' role in discerning emotional nuances from audio inputs. The second image depicts a neural network model predicting PHQ8 scores, integrating multimodal inputs: visual, audio, and text. This model demonstrates temporal modeling's ability to capture sequential dependencies and contextual information across data modalities, highlighting deep learning's versatility and efficacy in nuanced emotional and psychological assessments [53, 54].

AI-generated, for reference only.

3.4 Hybrid Learning Approaches

Hybrid learning approaches in audio analysis enhance emotion recognition and depression detection by integrating multiple deep learning models, exploiting their complementary strengths for improved accuracy and robustness. One example is combining self-attention networks (SAN) with deep convolutional neural networks (DCNN) to extract complementary speech data features, enhancing emotional and depressive state detection [35].

Integrating audio and text representations exemplifies hybrid approaches, utilizing multimodal data to classify individuals as depressed or non-depressed. This method capitalizes on rich speech and textual information, offering a nuanced understanding of mental health conditions [55]. Large language models (LLMs) in hybrid approaches outperform traditional methods, particularly in noisy datasets, achieving higher F1-scores and demonstrating better generalization and reliability [36].

Hybrid learning approaches significantly advance audio analysis for mental health diagnostics by integrating techniques like self-attention networks and deep convolutional neural networks to extract and combine relevant speech signal features. This multifaceted methodology enhances depression severity assessment accuracy, as demonstrated by studies utilizing diverse datasets and advanced machine learning models to improve mental health condition detection through speech analysis [38, 56, 34, 35]. By combining different deep learning models, these approaches harness each model's strengths, leading to more accurate and robust systems capable of addressing emotion recognition and depression detection complexities.

4 Emotion Recognition from Speech

4.1 Role of Audio Features

Audio features are pivotal in emotion recognition systems, providing critical data for identifying emotional states from speech. These features are divided into low-level descriptors (LLDs) like pitch, energy, and spectral features, which capture speech's core characteristics, and high-level statistical functions (HSFs), derived from LLDs, offering a comprehensive representation of emotional content [57]. The extraction and analysis of these features are essential for understanding the complexities of emotional expression.

Advanced methodologies such as dual-layer Long Short-Term Memory (LSTM) networks have been developed to capture long-term emotional dependencies more effectively. Stacking two LSTM layers enhances the modeling of temporal sequences, crucial for accurate emotion recognition from speech [58]. This underscores the importance of capturing both short-term and long-term dependencies in audio features to achieve high accuracy.

Integrating audio features with linguistic data significantly improves emotion recognition by utilizing deep learning techniques to automatically learn meaningful representations from spoken language and text. This approach enhances classification performance across diverse datasets and real-world scenarios [59, 60, 61, 53, 62]. By combining these modalities, systems gain a holistic understanding of emotional states, considering both verbal and non-verbal cues. This comprehensive analysis is particularly beneficial in contexts like depression detection, where audio and linguistic features provide critical insights.

Audio features are indispensable for accurately discerning emotional states in recognition systems. Ongoing advancements in feature extraction and integration techniques, including text and graph-based feature fusion, large language models, and hybrid networks for speech analysis, significantly enhance mental health diagnostic systems' effectiveness and reliability. These innovations facilitate earlier detection and more accurate assessments of mental health disorders like depression and anxiety by leveraging diverse data sources such as voice recordings and multimodal inputs, ultimately improving clinical outcomes and streamlining the diagnostic process [40, 63, 35, 36].

4.2 Challenges in Emotion Recognition

Emotion recognition from speech encounters several challenges that hinder the development of effective and reliable systems. A significant challenge in mental health research is extracting robust, discriminative, and depression-relevant features from lengthy audio recordings, crucial for

AI-generated, for reference only.

accurate speech-based depression assessment. This involves leveraging advanced techniques like hybrid networks that integrate self-attention mechanisms with deep convolutional neural networks to improve feature representation. Addressing segment-level labeling noise and ensuring model interpretability are vital for enhancing these tools' clinical applicability, as recent studies emphasize the importance of analyzing longer-duration speech for reliable depression detection [64, 65, 35]. The challenge lies in isolating emotional cues from features that may also convey the speaker's identity, raising privacy concerns while ensuring the relevance of captured features to the emotional states being analyzed.

Another critical challenge is the scarcity of studies that benchmark automatic categorization systems against human performance within the same experimental framework. This gap creates uncertainties regarding these systems' effectiveness, as they are often unassessed against human capabilities [53]. Without such comparative analyses, evaluating the true performance of automated systems in real-world scenarios becomes difficult.

The reliance on subjective self-reported questionnaires and clinical interviews further complicates emotion recognition, as these traditional methods can be influenced by various factors, limiting their reliability and complicating the acquisition of accurate ground truth labels for training and evaluating emotion recognition systems [66]. Thus, there is an urgent need for objective and consistent methods to capture emotional states.

Existing methods often overlook essential features in speech signals critical for emotion recognition, leading to systems that struggle to achieve high accuracy, especially in distinguishing subtle emotional nuances [47]. Traditional approaches, frequently reliant on speaker-dependent features, have demonstrated less than 70% accuracy, highlighting the limitations of current methodologies in effectively recognizing emotions from speech [29].

Addressing these challenges necessitates innovative feature extraction techniques, improved benchmarking against human performance, and establishing more objective methods for capturing emotional states. These efforts are essential for enhancing the accuracy and reliability of emotion recognition systems in real-world applications, particularly as advancements in deep learning and data augmentation techniques facilitate the extraction of meaningful features from diverse input sources such as speech, text, and audio-visual data. By leveraging large datasets and innovative algorithms, these systems can better identify and classify emotions, crucial for applications in healthcare, where early detection of conditions such as depression and stress can significantly enhance patient outcomes [59, 60, 67, 68, 18].

4.3 Dimensional Emotion Recognition

Dimensional emotion recognition signifies a shift from traditional categorical approaches, offering a more nuanced understanding of emotional states by mapping them onto continuous dimensions. This method typically assesses emotions along axes such as valence, arousal, and dominance, forming a multidimensional space for emotion representation [57]. This approach provides a comprehensive characterization of emotions, capturing the intensity and subtleties that categorical models often miss.

Applying dimensional emotion recognition in speech analysis allows for capturing dynamic emotional changes over time, crucial for understanding complex emotional states. This is particularly relevant in mental health contexts, where emotions can fluctuate significantly in response to various stimuli [58]. By employing dimensional models, researchers can gain deeper insights into individuals' emotional trajectories, facilitating more accurate and personalized assessments.

Advanced deep learning techniques, such as multilayer perceptrons, effectively model these continuous emotional dimensions. These models process a wide range of audio features to predict dimensional scores, enabling a detailed analysis of emotional states [57]. Integrating dimensional emotion recognition with multimodal data sources, including visual and textual inputs, further enhances the understanding and interpretation of complex emotional expressions.

The application of dimensional emotion recognition is particularly advantageous for mental health diagnostics, where understanding the intensity and evolution of emotions can yield valuable insights into conditions like depression and anxiety. Acknowledging the complex and continuous nature of emotions, this innovative approach improves the development of advanced diagnostic tools that more accurately capture the nuances of human emotional experiences, especially in identifying mental health conditions such as depression, anxiety, and stress. This is accomplished through sophisticated machine learning models analyzing emotional recall and cognitive networks, providing a deeper understanding of the semantic dimensions underlying emotional states and improving automated screening methods' effectiveness [34, 68, 17, 19].

5 Depression Detection from Speech

In recent years, the intersection of technology and mental health has garnered significant attention, particularly in the realm of depression detection. This section explores the various methodologies that have emerged, showcasing how advancements in artificial intelligence and machine learning are being harnessed to improve diagnostic accuracy and understanding of depressive disorders. By examining innovative approaches and their implications, we can better appreciate the potential of these technologies in addressing the complexities of depression. The following subsection delves into specific methodologies for depression detection, highlighting key techniques and their contributions to the field.

5.1 Methodologies for Depression Detection

Recent advancements in methodologies for detecting depression from speech have significantly leveraged deep learning, multimodal integration, and innovative feature extraction techniques to enhance diagnostic accuracy and interpretability. A prominent approach involves the use of Multi-Stage Dilated CNN-LSTM Models (MS-DCL), which effectively capture temporal dependencies in speech data to classify depression severity [4]. This technique underscores the importance of modeling temporal relationships within speech signals to improve classification outcomes.

The integration of multimodal data has been a key innovation, as exemplified by the Multimodal Fusion Framework (MMFF), which exploits modality interactions at multiple levels to enhance both the effectiveness and interpretability of depression detection models [1]. This framework combines audio, text, and visual features, allowing for a comprehensive analysis of depressive symptoms across different modalities.

Another effective methodology involves segmenting interviews based on identified topics and extracting audio, video, and semantic features for each topic. This approach constructs a feature vector that includes topic presence and key topics, facilitating a nuanced understanding of depressive states [6]. The ability to segment and analyze data based on thematic content enhances the depth of analysis and the accuracy of depression detection.

Deep learning models such as Long Short-Term Memory (LSTM) networks have been employed to analyze and predict depression from textual data, capturing temporal dependencies that are indicative of depressive symptoms [5]. The use of attention mechanisms further enhances feature representation and classification accuracy by integrating various speech features into a comprehensive model [3].

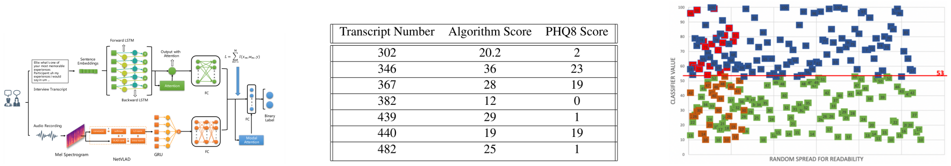
The challenge of data scarcity in detecting Major Depressive Disorder (MDD) from speech signals has been addressed by developing novel data augmentation techniques that preserve critical acoustic information, thereby enhancing the robustness of depression detection models [69]. These techniques are crucial in overcoming the limitations of conventional data augmentation methods, which often fail to maintain the integrity of essential speech features.

Moreover, methodologies that integrate audio and textual features in modeling depression relapse detection through similarity assessment have shown promise in capturing the nuances of depressive states [70]. The automatic prediction of PHQ-8 scores from various modalities, including speech, language, and visual features, has been benchmarked to enhance the detection and monitoring of depression severity [2].

The methodologies discussed in these references highlight a range of innovative strategies for utilizing speech and multimodal data—such as audio, text, and visual cues—to enhance the accuracy and effectiveness of depression detection. These approaches include advanced techniques like word-level multimodal fusion, topic-attentive transformer models, and sequence modeling of interviews, which collectively demonstrate the potential of integrating diverse data sources to improve diagnostic capabilities in automated depression assessment. [71, 15, 65, 17, 9]. The continuous refinement of

AI-generated, for reference only.

these approaches promises to enhance the accuracy and accessibility of depression detection systems, ultimately improving outcomes for individuals affected by this pervasive mental health disorder.



Transcript Number	Algorithm Score	PHQ8 Score
302	20.2	2
346	36	23
367	28	19
382	12	0
439	29	1
440	19	19
482	25	1

(a) A neural network model for multimodal speech recognition[55]

(b) Transcript Number and Scores[19]

(c) Scatter Plot of Random Spread for Readability vs. Classifier Value[34]

Figure 4: Examples of Methodologies for Depression Detection

As shown in Figure 4, The exploration of methodologies for depression detection from speech is a burgeoning area of research that leverages advanced computational techniques to identify signs of depression through vocal and textual data. The example provided showcases three distinct approaches within this domain. Firstly, a neural network model for multimodal speech recognition is illustrated, which ingeniously integrates both speech and text inputs to detect depression. This model is structured with various layers, including LSTM and attention layers, to effectively process and classify the data. Secondly, a tabular representation of transcript numbers and scores is presented, highlighting the correlation between algorithm-generated scores and PHQ8 scores, a widely used measure for assessing depression severity. Lastly, a scatter plot offers a visual depiction of the relationship between 'Random Spread for Readability' and 'Classifier Value', providing insights into the variability and effectiveness of the classification method. Together, these methodologies underscore the potential of combining speech and text analysis with machine learning to enhance the accuracy and reliability of depression detection systems. [**?**
]shen2022automaticdepressiondetectionemotional,salimath2018detectinglevelsdepressiontext,wok2021hybridapproachdetectings

5.2 Challenges in Depression Detection

Depression detection from speech is fraught with numerous challenges that hinder the development of effective diagnostic systems. A significant challenge is the reliance on clinical variables for predicting depression relapse, which often neglects critical factors such as speech patterns that could provide valuable insights into an individual's mental state [70]. Traditional methods predominantly focus on binary classification, which lacks the granularity required to discern varying levels of depression severity. This limitation impedes the ability to provide nuanced diagnoses and tailored interventions [4].

The processing of long audio and video data presents another challenge, as statistical functions applied to short-term features often result in the loss of crucial temporal information. This complicates the feature extraction process and can lead to inaccuracies in depression detection [6]. Moreover, many existing approaches are heavily reliant on specific datasets, raising concerns about the generalizability of these models across different languages and cultural contexts. Without proper adaptation, these models may fail to perform effectively in diverse settings [5].

The complexity and extended training times associated with multi-feature approaches also pose a significant barrier. While these methods can enhance model accuracy, they may not be suitable for all practical applications due to their computational demands [3]. This issue is exacerbated by the scarcity of large, diverse datasets that are essential for training robust models capable of generalizing across various populations.

To effectively tackle the challenges in mental health detection from social media data, it is essential to develop innovative solutions that enhance data diversity by creating representative corpora, improve model adaptability through the use of advanced hybrid and ensemble techniques, and streamline computational processes by leveraging sophisticated natural language processing methods and diverse linguistic features. [32, 18]. The development of more granular classification systems and the integration of culturally and linguistically diverse datasets are crucial steps toward overcoming the current limitations in depression detection from speech.

AI-generated, for reference only.

5.3 Innovations in Non-Invasive Detection

Recent advancements in non-invasive methods for depression detection have significantly improved the precision and applicability of diagnostic systems, offering valuable alternatives to traditional invasive techniques. A notable innovation is the Multi-aspect Depression Severity Assessment (MaDSA) method, which enhances depression severity evaluation across multiple dimensions, outperforming standard baseline methods by providing a more nuanced understanding of depressive symptoms [72]. This method leverages comprehensive data analysis to facilitate precise evaluations, thereby advancing the field of mental health diagnostics.

The integration of cross-task attention transfer processes represents a significant breakthrough in depression analysis. As demonstrated by Zhao et al., this method operates at both frame and sentence levels, utilizing attention mechanisms to focus on the most relevant segments of speech data [49]. This approach not only enhances the interpretability of models but also improves the accuracy of depression assessments by highlighting critical speech features indicative of depressive states.

Graph Neural Networks (GNNs) have been applied in video-based depression analysis, marking a substantial improvement over traditional single-stage methods. The two-stage temporal modeling framework using GNNs captures complex temporal dependencies in video data, providing a more detailed and context-rich analysis of depressive symptoms [73]. This integration of video analysis with conventional speech data offers a comprehensive perspective on depression, enriching the diagnostic process.

The Emotional Audio-Textual Depression Corpus (EATD-Corpus) serves as a crucial resource for future research, enabling the development of robust depression detection models that effectively utilize both audio and textual data [55]. This corpus supports the training of models capable of detecting subtle emotional cues associated with depression, thereby enhancing diagnostic accuracy.

Innovative data augmentation techniques, such as FrAUG, address the challenge of data scarcity by varying frame-width and frame-shift parameters during feature extraction. This method generates augmented data samples that retain essential acoustic characteristics, thereby improving the robustness and generalizability of depression detection models [69]. These techniques ensure that models maintain critical features necessary for effective depression detection across diverse datasets.

The application of bidirectional learning and self-attention mechanisms in Bidirectional Representation Learning Transformers (BRLTM) has addressed data heterogeneity and enhanced model interpretability, outperforming existing approaches [74]. These mechanisms enable models to capture complex relationships within data, providing a more comprehensive analysis of depressive states.

Furthermore, the integration of multimodal data at the word level, as proposed by Rohanian et al., allows for a more nuanced understanding of depression indicators [9]. This approach effectively combines audio, visual, and textual data to enhance the detection of depression symptoms, demonstrating the potential of advanced deep learning techniques in non-invasive depression diagnostics.

Overall, these innovations highlight the potential of deep learning and multimodal data integration in enhancing the accuracy and applicability of non-invasive depression detection methods. The ongoing enhancement of automated depression detection methodologies, such as those utilizing audio and text sequence modeling, chain-of-thought prompting, and multi-modal analysis of spoken language and facial expressions, holds significant potential for revolutionizing mental health screening and intervention strategies. These advancements aim to create more accessible and effective tools for identifying depression, addressing barriers to care like social stigma and financial constraints, and ultimately facilitating timely intervention for individuals who may otherwise remain undiagnosed. [19, 75, 22, 34, 17]

6 Datasets and Evaluation Metrics

The selection of suitable datasets is fundamental to advancing emotion and depression detection research, providing the empirical foundation for model development and evaluation. This section examines key datasets and the metrics used to evaluate model performance.

6.1 Overview of Commonly Used Datasets

Numerous datasets significantly enhance emotion and depression detection from speech. The DAIC-WOZ dataset, with clinically annotated audio and video from human-agent interactions, is crucial for diagnosing psychological distress and estimating depression severity [70, 26]. Complementing this, the AVEC series, including AVEC 2017, offers a comprehensive collection of multimodal recordings with PHQ-8 scores [2].

E-DAIC-WOZ and CMDC datasets provide multimodal data for depression diagnosis, integrating audio, video, and text [1]. The EATD-Corpus, the first public depression dataset in Chinese, adds linguistic diversity to the field [15].

Social media datasets, derived from tweets with depression-related hashtags and standard essay corpora, offer insights into depression-related language on these platforms [18]. Additionally, a dataset from ung.no, a Norwegian public information channel, reflects cultural nuances relevant to depression detection [5]. A dataset featuring 961 vlogs from 816 speakers provides extensive emotional expression data for model training and evaluation [76].

These datasets collectively enhance emotion and depression detection research, facilitating the development of models that address mental health diagnostic complexities. Integrating audio and text features in automated algorithms underscores the potential for passive monitoring and sentiment analysis to improve early depression detection, overcoming traditional diagnostic barriers [17, 30].

6.2 Specialized Datasets for Depression Detection

Specialized datasets are critical for advancing diagnostic models in depression detection. A notable dataset includes over 5 million posts from users with depression and approximately 1,075,000 posts from control users, providing insights into depression and offensive language [21].

The DAIC-WOZ dataset, integral to the AVEC series, facilitates depression level identification through its rich corpus of audio, video, and text from interviews lasting 7 to 33 minutes, supporting advanced machine learning model development for psychological distress assessment [6, 77].

The EATD-Corpus, as the first public dataset with audio and text data in Chinese, enhances research diversity and model generalizability across languages [15].

These specialized datasets enable robust depression detection systems, reflecting depressive disorder complexities. By employing methodologies such as chain-of-thought prompting and large language models, researchers can enhance diagnostic model precision, fostering accurate clinical evaluations and targeted interventions [22, 21, 36].

6.3 Datasets for Emotion Recognition

Emotion recognition datasets are vital for developing models that accurately identify emotional states from speech. The IEMOCAP database, with around 12 hours of audio-visual data from scripted and improvised scenarios, is a rich resource for emotion recognition research [57].

The RAVDESS features recordings from 24 professional actors, providing valuable data for differentiating subtle emotional nuances in vocal and musical contexts [58].

In multilingual contexts, the MELD includes conversations from "Friends," annotated with emotion labels, supporting model generalization across languages and cultures [55]. The SAVEE database, with recordings of four male actors expressing seven emotions, aids in exploring audio-visual integration in emotion recognition tasks [15].

These datasets offer a robust foundation for emotion recognition research, enabling sophisticated model development that addresses emotional expression complexities across modalities. Leveraging advanced deep learning techniques and data augmentation, researchers can enhance emotion recognition system precision for real-world applications, such as healthcare surveillance, where timely emotional state detection can facilitate early intervention for conditions like depression and stress [59, 67, 60].

Benchmark	Size	Domain	Task Format	Metric
SMCD[18]	1,000	Mental Health	Keyword Identification	Correlation Coefficient, Precision
OLID-MH[21]	1,075,000	Mental Health	Offensive Language Detection	F1 score, Precision
DEM3C[78]	365	Neuropsychology	3-class Classification	UAR
MMPsy[79]	12,024	Psychology	Mental Health Assessment	Accuracy, F1
DepDement[80]	2,448	Speech Analysis	Clustering	Silhouette score, Elbow method
LLM-MH[33]	3,503	Psychology	Multi-label Classification	Accuracy, F1-score
DEAR[59]	1,000	Speech Emotion Recognition	Emotion Classification	Accuracy, F1-score
MHD-Bench[38]	10,120	Mental Health	Speech Analysis	F1-score

Table 1: This table provides an overview of various benchmarks used in the evaluation of deep learning models for emotion recognition and depression detection. It details the size, domain, task format, and evaluation metrics for each benchmark, highlighting the diversity and specificity of datasets employed in this research area.

6.4 Evaluation Metrics and Datasets

Evaluating deep learning models in emotion recognition and depression detection requires comprehensive metrics to ensure reliability across datasets. Key metrics include accuracy, precision, recall, and F1-score, providing insights into model sensitivity and specificity. The F1-score is crucial for mental health detection tasks [6]. The UAR is also used for balanced evaluation across classes, especially in imbalanced datasets [4].

For regression tasks, metrics like RMSE and MAE assess prediction accuracy, particularly in studies using multimodal data to predict depression severity scores from the PHQ-8 questionnaire. Recent research showed visual features from facial landmarks achieved an MAE of 4.66, while audio-derived features yielded an MAE of 4.11, highlighting multimodal approach effectiveness [26, 2, 68, 19].

Advanced metrics like AUC measure models' discriminative power in classification tasks [3]. The PAM captures overall model performance across multiple dimensions [24].

Cross-validation techniques, including five-fold cross-validation, ensure model robustness and generalizability [3]. This evaluation framework allows comprehensive model effectiveness assessments across data splits and architectures. Grid search methods further optimize model parameters and enhance performance evaluations [6].

In social media-based models, comparing keyword frequency against a standard corpus using statistical measures like the Correlation Coefficient and Precision ensures accuracy across datasets [18].

This multifaceted evaluation approach in emotion recognition and depression detection ensures model reliability across datasets and applications. Table 1 presents a comprehensive summary of benchmarks utilized in the assessment of deep learning models for tasks related to emotion recognition and mental health detection. This rigorous process is vital for developing advanced systems capable of accurately identifying emotional states and diagnosing depression through speech analysis, ultimately enhancing mental health care accessibility and effectiveness. By leveraging sophisticated algorithms and multimodal data, these systems can provide automated assessments, improving early intervention and treatment outcomes for individuals with depression [15, 17, 35, 55, 44].

7 Challenges and Future Directions

7.1 Challenges in Current Methods

Current methods in emotion recognition and depression detection from speech face significant hurdles. One major issue is the reliance on structured clinical interviews and questionnaires, which often miss the nuanced expressions of depression found in informal speech and text [5]. This dependency can lead to insensitivity in detecting subtle emotional cues crucial for accurate diagnosis. The quality and completeness of multimodal data also impact model performance. The Multimodal Fusion Framework (MMFF) is particularly sensitive to input data quality, where inconsistencies can reduce efficacy [1]. Similarly, the SadTime network's performance is affected by variations in sample size and feature dimensions across datasets, highlighting the need for robust data handling techniques [24].

Another challenge is speaker-dependent variability, which affects model generalizability across individuals. This is particularly problematic in depression detection, where distinguishing between similar emotional states is necessary [4]. The dynamic nature of emotional expressions, especially in text, complicates detection as current methods struggle to adapt to fluctuations [27]. Social media platforms add complexity due to the informal and evolving nature of language, with current benchmarks failing to account for these variations, complicating accurate mental health issue detection [18]. Limited data samples can hinder the generalization of methods like topic modeling-based multimodal analysis [6].

Innovative approaches such as ABAFnet effectively integrate features and dynamically adjust their importance, improving depression detection accuracy [3]. However, reliance on advanced data augmentation strategies, like FrAUG, underscores the ongoing need for high-quality data to enhance classification performance [69]. Techniques like DepressionNet, which filter irrelevant content and emphasize salient features, reveal the limitations of current methodologies in handling diverse datasets [81].

Addressing these challenges requires developing adaptable and generalizable models, implementing advanced data augmentation techniques, and enhancing feature extraction methods. Hybrid networks combining self-attention mechanisms and deep convolutional neural networks could improve the extraction of depression-relevant features from speech signals. Integrating large language models and encoder-based approaches has shown potential in enhancing classification accuracy, particularly in diverse and noisy datasets, thus addressing current methodologies' limitations in mental health assessment [36, 35].

7.2 Data Scarcity and Quality

Data scarcity and quality critically challenge emotion recognition and depression detection from speech, affecting diagnostic models' reliability and generalizability. Limited dataset sizes and diversity often hinder robust model development, as small sample sizes and inadequate physiological data impede effective diagnostic tool creation. This scarcity is compounded by variability in personal attributes and speech patterns, necessitating datasets that capture a broad range of emotional expressions and behaviors [82].

Reliance on specific data sources, such as social media, often fails to represent linguistic diversity across contexts and groups, limiting model applicability [30]. Dependency on high-quality audio recordings and variability in individual speech patterns further complicate reliable model development [82]. Existing methodologies may struggle with datasets lacking sufficient affective annotations or where language does not align with employed lexica, affecting emotion recognition effectiveness [83]. The quality and availability of training data are crucial for model performance, as models depend on the diversity and representativeness of input data.

Future research should enhance dataset quality through privacy computing techniques that align social network data with professional insights, addressing data scarcity and quality issues in emotion and depression detection [30]. Developing gender-specific models for mood prediction can also improve emotion recognition capabilities, providing more tailored assessments [39]. Integrating multiple data modalities, as demonstrated by Stepanov et al., can improve accuracy in predicting depression severity compared to unimodal methods [2]. This approach emphasizes the importance of combining diverse data sources to enhance model robustness and generalizability. Additionally, synthetic data generation and augmentation techniques present promising avenues for improving dataset balance and preserving privacy, ultimately enhancing the reliability of emotion and depression detection systems.

Advancing emotional state detection is crucial for mental health diagnostics, facilitating the development of sophisticated models that accurately identify conditions like depression and anxiety. Recent studies indicate that large language models (LLMs) outperform traditional machine learning methods, especially with noisy and diverse datasets. Innovative approaches utilizing audio and text sequence modeling, along with advanced topic modeling techniques, have demonstrated comparable effectiveness to established methods, enhancing emotional state identification accuracy and promoting passive monitoring solutions that mitigate barriers to mental health care access [6, 34, 17, 36].

7.3 Privacy and Ethical Concerns

The use of speech data for emotion recognition and depression detection raises significant privacy and ethical concerns, necessitating robust measures to protect user data. Continuous audio recording, often coupled with visual data, presents ethical challenges regarding unauthorized use and potential exposure of sensitive personal information [84]. Ensuring individual privacy is paramount, as ethical considerations surrounding AI in mental health diagnostics highlight the need to address privacy concerns related to speech data usage [85].

Centralized approaches for depression assessment can exacerbate privacy issues, emphasizing the need for privacy-preserving methods that protect user data while enabling effective diagnostics [86]. Moreover, reliance on the quality and availability of multimodal data introduces additional ethical considerations, underscoring the importance of safeguarding data privacy in system development [7].

Innovative strategies, such as deploying open-source models locally through Federated Learning, can significantly enhance privacy protections by minimizing data sharing and centralization. This decentralized, privacy-preserving analysis is particularly beneficial in sensitive applications like mental health assessment. Federated Learning enables real-time, continuous evaluation of conditions such as depression while maintaining user privacy, as demonstrated in speech analysis and social media data applications. By reducing data transfer and processing on central servers, these models protect individual privacy while maintaining robust performance with minimal accuracy loss compared to traditional centralized methods [23, 86, 30]. However, the ethical implications of using AI models, particularly concerning biases and the potential for misinterpretation of emotional states, must be addressed to ensure fairness and accuracy in diagnostics. Future research should continue exploring privacy-preserving techniques and ethical frameworks that safeguard user data, ensuring that advancements in emotion recognition and depression detection do not compromise individual rights and privacy.

7.4 Model Generalizability and Robustness

Achieving model generalizability and robustness in emotion recognition and depression detection from speech remains a significant challenge, primarily due to variability in speech data across individuals and contexts. A key issue is dependency on speaker identity, which can lead to models overly tailored to specific individuals, limiting broader applicability. The Non-Uniform Speaker Disentanglement (NUSD) approach addresses this by enhancing depression detection accuracy while reducing reliance on speaker identity, thus improving patient privacy and model generalizability [87].

Integrating diverse datasets is crucial for developing models that generalize across populations and conditions. Future research should explore using kinemes for continuous prediction of depression severity, enhancing model generalizability across datasets. Combining kinemes with other behavioral markers represents a promising avenue for improving the robustness of depression detection systems [88].

Despite advances in multimodal data integration, models often struggle to generalize to unseen data, particularly in visual feature extraction. This limitation underscores the need for methodologies that effectively capture and generalize complex patterns across various data modalities [2]. Enhancing model robustness requires continuous refinement of feature extraction techniques and the development of comprehensive datasets reflecting the diversity of human emotional expressions.

To address challenges in detecting depression through social media analysis, future research should prioritize developing adaptive learning frameworks that dynamically adjust to evolving data and contextual variations, enhancing accuracy and relevance in real-time mental health assessments. This focus is crucial given the complexities of natural language use on social media, which often includes non-standard and noisy language, necessitating models that balance classification performance with computational efficiency [89, 18, 90, 45]. Additionally, exploring transfer learning techniques can facilitate model application to new domains with limited labeled data, enhancing generalizability and robustness. These efforts are essential for advancing the field and ensuring that emotion recognition and depression detection systems are effective and reliable across diverse applications.

AI-generated, for reference only.

7.5 Future Directions and Impact

The future trajectory of emotion recognition and depression detection from speech is set for transformative advancements, driven by key research directions. A primary focus is expanding and diversifying datasets, crucial for enhancing model generalization and robustness across diverse populations and contexts. Collaborative efforts to amass larger datasets and improve model generalization are essential, as is exploring integrating EEG diagnostics into clinical practice [91]. Expanding the corpus size and validating it against clinical standards, particularly across social media platforms, will enhance detection methods and applicability [18].

Advancements in feature extraction methods and exploration of novel fusion techniques for combining audio and visual data are expected to propel the field further. Future research should refine feature extraction techniques and explore multimodal approaches that integrate remote Photoplethysmography (rPPG) with other data types, enhancing the robustness of depression detection models [31]. Additionally, refining classification models and exploring cross-platform linguistic features will contribute to more nuanced emotion recognition systems. Integrating multimodal data sources, such as textual information from speech, holds promise for a deeper understanding of depression severity and its indicators [4].

The incorporation of complementary AI strategies, such as natural language processing and smart technologies, offers potential for passive monitoring of depression symptoms, providing a more nuanced understanding of mental health conditions. Future research could explore hybrid models incorporating additional data sources and larger datasets to enhance the robustness and applicability of the depression detection framework [27]. The application of FrAUG to features like voice quality, combined with other augmentation techniques, could benefit various paralinguistic applications, further enhancing detection capabilities [69].

Enhancing model interpretability and integrating additional emotional and contextual data will bolster model robustness. Applying explainability techniques and enhancing transformer models with external knowledge will improve diagnostic system transparency and reliability. Furthermore, exploring additional data modalities and refining knowledge infusion techniques will enhance the detection of less correlated symptoms, improving overall model performance [6].

Future research should also explore the applicability of these methods on diverse datasets and integrate them into mental health applications, such as intelligent chatbots for real-time support [5]. Leveraging intelligent systems can achieve real-time monitoring and intervention, offering timely support to individuals experiencing mental health challenges.

8 Conclusion

The exploration of deep learning methodologies has underscored their pivotal role in enhancing emotion recognition and depression detection from speech. The integration of diverse data modalities, including audio, text, and visual inputs, is essential for developing robust diagnostic models. A notable advancement is the implementation of hybrid networks that integrate self-attention mechanisms with deep convolutional architectures, which have surpassed existing benchmarks in depression severity detection. This highlights the potential of leveraging multimodal data for early and accurate diagnosis of depressive disorders.

Multimodal machine learning approaches have emerged as a promising avenue for improving mental health diagnostics and treatment strategies. By effectively managing data organization and addressing scarcity, these approaches have set new performance standards for depression and anxiety detection across various datasets, particularly emphasizing the utility of large language models and clinically validated data. The incorporation of dialogue structure and emotional context further refines detection accuracy, illustrating the importance of contextual understanding in model performance.

Advancements in ensemble methodologies have significantly enhanced the precision of depression detection through speech analysis, achieving superior performance metrics. Techniques such as adversarial disentanglement have improved detection accuracy while preserving speaker privacy, underscoring the need for ongoing research to tackle ethical considerations and establish standardized evaluation frameworks. The application of Bayesian Networks as predictive tools for Major Depressive Disorder exemplifies the potential of sophisticated statistical models to support clinical decision-making across diverse mental health conditions.

AI-generated, for reference only.

This survey calls for the continued exploration of innovative methodologies and technologies, which are poised to transform the landscape of mental health diagnostics. Progress in emotion recognition and depression detection systems promises to enhance diagnostic accuracy and treatment outcomes, contributing to more effective and accessible mental health care solutions.

References

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

[2] 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.

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

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

[5] Md Zia Uddin, Kim Kristoffer Dysthe, Asbjørn Følstad, and Petter Bae Brandtzaeg. Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*, 34(1):721–744, 2022.

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

[7] 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.

[8] Sultan Ahmed, Salman Rakin, Mohammad Washeef Ibn Waliur, Nuzhat Binte Islam, Billal Hossain, and Md. Mostofa Akbar. Depression detection from social media bangla text using recurrent neural networks, 2024.

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

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

[11] 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.

[12] Sri Harsha Dumpala, Katerina Dikaios, Abraham Nunes, Frank Rudzicz, Rudolf Uher, and Sageev Oore. Self-supervised embeddings for detecting individual symptoms of depression, 2024.

[13] 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.

[14] 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.

[15] 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.

[16] 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

AI-generated, for reference only.

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.

[17] 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.

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

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

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

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

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

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

[24] 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.

[25] 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.

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

[27] 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.

[28] Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, Wei Dang, and Xiaoying Pan. Deep learning for depression recognition with audiovisual cues: A review, 2021.

[29] Damian Campo, Manuela Bastidas, and Olga Lucía Quintero. Multiresolution analysis (discrete wavelet transform) through daubechies family for emotion recognition in speech, 2019.

[30] 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.

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

[32] Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. Exploring hybrid and ensemble models for multiclass prediction of mental health status on social media, 2022.

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

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

[35] 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.

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

[37] Susmita Bhaduri, Anirban Bhaduri, and Rajib Sarkar. Language independent speech emotion and non-invasive early detection of neurocognitive disorder, 2021.

[38] 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.

[39] Zakaria Aldeneh, Mimansa Jaiswal, Michael Picheny, Melvin McInnis, and Emily Mower Provost. Identifying mood episodes using dialogue features from clinical interviews, 2022.

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

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

[42] 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.

[43] 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, 2020.

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

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

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

[47] Shuanglin Li, Siyang Song, Rajesh Nair, and Syed Mohsen Naqvi. A frequency-aware augmentation network for mental disorders assessment from audio, 2025.

[48] Karol Chlasta, Krzysztof Wołk, and Izabela Krejtz. Automated speech-based screening of depression using deep convolutional neural networks, 2019.

[49] Ziping Zhao, Zhongtian Bao, Zixing Zhang, Nicholas Cummins, Haishuai Wang, and Björn Schuller. Hierarchical attention transfer networks for depression assessment from speech. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7159–7163. IEEE, 2020.

[50] Nawshad Farruque, Chenyang Huang, Osmar Zaiane, and Randy Goebel. Basic and depression specific emotion identification in tweets: Multi-label classification experiments, 2021.

[51] 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.

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

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

[54] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, pages 81–88, 2019.

[55] Ying Shen, Huiyu Yang, and Lin Lin. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model, 2022.

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

[57] Bagus Tris Atmaja and Masato Akagi. Deep multilayer perceptrons for dimensional speech emotion recognition, 2020.

[58] Xiaoran Yang, Shuhan Yu, and Wenxi Xu. Improvement and implementation of a speech emotion recognition model based on dual-layer lstm, 2024.

[59] Ravi Shankar, Abdouh Harouna Kenfack, Arjun Somayazulu, and Archana Venkataraman. A comparative study of data augmentation techniques for deep learning based emotion recognition, 2022.

[60] 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.

[61] Subham Banga, Ujjwal Upadhyay, Piyush Agarwal, Aniket Sharma, and Prerana Mukherjee. Indian emospeech command dataset: A dataset for emotion based speech recognition in the wild, 2019.

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

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

[64] 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.

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

[66] Aditya Parikh, Misha Sadeghi, and Bjorn Eskofier. Exploring facial biomarkers for depression through temporal analysis of action units, 2024.

[67] Marwan Dhuheir, Abdullatif Albaseer, Emna Baccour, Aiman Erbad, Mohamed Abdallah, and Mounir Hamdi. Emotion recognition for healthcare surveillance systems using neural networks: A survey, 2021.

[68] 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.

[69] 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.

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

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

[72] Chaebin Lee, Seungyeon Seo, Heejin Do, and Gary Geunbae Lee. Multi-aspect depression severity assessment via inductive dialogue system, 2024.

[73] Jiaqi Xu, Siyang Song, Keerthy Kusumam, Hatice Gunes, and Michel Valstar. Two-stage temporal modelling framework for video-based depression recognition using graph representation, 2021.

[74] Yiwen Meng, William Speier, Michael K. Ong, and Corey W. Arnold. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression, 2021.

[75] Albert Haque, Michelle Guo, Adam S Miner, and Li Fei-Fei. Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*, 2018.

[76] 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.

[77] Shubham Dham, Anirudh Sharma, and Abhinav Dhall. Depression scale recognition from audio, visual and text analysis, 2017.

[78] 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.

[79] 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.

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

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

[82] 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.

[83] D. Xezonaki, G. Paraskevopoulos, A. Potamianos, and S. Narayanan. Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews, 2020.

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

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

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

[87] Jinhan Wang, Vijay Ravi, and Abeer Alwan. Non-uniform speaker disentanglement for depression detection from raw speech signals, 2023.

[88] Monika Gahalawat, Raul Fernandez Rojas, Tanaya Guha, Ramanathan Subramanian, and Roland Goecke. Explainable depression detection via head motion patterns, 2023.

[89] Eduardo Garcia, Juliana Gomes, Adalberto Barbosa Júnior, Cardeque Borges, and Nádia da Silva. Deeplearningbrasil@lt-edi-2023: Exploring deep learning techniques for detecting depression in social media text, 2023.

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

[91] Milena Cukic Radenkovic. Machine learning approaches in detecting the depression from resting-state electroencephalogram (eeg): A review study, 2019.

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.