Multimodal Machine Learning in Mental Health: A Survey of Data, Algorithms, and Challenges

Zahraa Al Sahili[1]*

Queen Mary University of London, z.alsahili@qmul.ac.uk

Ioannis Patras

Queen Mary University of London, i.patras@qmul.ac.uk

Matthew Purver

Queen Mary University of London, Jožef Stefan Institute, m.purver@qmul.ac.uk

The application of machine learning (ML) in detecting, diagnosing, and treating mental health disorders is garnering increasing attention. Traditionally, research has focused on single modalities, such as text from clinical notes, audio from speech samples, or video of interaction patterns. Recently, multimodal ML, which combines information from multiple modalities, has demonstrated significant promise in offering novel insights into human behavior patterns and recognizing mental health symptoms and risk factors. Despite its potential, multimodal ML in mental health remains an emerging field, facing several complex challenges before practical applications can be effectively developed. This survey provides a comprehensive overview of the data availability and current state-of-the-art multimodal ML applications for mental health. It discusses key challenges that must be addressed to advance the field. The insights from this survey aim to deepen the understanding of the potential and limitations of multimodal ML in mental health, guiding future research and development in this evolving domain.

CCS CONCEPTS • Artificial intelligence • Machine learning •Applied computing •Life and medical sciences

Additional Keywords: multimodal machine learning, mental health, healthcare

1 INTRODUCTION

Mental health disorders represent a significant global health concern, with approximately 1 in 8 people worldwide affected by conditions such as anxiety and depression, accounting for an estimated 970 million individuals in 2019 alone [1]. Despite the prevalence of these disorders, there is a substantial treatment gap, particularly in low- and middle-income countries where up to 85% of individuals with mental health conditions receive no treatment [2]. This gap underscores the urgent need for innovative solutions to enhance mental health care access and efficacy.

In recent years, digital health interventions have emerged as a promising avenue to address this need. The global digital mental health market, valued at $5.1 billion in 2020, is projected to expand at a compound annual growth rate of 21.0% from 2021 to 2028 [3]. Among these interventions, multimodal machine learning—an approach that integrates diverse data types such as text, audio, and video—has shown considerable promise. Research indicates that combining indicates t multiple data types (e.g., audio, video, text) leads to better accuracy and robustness in detecting and assessing mental health conditions compared to using a single data type[4]. For example, integrating facial expressions, speech, and physiological signals can enhance the detection of emotions

---

[1]* corresponding author.

and mental states, offering more comprehensive insights for interventions speech and facial recognition data can significantly improve the accuracy of detecting mental health conditions, such as depression, compared to single data modality approaches [4].

This survey provides a comprehensive overview of multimodal machine learning in mental health research. We review the various types of multimodal data commonly used, including text, audio, video, and physiological data. We present an overview of available datasets and the specific mental health conditions they address, such as depression, stress, bipolar disorder, and PTSD. Additionally, we discuss state-of-the-art machine learning techniques employed to analyze these data, encompassing RNN/CNN-based, transformer-based, and graph neural network-based algorithms.

Furthermore, we explore the challenges and opportunities in multimodal machine learning for mental health, including issues related to data availability, privacy, bias, benchmarking, complexity and evaluation. Despite these challenges, the potential benefits of multimodal machine learning for mental health research are vast, promising significant improvements in diagnosis and treatment, ultimately enhancing the lives of millions.

2 DATA TYPES AND DATASETS

2.1 Data Types

In multimodal machine learning for mental health applications, a variety of data types are commonly used, including text, audio, video, and physiological data.

Text Data: such as social media posts, clinical notes, and survey responses, can provide insight into an individual's thoughts and emotions. Text data can be analyzed using natural language processing (NLP) techniques, such as sentiment analysis and emotion detection, to extract information about an individual's mental state.

Audio Data: such as speech and nonverbal cues, can provide information on an individual's tone of voice and emotional state. Audio data can be analyzed using techniques such as speech recognition, emotion recognition, and speaker identification to extract information about an individual's mental state.

Video Data: such as facial expressions and body language, can provide information on an individual's nonverbal cues and emotional state. Video data can be analyzed using techniques such as facial expression recognition, body language analysis, and gaze tracking to extract information about an individual's mental state.

Physiological Data: such as heart rate, skin conductance, and electroencephalography (EEG), can provide objective measures of an individual's physiological responses to different stimuli. Physiological data can be analyzed using techniques such as signal processing and pattern recognition to extract information about an individual's mental state.

To analyze these data, it is important to have large, diverse, and representative datasets. However, collecting such data can be challenging becuase of patients privacy restrictions.

2.2 Datasets

Multimodal mental health datasets have become increasingly valuable for researchers aiming to investigate the underlying mechanisms and treatment options for various mental health disorders. These datasets comprise a range of modalities, such as video, audio, text, and physiological signals, offering a comprehensive understanding of the complex nature of mental health disorders. In this section, we present an overview of the available multimodal mental health datasets that are currently available [5-30], highlighting the types of data included in each dataset as well as their applications as shown Table 1.

The mental health disorders studied include depression, PTSD, bipolar disorder, stress, and emotion recognition (Table 1). However, certain mental illnesses, such as schizophrenia, were not included due to the sensitivity of the data and the complexity of the disorders. Additionally, we provided information on the gender balance and the countries where the data was collected, to aid in mitigating bias and promoting fairness in research.
It is also important to note that the datasets included in this survey have access restrictions due to privacy concerns and ethical considerations. It is imperative that any research conducted using these datasets is done under strict ethical guidelines and to improve mental health outcomes for all individuals.

The datasets cover a variety of mental health disorders, with a notable focus on depression and stress, which is the most commonly targeted disease. Specifically, seven datasets address depression, three datasets focus on PTSD, ten on stress, two on bipolar disorder, one on behavioral disorders, and six on emotion recognition. This prevalence of depression and stress datasets reflects its high impact on global mental health.

In terms of data modalities, the majority of datasets incorporate video and audio data, with 16 datasets each. Text data is included in 11 datasets, and physiological data is found in 13 datasets. The integration of these multiple modalities enhances the robustness and accuracy of machine learning models used in mental health research. However, the combination of physiological data with other modalities is less common.

The integration of multiple modalities is common, which strengthens the reliability of machine learning models. Nonetheless, the use of physiological data is less prevalent despite its potential benefits. When examining multimodal combinations, six datasets include video, audio, and text; another six combine video and audio. Four datasets focus solely on physiological data, while two datasets each combine text and physiology, video and physiology, audio and text and physiology, and video and text and physiology. Additionally, one dataset each combines audio and physiology, video, audio, and physiology, and video, text, and physiology.Moreover, the datasets vary widely in size and composition, ranging from large-scale datasets like E-DAIC to smaller, more specific ones such as the Chinese Multimodal Depression Corpus.

The analysis of these 24 datasets highlights the critical role of multimodal data in advancing mental health research through machine learning. While significant progress has been made, there remains a need for more comprehensive and diverse datasets to fully realize the potential of these technologies. By providing a comprehensive understanding of the complex nature of mental health disorders, these datasets have the potential to drive important advancements in the field and improve outcomes for individuals struggling with mental health issues.To enhance the value of future datasets, it is crucial to include a broader range of demographic information, ensuring that models can generalize well across different populations. In addition, ensuring that datasets are large and diverse could improve the generalizability of models. Moreover, expanding the inclusion of physiological data can provide additional insights, potentially leading to more accurate mental health assessments.

Table 1: Disorders, Modalities, and Details in Mental Health Multimodal Datasets

reference	Dataset	disease	# modalties	video	audio	text	Pysiology	# Pysiology	more info	Gender	country
[5]	DAIC-WOZ	depression PTSD	3	x	x	x			189 sessions	-	-
[6]	E-DAIC dataset	depression PTSD	3	x	x	x			275 subjects -70 hours	-	-
[7]	AVEC 2013	depression	2	x	x				340 videos	-	-
[8]	A multimodal	depression	2		x		x	1	52-55 particip	-	China

	open dataset for mental disorder analysis								ants per state		
[9]	SWELL	stress	5				x	5	25 participants	32% F	Netherlands
[10]	D-vlog	depression	2	x	x				-	-	YouTube
[11]	Chinese Multimodal Depression Corpus	depression	2	x	x				78 folders for each subject. 19 MDD and 26 HC	-	China
[12]	Bipolar Disorder and Cross-Cultural Affect Recognition	bipolar/ behavior	3	x	x		x	1	BPC +SEWA + RECOLA databases	-	Turkey +Germany+Hungry+ France
[13]	Turkish Bipolar Disorder Corpus	bipolar	2	x	x				51 participants	31% F	Turkey
[14]	PTSD in the Wil	PTSD	3	x	x	x			634 participants	-	-
[15]	BIRAFFE 2	emotion	2			x	x	2	103 participants	33% F	Poland
[Koelstr,2012]	Deap	emotion	3	x			x	2	32 participants		
[17]	MuSE	stress	5	x		x	x	3	28	32% F	US
[18]	MIREX	emotion	3	x	x	x			193 samples	-	-
[19]	MELD	emotion	3	x	x	x			13,000 utterances	-	-
[20]	WEMAC	emotion	5		x	x	x	3	47 women	100% F	Spain
[21]	EmoReact:	emotion	2						63 children	51% F	YouTube
[22]	WESAD	stress and affect	3				x	3	25 students	-	-

[23]	A multimodal sensor dataset for continuous stress detection of nurses in a hospital	stress	6				x	6	15 participants	-	-
[24]	MuSe-Stress	stress	2	x	x				105	69.9% F	-
[25]	Multimodal stress detection	stress	6		x	x	x	4	80 participants	58.7% F	-
[26]	Empathic School	stress	4	x			x	3	20 participants	-	Finland & US
[27]	Automatic Detection of Self-Adaptors for Psychological Distress	stress	3	x	x	x			35 participants	-	-
[28]	CLAS	stress	5			x	x	4	62 healthy	-	-
[29]	MuSe 2022	stress	2	x	x				105 participants	-	-
[30]	UBFC-PHYS	stress	2				x	2	56 participants	-	-

3 ALGORITHMS

A variety of machine learning algorithms have been applied to multimodal data for mental health applications.Historically, Support Vector Machines (SVMs) and Random Forests have been used to classify multimodal data, such as text and physiological data. In addition, SVMs were particularly useful for high-dimensional data and have been applied to classify mental health conditions, such as depression and PTSD. Moreover, Random Forest algorithm was used for a wide range of multimodal mental health data as classifying depression and anxiety through aggregating multiple decision trees to increase the accuracy of the classifier. Recent research has focused on advanced deep learning techniques, particularly CNN/RNN-based models, Transformer models, and Graph Convolutional Networks (GCNs).

3.1   CNN/RNN-Based Models

Multimodal algorithms leveraging Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown significant promise in integrating information from various modalities—such as text, image, audio, and video—for diverse tasks. CNNs are adept at extracting high-level visual features from image and video data, while RNNs, particularly Long Short-Term Memory (LSTM) networks, are proficient in modeling temporal dependencies within sequential data like text and speech. In these multimodal frameworks, CNNs process visual inputs, extracting salient features, and RNNs handle sequential data, capturing contextual dependencies. Outputs from both networks are typically combined in a fusion layer to make final predictions, improving the performance over traditional methods.

For instance, Man et al. demonstrated that a multimodal pipeline combining ResNet architectures for imaging with ALMO, FastText, and BOW for text significantly outperformed six different unimodal approaches in detecting depressive symptoms from social media data [31]. In addition, Gui et al. developed a decentralized execution approach using COMMA policy gradients, which employed GRU for text and VGG-Net for image features. Their model used policy gradient agents to decide feature selection, resulting in enhanced depression detection performance [32]. Similarly, Rohanian et al. proposed an LSTM-based word-level multimodal system that incrementally fused modalities and introduced fusion gates to control the contributions of audio and visual inputs, effectively mitigating noise [33]. To add, Zhang et al. introduced a trilateral bimodal encoding model (MEN) incorporating attentional decision fusion (ADF) and a hybrid fusion approach for depression detection [34]. Their model combined early intra-modality and late inter-modality fusion, achieving superior performance on the AVEC2013 and AVEC2014 datasets, with notable improvements in MAE and RMSE scores [34]. Moreover, Pradhan et al. utilized a hybrid network combining Complex dual Tree with Fast Lifting Wavelet Transform (CTFL-WT) for pre-processing and Channel Attentive SqueezeNet (CASN) for feature extraction to predict depression risk [35]. Their model showed improved classification performance on the WESAD dataset by using Adaptive Arithmetic Optimization Algorithm (AAOA) and a Hybrid DenseNet with LSTM (DLSTM) [35].

Das et al. proposed an audio-based depression detection method, using neural networks to classify audio spectrogram features [36]. Their model, combining MFCC and spectrogram features with a novel CNN architecture, achieved over 90% accuracy on the DAIC-WOZ and MODMA datasets, outperforming state-of-the-art methods [36]. In addition, Zhang et al. developed the Audio, Video, and Text Fusion-Three Branch Network (AVTF-TBN), which integrated auditory, visual, and textual cues, achieving high performance on depression risk analysis tasks [37]. Moreover, Lilhore et al. introduced a hybrid model for predicting Postpartum Depression Disorder (PPDD), combining Improved Bi-directional Long Short-Term Memory (IBi-LSTM) with Transfer Learning (TL) for CNN-text and CNN-audio architectures [38]. Their model demonstrated superior precision, recall, accuracy, and F1-score on a UCI dataset, highlighting its potential for early PPDD risk assessment [38]. Furthermore, Beniwal et al. explored the relationship between social media use and depression, developing a multimodal dataset comprising text, emoticons, and images [39]. Their hybrid BERT-CNN model achieved a remarkable 99% accuracy in categorizing users as depressive or non-depressive, showcasing the effectiveness of combining text and image data for mental health detection [39]. In a similar manner, Aina et al. introduced a pipeline for analyzing facial expressions using AffectNet and 2013 Facial Emotion Recognition (FER) datasets [40]. Their hybrid architecture, utilizing the YOLOv8 algorithm and an ensemble of CNNs and Visual Transformer (ViT) models, achieved an 81% accuracy in predicting mental disorders. Techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) and saliency maps were integrated to enhance interpretability, aiding healthcare professionals in understanding model predictions [40].

3.2   Multimodal Transformer Models

Transformer models have significantly advanced multimodal data integration, enabling comprehensive analysis across text, image, and audio data. Yoon et al. proposed a method utilizing unimodal transformers for video and audio processing, followed by a multimodal transformer and a depression detection layer, demonstrating the approach on their Dvlog dataset [10]. Similarly, Bucur et al. presented a versatile, time-enhanced multimodal transformer architecture for identifying depression from social media posts [41]. Their model used pre-trained embeddings for images and text, integrated time2vec positional embeddings to account for time differences

between posts, and incorporated EmoBERTa and CLIP embeddings, achieving robust performance even with unstructured datasets. Additionally, Yang et al. introduced the Multimodal Purification Fusion (MMPF) model for automatic depression detection (ADD) [41]. MMPF incorporated clinical constraints into the learning process through a prior constraint gating (PCG) strategy and employed text and audio encoders to extract features. The model used a multimodal purification refinement process and multiperspective contrastive learning (MCL) strategy, achieving promising results on the DAIC-WOZ dataset [42].

The TOpic ATtentive transformer-based ADD (TOAT) model by Guo et al. addressed challenges in organizing and processing varied audio and textual data lengths, and the scarcity of training samples due to privacy concerns[43]. TOAT utilized a topic attention module and a two-branch architecture with late fusion to independently encode audio and textual data, achieving state-of-the-art performance on the DAIC-WOZ dataset [43]. In addition, Tao et al. proposed the DepMSTAT framework, designed to detect depression using acoustic and visual features from social media data [44]. This framework included modules for data preprocessing, token generation, and a Spatial-Temporal Attentional Transformer (STAT) for capturing spatial and temporal correlations. DepMSTAT demonstrated high accuracy (71.53%) on the D-Vlog depression database, outperforming existing models [44].

On the other side, Yoa et al. explored the interconnectedness of stress detection and emotion recognition using textual and acoustic features [45]. Their MUSER architecture, a transformer-based model, employed a speed-based dynamic sampling approach for multi-task learning, demonstrating effective stress detection on the MuSE dataset and external validation with the OMG-Emotion dataset [45].Moreover, Lee et al. addressed mood disorder detection by integrating brain structural MRI scans with DNA whole-exome sequencing data [46]. Their fusion model, combining Vision Transformer (ViT) and XGBoost, significantly improved the area under the curve (AUC) and accuracy, demonstrating the potential of multimodal integration for refining mood disorder diagnostics [46].Furthermore, Malhotra et al. emphasized the application of explainable AI (XAI) techniques to enhance the interpretability of Transformer-based models in mental healthcare diagnostics [47]. By employing SHAP and LIME, they conducted extensive experiments with large language models (LLMs) fine-tuned for detecting depressive and suicidal behavior from online content [47]. Their study highlighted the importance of XAI in understanding the psychological states and behaviors of individuals.Additionally, Moon et al. addressed the limitations of controlled laboratory environments by introducing the Extended D-vlog dataset, which includes real-life YouTube vlogs [48]. They developed a virtual agent to provide Cognitive Behavioral Therapy (CBT)-based responses, achieving impressive performance in real-world depression detection and intervention with the TVLT model on the Multimodal Extended D-vlog Dataset [48].

3.3 Graph Convolutional Networks (GCNs)

Graph Convolutional Networks (GCNs) have been applied to multimodal data to effectively capture complex and relational information. Zheng et al. proposed a graph attention model that integrates multimodal knowledge for depression detection [49]. Their approach, leveraging audio, video, and text modalities, demonstrated superior performance on the DIAC-WOZ dataset by learning effective embeddings and incorporating medical knowledge [49]. In addition, Xia et al. (2024) addressed the few-shot learning problem in depression detection with a multimodal GNN-based model [50]. By integrating a GNN for recursive aggregation and transformation of node representations, their approach effectively combined audio, text, and video features in a Bi-LSTM fusion network [50]. This method achieved an accuracy of 0.861 on the DAIC-WOZ dataset, significantly outperforming baseline models. Also, Cha et al. highlighted the importance of early depression detection and introduced the Multimodal Object-Oriented Graph Attention Model (MOGAM) [51]. This model, applicable to various data types including text, images, and videos, employed a cross-attention mechanism to aggregate multimodal features [51]. MOGAM achieved an accuracy of 0.871 and an F1-score of 0.888, proving its scalability and effectiveness on social media datasets [51].In addition, Shen et al. introduced a graph-based multimodal fusion approach for automatic depression assessment (ADA) [52]. Their model used undirected edges to link temporally continuous features within each modality and across different modalities, ensuring seamless global information propagation. This

approach achieved a root mean square error (RMSE) of 4.80 and a CCC value of 0.563 on the E-DAIC dataset [52].Furthermore, Noman et al. proposed a graph deep learning framework for classifying brain networks in major depressive disorder (MDD) [53]. Their graph autoencoder (GAE) architecture embedded the topological structure and node content of fMRI networks into low-dimensional representations, significantly outperforming state-of-the-art methods in brain connectome classification[53].

On the other side, Pann et al. introduced MAMF-GCN, a multi-scale adaptive multi-channel fusion deep graph convolutional network designed to predict mental disorders [54]. This model extracts both specific and shared features from imaging data, phenotypic information, and other sources, using an attention mechanism to learn adaptive weights for feature fusion, thereby enhancing diagnostic accuracy for brain disorders [54].Moreover, Hu et al. developed MMGCN, a multimodal fusion framework employing deep graph convolution networks for emotion recognition in conversations [55]. Their model utilized a modality encoder and a multimodal spectral-based GCN, achieving state-of-the-art performance on the IEMOCAP and MELD datasets [55]. Similarly, Joshi et al. proposed COGMEN, a contextualized GNN-based multimodal emotion recognition model that combines a recurrent GNN with a graph transformer network [56].In addition, Cabral et al. proposed a novel multi-label emotion graph representation for classifying mental health based on social media posts [57]. They constructed a word–document graph tensor to describe emotion-based contextual representation, demonstrating significant performance improvements across three public datasets. Moreover, Wang et al. introduced the Multimodal Transformer Graph Convolution Attention Isomorphism Network (MTGCAIN) for diagnosing intellectual disability (ID) [58]. MTGCAIN leveraged graph convolution attention to extract brain connectivity features and used multiple brain network atlases for comprehensive information exchange. This approach achieved an accuracy of 81.29% and an AUC of 0.8760 [58]. Likewise, Liu et al. developed a deep GCN based on variable multi-graph and multimodal data (VMM-DGCN) for autism spectrum disorder (ASD) diagnosis[59]. Their algorithm, evaluated on the ABIDE I dataset, achieved a 91.62% accuracy and a 95.74% AUC, outperforming other diagnostic algorithms by effectively capturing multi-scale feature representations and incorporating non-imaging information into feature representations[59].

The potential for multimodal transformers and graph neural networks in mental health research is substantial. GNNs excel in capturing complex relationships inherent in mental health conditions, while transformers are effective in pretraining and transfer learning, beneficial in data-scarce scenarios. Future research could explore graph transformers and spatio-temporal graph neural networks, as well as the integration of GNN-based fusion within multimodal transformers.

4   OPEN CHALLENGES

This section discusses the open challenges in the era of multimodal machine learning for mental health including the need to develop robust and explainable multimodal machine learning models for mental health applications that are ethically sound and socially responsible.

4.1   Data availability
One of the primary challenges in multimodal machine learning for mental health applications is the need for extensive and diverse data collection. Effective training and validation of machine learning models require access to large, representative datasets. However, collecting such data is challenging due to ethical considerations, including data privacy and informed consent, and logistical issues related to the inclusion of marginalized communities. This lack of diverse data can lead to biased models that do not accurately represent all segments of the population, thereby impacting the fairness and accuracy of mental health assessments.

4.2   Benchmarks
Establishing benchmarks in multimodal mental health research is another significant challenge. Standardized and reliable methods for evaluating the performance of various models and techniques are essential for the advancement of the field. There is a need for a common framework that accurately reflects real-world scenarios

and accounts for the complexity of mental health conditions. Effective benchmarks will facilitate the comparison of different approaches and ultimately lead to improved mental health diagnostics and treatments.

4.3 Explainability

Explainability is a critical challenge in the development of multimodal machine learning models for mental health. Mental health assessments are inherently subjective and complex, making it difficult to create models that are both robust and interpretable. For patients and healthcare providers to trust the outcomes and recommendations of these models, clear and transparent explanations of the decision-making processes are essential. Additionally, explainability can help identify and mitigate biases in the data and models, thereby enhancing the accuracy and fairness of mental health diagnoses and treatments.

4.5 Bias and Fairness

Bias and fairness are crucial considerations in the development of multimodal models for mental health. These models can perpetuate and amplify existing biases in the data and healthcare system, leading to unfair or inaccurate diagnoses and treatments for certain populations, particularly those from underrepresented groups. It is essential to identify and mitigate any biases present in the data and models to ensure that these tools are reliable and equitable for all patients. This involves developing robust methods for evaluating and addressing bias and using diverse and representative training data.

4.6 Privacy

Privacy is a major concern in the development of multimodal models for mental health, as these models typically require access to sensitive and personal information, such as medical records, mental health assessments, and audio and video recordings. Ensuring the privacy of this information is crucial to building trust with patients and healthcare providers and avoiding potential harm. This necessitates the development of secure, privacy-preserving techniques for collecting, storing, and utilizing sensitive data in the creation and deployment of multimodal algorithms.

4.7 Evaluation

Robust evaluation of multimodal machine learning models for mental health applications is an ongoing challenge. Traditional evaluation metrics, such as accuracy and precision, may not adequately capture the complexity and subjectivity of mental health assessments. Therefore, there is a need to develop new evaluation methods that consider the nuanced nature of mental health conditions and the impact of multimodal data integration.

4.8 Complexity and Multifactorial Nature

The inherent complexity and multifactorial nature of mental health disorders present significant challenges in model development. Mental health conditions are influenced by a wide range of factors, including biological, psychological, and social elements, making it difficult to create models that accurately capture these complexities. Additionally, some disorders, such as schizophrenia and bipolar disorder, require further research to understand their underlying mechanisms before machine learning techniques can be effectively applied.

Overall, addressing these challenges requires further research to develop robust, explainable, and ethically sound multimodal machine learning models for mental health applications. Ensuring that these models are socially responsible and capable of providing accurate and fair assessments across diverse populations is essential for advancing the field and improving mental health outcomes.

5 CONCLUSION

The integration of artificial intelligence (AI) and multimodal machine learning (ML) in mental health research presents significant potential for improving the identification, diagnosis, and treatment of mental health disorders. Our survey demonstrates that multimodal ML can enhance the accuracy and robustness of mental health

assessments by integrating diverse data sources such as video, audio, text, and physiological signals. This comprehensive approach offers a more complete understanding of an individual's mental health, leading to more effective treatments.

However, several challenges need to be addressed to fully leverage the benefits of multimodal ML in mental health. These include the need for extensive and diverse data collection, ethical considerations regarding data privacy, and the importance of addressing bias and fairness to ensure equitable model outcomes,added to the development of standardized benchmarks and evaluation methods.

Despite these challenges, the potential benefits of multimodal ML in mental health research are immense. Continued research and collaboration are essential to overcome these obstacles, leading to significant advancements in mental health diagnostics and treatments. By addressing these issues, we can ultimately improve the lives of millions of individuals worldwide, paving the way for a future where mental health care is more accurate, effective, and accessible.

REFERENCES

[1] World Health Organization. (n.d.). *Mental health*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

[2] World Health Organization. (n.d.-a). *Mental health atlas 2020*. World Health Organization. https://www.who.int/publications/i/item/9789240036703

[3] *Mental health apps market size and Share report, 2030*. Mental Health Apps Market Size And Share Report, 2030. (n.d.). https://www.grandviewresearch.com/industry-analysis/mental-health-apps-market-report

[4] Ahmad, A., Singh, V., Upreti, K. (2024). A Systematic Study on Unimodal and Multimodal Human Computer Interface for Emotion Recognition. In: García Márquez, F.P., Jamil, A., Ramirez, I.S., Eken, S., Hameed, A.A. (eds) Computing, Internet of Things and Data Analytics. ICCIDA 2023. Studies in Computational Intelligence, vol 1145. Springer, Cham. https://doi.org/10.1007/978-3-031-53717-2_35

[5] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel

[6] Ringeval, Fabien, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt et al. " AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition, " in Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop , pp. 3-12. ACM, 2019.

[7] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition

challenge. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge (AVEC '13). Association for Computing Machinery, New York, NY, USA, 3–10. https://doi.org/10.1145/2512530.2512533

[8] Cai, H., Yuan, Z., Gao, Y. et al. A multi-modal open dataset for mental-disorder analysis. Sci Data 9, 178 (2022). https://doi.org/10.1038/s41597-022-01211-x

[9] Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M., & Kraaij, W. (2014). The SWELL Knowledge Work Dataset for Stress and User Modeling Research. To appear in: Proceedings of the 16th ACM International Conference on Multimodal Interaction (ICMI 2014) (Istanbul, Turkey, 12-16 November 2014)

[10] Yoon, J., Kang, C., Kim, S., & Han, J. (2022). D-vlog: Multimodal Vlog Dataset for Depression Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11), 12226-12234. https://doi.org/10.1609/aaai.v36i11.21483

[11] Bochao Zou, May 9, 2022, "Chinese Multimodal Depression Corpus", IEEE Dataport, doi: https://dx.doi.org/10.21227/jng1-rn06.

[12] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, Elvan Ciftçi, Hüseyin Güleç, Albert Ali Salah, and Maja Pantic. 2018. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop (AVEC'18). https://doi.org/10.1145/3266302.3266316

[13] Elvan C¸ ift¸ci, Heysem Kaya, H¨useyin G¨ule¸c, and Albert Ali Salah. The turkish audiovisual bipolar disorder corpus. In 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), pages 1–6. IEEE, 2018.

[14] Sawadogo, M. A. L., Pala, F., Singh, G., Selmi, I., Puteaux, P., & Othmani, A. (2022). PTSD in the Wild: A Video Database for Studying Post-Traumatic Stress Disorder Recognition in Unconstrained Environments. arXiv preprint arXiv:2209.14085.

[15] Kutt, K., Drążyk, D., Żuchowska, L. et al. BIRAFFE2, a multimodal dataset for emotion-based personalization in rich affective game environments. Sci Data 9, 274 (2022). https://doi.org/10.1038/s41597-022-01402-6

[16] S. Koelstra S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," in IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18-31, Jan.-March 2012, doi: 10.1109/T-AFFC.2011.15.

[17] Mimansa Jaiswal, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2020. MuSE: a Multimodal Dataset of Stressed Emotion. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 1499–1510, Marseille, France. European Language Resources Association.

[18] Panda R., Malheiro R., Rocha B., Oliveira A. & Paiva R. P. (2013). "Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis". 10th International Symposium on Computer Music Multidisciplinary Research -- CMMR'2013, Marseille, France.

[19] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation. ACL 2019.

[20] Miranda, J. A., Rituerto-González, E., Gutiérrez-Martín, L., Luis-Mingueza, C., Canabal, M. F., Bárcenas, A. R., ... & López-Ongil, C. (2022). WEMAC: Women and Emotion Multi-modal Affective Computing dataset. arXiv preprint arXiv:2203.00456.

[21] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E. Hughes, and Louis-Philippe Morency. 2016. EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16). ACM, NY, USA, 137–144. https://doi.org/10.1145/2993148.2993168

[22] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, Kristof Van Laerhoven, "Introducing WESAD, a multimodal dataset for Wearable Stress and Affect Detection", ICMI 2018, Boulder, USA, 2018

[23] Hosseini S, Gottumukkala R, Katragadda S, Bhupatiraju RT, Ashkar Z, Borst CW, Cochran K. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. Sci Data. 2022 Jun 1;9(1):255. doi: 10.1038/s41597-022-01361-y. PMID: 35650267; PMCID: PMC9159985.

[24] Stappen, Lukas, Baird, Alice, Christ, Lukas, Meßner, Eva-Maria, & Schuller, Björn. (2021). MuSe-Stress: Multimodal Emotional Stress (MuSe2021) (Version V1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4659752

[25] Pisanski, Katarzy et al. (2018), "Data for: Multimodal stress detection: Testing for covariation in vocal, hormonal and physiological responses to Trier Social Stress Test ", Mendeley Data, V2, doi: 10.17632/frstygc2vp.2

[26] Hosseini, M., Sohrab, F., Gottumukkala, R., Bhupatiraju, R. T., Katragadda, S., Raitoharju, J., ... & Gabbouj, M. (2022). EmpathicSchool: A multimodal dataset for real-time facial expressions and physiological data analysis under different stress conditions. *arXiv preprint arXiv:2209.13542*.

[27] W. Lin, I. Orton, M. Liu and M. Mahmoud, "Automatic Detection of Self-Adaptors for Psychological Distress," 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 2020, pp. 371-378, doi: 10.1109/FG47880.2020.00032.

[28] Markova, Valentina (2020), "CLAS: A Database for Cognitive Load, Affect and Stress Recognition", Mendeley Data, V1, doi: 10.17632/8hm59ryzb8.1

[29] Christ, L., Amiriparian, S., Baird, A., Tzirakis, P., Kathan, A., Müller, N., ... & Schuller, B. W. (2022). The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress.

[30] Rita Meziati, Yannick Benezeth, Pierre De Oliveira, Julien Chappé, Fan Yang, March 3, 2021, "UBFC-Phys", IEEE Dataport, doi: https://dx.doi.org/10.21227/5da0-7344.

[31] Mann, P., Paes, A., & Matsushima, E. H. (2019). See and read: Detecting depression symptoms in higher education students using multimodal social media data. *arXiv preprint arXiv:1912.01131*.

[32] Gui, T., Zhu, L., Zhang, Q., Peng, M., Zhou, X., Ding, K., & Chen, Z. (2019). Cooperative Multimodal Approach to Depression Detection in Twitter. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 110-117. https://doi.org/10.1609/aaai.v33i01.3301110

[33] Rohanian M, Hough J, Purver M (2019) . Detecting depression with word-level multimodal fusion . Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2019-September , 1443 - 1447

[34] Xiufeng Zhang, Bingyi Li, Guobin Qi,A novel multimodal depression diagnosis approach utilizing a new hybrid fusion method,Biomedical Signal Processing and Control,Volume 96, Part A,2024,106552,ISSN 1746-8094

[35] Pradhan, A., Srivastava, S. Hybrid densenet with long short-term memory model for multi-modal emotion recognition from physiological signals. *Multimed Tools Appl* **83**, 35221–35251 (2024). https://doi.org/10.1007/s11042-023-16933-2

[36] Das, A. K., & Naskar, R. (2024). A deep learning model for depression detection based on MFCC and CNN generated spectrogram features. *Biomedical Signal Processing and Control*, *90*, 105898. https://doi.org/10.1016/j.bspc.2023.105898

[37] Zhang Z, Zhang S, Ni D, Wei Z, Yang K, Jin S, Huang G, Liang Z, Zhang L, Li L, et al. Multimodal Sensing for Depression Risk Detection: Integrating Audio, Video, and Text Data. *Sensors*. 2024; 24(12):3714. https://doi.org/10.3390/s24123714

[38] Lilhore, U. K., Dalal, S., Varshney, N., Sharma, Y. K., Rao, K. B., Rao, V. V., Alroobaea, R., Simaiya, S., Margala, M., & Chakrabarti, P. (2024). Prevalence and risk factors analysis of postpartum depression at early stage using hybrid deep learning model. *Scientific Reports*, *14*(1), 1-24. https://doi.org/10.1038/s41598-024-54927-8

[39] Beniwal, R., & Saraswat, P. (2024). A Hybrid BERT-CNN Approach for Depression Detection on Social Media Using Multimodal Data. *The Computer Journal*, *67*(7), 2453-2472. https://doi.org/10.1093/comjnl/bxae018

[40] J. Aina, O. Akinniyi, M. M. Rahman, V. Odero-Marah and F. Khalifa, "A Hybrid Learning-Architecture for Mental Disorder Detection Using Emotion Recognition," in IEEE Access, vol. 12, pp. 91410-91425, 2024, doi: 10.1109/ACCESS.2024.3421376.

[41] Bucur, A. M., Cosma, A., Rosso, P., & Dinu, L. P. (2023). It's Just a Matter of Time: Detecting Depression with Time-Enriched Multimodal Transformers. *arXiv preprint arXiv:2301.05453*

[42] B. Yang *et al.*, "MMPF: Multimodal Purification Fusion for Automatic Depression Detection," in *IEEE Transactions on Computational Social Systems*, doi: 10.1109/TCSS.2024.3411616.

[43] Guo, Y., Zhu, C., Hao, S., & Hong, R. (2022). A Topic-Attentive Transformer-based Model For Multimodal Depression Detection. arXiv preprint arXiv:2206.13256..

[44] Y. Tao, M. Yang, H. Li, Y. Wu and B. Hu, "DepMSTAT: Multimodal Spatio-Temporal Attentional Transformer for Depression Detection," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 2956-2966, July 2024, doi: 10.1109/TKDE.2024.3350071.

[45] Yao, Y., Papakostas, M., Burzo, M., Abouelenien, M., & Mihalcea, R. (2021). Muser: Multimodal stress detection using emotion recognition as an auxiliary task. *arXiv preprint arXiv:2105.08146*.

[46] Lee, S., Cho, Y., Ji, Y., Jeon, M., Kim, A., Ham, B., & Joo, Y. Y. (2024). Multimodal integration of neuroimaging and genetic data for the diagnosis of mood disorders based on computer vision models. *Journal of Psychiatric Research*, *172*, 144-155. https://doi.org/10.1016/j.jpsychires.2024.02.036

[47] Malhotra, A., & Jindal, R. (2024). XAI Transformer based Approach for Interpreting Depressed and Suicidal User Behavior on Online Social Networks. *Cognitive Systems Research*, *84*, 101186. https://doi.org/10.1016/j.cogsys.2023.101186

[48] Moon, P., & Bhattacharyya, P. (2024). We Care: Multimodal Depression Detection and Knowledge Infused Mental Health Therapeutic Response Generation. *arXiv preprint arXiv:2406.10561*.

[49] W. Zheng, L. Yan, C. Gou and F. -Y. Wang, "Graph Attention Model Embedded With Multi-Modal Knowledge For Depression Detection," *2020 IEEE International Conference on Multimedia and Expo (ICME)*, London, UK, 2020, pp. 1-6, doi: 10.1109/ICME46284.2020.9102872.

[50] Xia, Y., Liu, L., Dong, T. *et al.* A depression detection model based on multimodal graph neural network. *Multimed Tools Appl* **83**, 63379–63395 (2024). https://doi.org/10.1007/s11042-023-18079-7

[51] Cha, J., Kim, S., Kim, D., & Park, E. (2024). MOGAM: A Multimodal Object-oriented Graph Attention Model for Depression Detection. *arXiv preprint arXiv:2403.15485*.

[52] Shen, H., Song, S., & Gunes, H. (2024). *Multi-modal Human Behaviour Graph Representation Learning for Automatic Depression Assessment*. Apollo - University of Cambridge Repository. https://doi.org/10.17863/CAM.108413

[53] F. Noman, C. -M. Ting, H. Kang, R. C. . -W. Phan and H. Ombao, "Graph Autoencoders for Embedding Learning in Brain Networks and Major Depressive Disorder Identification," in *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1644-1655, March 2024, doi: 10.1109/JBHI.2024.3351177.

[54] Pan J, Lin H, Dong Y, Wang Y, Ji Y. MAMF-GCN: Multi-scale adaptive multi-channel fusion deep graph convolutional network for predicting mental disorder. Comput Biol Med. 2022 Sep;148:105823. doi: 10.1016/j.compbiomed.2022.105823. Epub 2022 Jul 6. PMID: 35872410.

[55] Hu, J., Liu, Y., Zhao, J., & Jin, Q. (2021). Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779*.

[56] Joshi, A., Bhat, A., Jain, A., Singh, A., & Modi, A. (2022). COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. *ArXiv*, *abs/2205.02455*.

[57] Cabral, R. C., Han, S. C., Poon, J., & Nenadic, G. (2024). MM-EMOG: Multi-Label Emotion Graph Representation for Mental Health Classification on Social Media. *Robotics*, *13*(3), 53. https://doi.org/10.3390/robotics13030053

[58] Wang, Y., Ren, Y., Bi, Y., Zhao, F., Bai, X., Wei, L., Liu, W., Ma, H., & Bai, P. (2024). Multimodal transformer graph convolution attention isomorphism network (MTCGAIN): A novel deep network for detection of insomnia

disorder. *Quantitative Imaging in Medicine and Surgery*, *14*(5), 3350-3365. https://doi.org/10.21037/qims-23-1594

[59] Liu, S., Wang, S., Sun, C., Li, B., Wang, S., & Li, F. DeepGCN based on variable multi-graph and multimodal data for ASD diagnosis. *CAAI Transactions on Intelligence Technology*. https://doi.org/10.1049/cit2.12340