American Journal of Psychiatric Rehabilitation

Expert Opinion Article

Leveraging Deep Learning For Automated Detection Of Mental Disorders: A Survey And Future Directions

Rahul Jena[1*], Utsav Muley[2], Trilok Gandhi[3], Ankesh Prajapati[4], Brijesh Parmar[5], Jayana Trivedi[6], Vivek Dave[7]

[1*]Faculty of IT & Computer Science, Parul University, Vadodara, India. rahuljena159@gmail.com
[2]Faculty of IT & Computer Science, Parul University, Vadodara, India. utsavmuley@gmail.com
[3]Faculty of IT & Computer Science, Parul University, Vadodara, India. trilokgandhi2001@gmail.com
[4]Faculty of IT & Computer Science, Parul University, Vadodara, India. ankeshprajapati217@gmail.com
[5]Faculty of IT & Computer Science, Parul University, Vadodara, India. brijeshpamarv@gmail.com
[6]Faculty of IT & Computer Science, Parul University, Vadodara, India. jayana10302@gmail.com
[7]Faculty of IT & Computer Science, Parul University, Vadodara, India. vivek.dave@paruluniversity.ac.in

***Abstract*—**The worldwide burden of mental disorders such as bipolar disorder, schizophrenia, depression, and anxiety now impacts millions of people, making them a major global health issue. The diagnosis of mental illnesses is a critical determinant of effective treatment and therapy strategies require initiation before accurate identification can be made. Artificial intelligence and in particular, deep learning have recently been adopted to carry out the automation of mental disorder identification from multimodal data including text, speech, and neuroimaging. This survey presents a comprehensive review of current approaches in deep learning techniques for mental health analysis and mitigation, including transformer models, recurrent neural networks (RNNs), convolutional neural networks (CNNs) and hybrid architectures. We review their application in different data sources including audio, medical documents, online social network posts, and brain signals such as EEG and MRI. We also outline some major challenges including data scarcity, ethical concerns, model interpretability issues and generalization challenges. In addition, we outline future research directions which include multimodal fusion, explainable AI, privacy preserving federated learning and real time mental health monitoring. The aim of this work is to serve as a starting point for scholars and practitioners on the application of deep learning to improve the diagnosis of mental health. This paper aims to present an overview of the deep learning methods for detection of mental disorders, their application, challenges and the possible future directions.

Keywords—Mental Disorder, Mental Health, Schizophrenia, Bipolar Disorder, Depression, Anxiety, Artificial Intelligence, Deep Learning, Convolutional Neural Networks, Recurrent Neural Networks, Explainable AI, Transformers, Autoencoders, EEG, MRI.

I. INTRODUCTION

A mental disorder can be defined as clinically significant behavioral, emotional, or thought disorder. It is generally marked by distress or impairment in significant areas of functioning, and there is a wide variety of mental disorders. In 2019, 970 million people, or 1 in every 8 people globally, had a mental disorder, of which anxiety and depressive disorders were the most prevalent [1]. One in two people worldwide will experience a mental disorder during their lifetime, reports a massive study conducted by Harvard Medical School and the University of Queensland researchers together. The study results are based on face-to-face, structured interviews of more than 150,000 adults in 29 countries representing the different levels of affluence from all parts of the world [2].

Mental health conditions have been strongly linked to a higher risk of death and cardiovascular issues. People with mental health issues have been found to be more likely to develop heart diseases, which can result in more serious medical complications and early death. A recent study examined more than 2.5 million cardiovascular deaths in the United States over a 20-year period [3]. The findings indicated that patients with mental disorders had higher age-adjusted cardiovascular mortality rates. Notably, patients with mental and behavioral disorders brought on by the use of psychoactive substances, patients who were male, and those who lived in rural areas experienced the largest increase.

Deep learning, a branch of artificial intelligence, has become a game-changing force in diagnosing mental disorders by solving long-standing issues in the profession. The conventional diagnostic approach is typically subjective scores, lengthy clinical interviews, and self-report data, which can result in misdiagnosis and untimely treatment. Deep learning employs computational power to process difficult, high-dimensional data such as neuroimaging scans, speech patterns, and text data from electronic health records or social media. Models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have attained staggering diagnostic performance, identifying mental health conditions such as schizophrenia, bipolar disorder, and depression with accuracy rates of more than 80% in most cases [4], [5]. Deep Learning enables

faster, more accurate, and objective mental health assessment by recognizing finer patterns and features which are difficult for clinicians to detect. In addition to improving diagnostic accuracy, deep learning has also introduced innovative tools for real-time monitoring and early intervention. Wearable devices integrated with deep learning models analyze physiological signals, such as heart rate variability and skin conductance, to detect stress and anxiety in real time [6]. At the same time, natural language processing (NLP) techniques, such as those powered by DL models like BERT, analyze social media and speech to pick up on warning signs of depression and suicidal ideation [5], [7]. Such new methods not only diagnose but also enable scalable and personalized mental health care solutions. Data privacy, model interpretability, and generalizability remain significant hurdles to broad deployment. All this considered, the integration of deep learning into mental health diagnosis is a paradigm shift, with enormous potential to enhance patient outcomes and revolutionize mental health care [4], [5].

The primary goal of this paper is to provide a comprehensive review of the current literature on the use of deep learning (DL) for diagnosing mental disorders. Mental disorders, including depression, anxiety, bipolar disorder, and schizophrenia, pose significant challenges towards effective and timely diagnosis owing to their subjective nature and overlapping presentation in clinics. With advanced artificial intelligence (AI), in particular deep learning, there has been a move away from the conventional ways of studying and diagnosing such disorders. In this paper, efforts will be made to review existing attempts, approaches, and uses of deep learning for this end as well as pointing out the limitations and problems that still beset real-world implementations.

By reviewing studies across various modalities, including neuroimaging, EEG, speech, text, and multimodal data, the paper highlights the manner in which DL models have reached all-time high levels of accuracy and reliability in the detection of mental health disorders. For instance, models like CNNs have been used to examine neuroimaging data for prediction of schizophrenia and bipolar disorder, while RNNs and transformers have demonstrated excellent performance in examining speech and text data for detecting depression [8,9]. The integration of multimodal data has further enhanced diagnostic capability by fusing behavioral, linguistic, and physiological inputs within a unified framework. Such advances demonstrate the revolutionizing capability of DL in creating scalable, automated, and objective diagnostic systems [9].

Despite these successes, interpretability of deep learning models, data availability with limited datasets, and data privacy remain the biggest hurdles to clinical adoption. The paper also attempts to identify such gaps in current research and offer directions on how they can be bridged in the future. For example, improving explainable AI (XAI) methods is one way to build trust with clinicians, while federated learning methods can assist in solving privacy concerns through decentralized data analysis [10,11]. The paper also emphasizes the importance of creating standardized, representative datasets that reflect global populations in order to mitigate biases during model training and testing [5].

This paper seeks to explore the ways in which deep learning (DL) technologies are transforming the diagnostics of mental health through critical literature review. The objective is to review DL model-based studies in processing data modalities like neuroimaging (MRI, EEG), speech, text, and multimodal data to detect and diagnose mental illnesses like depression, anxiety, schizophrenia, and bipolar disorder. The paper outlines the most popular deep learning architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), transformers, and autoencoders, and how they are helpful in enhancing diagnostic accuracy, scalability, and objectivity. The paper touches upon wearable technologies and real-time applications of DL in mental health care and highlights new ways and case studies [8].

This paper is intended for researchers, clinicians, and AI practitioners who are interested in knowing the potential of deep learning to transform mental health care. It is not a guidebook on how to use DL algorithms or conduct experiments. By establishing the scope of the discussion, this paper aims to present a clear and transparent summary of the impact of DL technologies with space for further research in ethics, model deployment, and technical complexity in other studies.

II. Data Modalities Used in Deep Learning for Mental Health

A. Text-based Approaches

Text-based data is an essential tool for mental health diagnosis, utilizing online social media platforms like Twitter, Reddit, and Facebook, where people often post their experiences and feelings. Deep learning models such as BERT and GPT examine this textual information to identify linguistic cues of mental disorders, i.e., depressive speech, emotional volatility, and suicidal thoughts. Fine-tuning these models with domain-specific datasets has been found effective in extracting salient patterns from casual texts [13]. Clinical summaries and electronic health records (EHRs) are also an important source of unstructured data, documenting symptoms, diagnoses, and treatments that deep learning algorithms parse to aid clinicians in determining early signs of disorders [4].

B. Speech-based Approaches

Speech signals provide a rich modality for mental health condition detection, since voice features such as tone, pitch, and rhythm can pick up on underlying emotional and cognitive states. Depression and schizophrenia and related disorders tend to appear in alteration of these voice biomarkers. More advanced models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have proved effective at processing time-series speech signals, extracting temporal patterns that allow for accurate diagnoses [9]. These methods facilitate real-time, non-invasive examination of speech signals obtained through clinical interviews or from digital health apps.

C. EEG and Neuroimaging-based Approaches

Electroencephalography (EEG) and Magnetic Resonance Imaging (MRI) are pivotal equipment in the realm of brain activity and structure analysis for mental health diagnosis. EEG signals record brainwave patterns that are applicable to identify depression, anxiety, or other diseases. Deep learning architectures such as CNNs and autoencoders are particularly good at processing EEG data, recognizing irregular patterns associated with mental health diseases. For instance, autoencoders have been employed to remove noise from EEG signals, enhancing signal quality for diagnostic purposes [14]. Likewise, MRI data processed by CNNs can identify structural abnormalities in brain areas related to disorders such as bipolar disorder and schizophrenia, yielding objective diagnostic information [8].

D. Multimodal Approaches

Multimodal methods combine information from various modalities—e.g., text, speech, and neuroimaging—to enhance diagnostic performance. By aggregating complementary sources of data, these systems account for the multi-faceted nature of mental illness. For example, a multimodal deep learning system could process speech characteristics, text-based linguistic features, and EEG signals simultaneously, offering a comprehensive picture of a patient's state. In [9], such systems have been demonstrated to perform better than unimodal systems in diagnosing complicated mental illness.

III. Deep Learning Techniques for Mental Disorder Detection

A. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have proven to be a revolutionary technology in neuroimaging data analysis, providing more precise and automated diagnosis of mental health illnesses. Neuroimaging modalities such as Magnetic Resonance Imaging (MRI), functional MRI (fMRI), and Electroencephalography (EEG) generate high-dimensional, complex data that is difficult to interpret by hand. CNNs are very good at dealing with such data because they can learn hierarchical spatial and structural patterns representations directly from raw input without requiring manual feature extraction.

MRI and fMRI scans are in common use for examining structural and functional abnormalities of the brain. CNNs have been found to be most efficient at examining these modalities by identifying important biomarkers like changes in grey matter volume, white matter integrity, and patterns of functional connectivity between brain regions. For instance, studies have identified that CNN-based models can differentiate between schizophrenia, bipolar disorder, and healthy controls with great accuracy. Through the detection of subtle differences in neuroanatomy, CNNs shed light on the biological origins of such disorders [8]. In fMRI, CNNs process time-series data to investigate dynamic brain activity, disclosing connectivity network abnormalities that are usually related to mental illness such as depression and PTSD.

EEG signals that quantify electrical activity of the brain are another paramount modality of mental health diagnosis. CNNs are well-fitted for processing EEG since CNNs can deal with spatial interrelationships (e.g., electrode placements) as well as with temporal dynamics (e.g., brainwave patterns over time). For instance, CNNs have been used to differentiate mental disorders like epilepsy, depression, and anxiety by recognizing characteristic patterns in EEG frequency bands like delta, theta, alpha, beta, and gamma. This obviates manual preprocessing and feature engineering since CNNs learn discriminative features directly from raw EEG signals [10].

The CNNs' layered architecture which consists of convolutional layers, pooling layers, and fully connected layers allows them to extract both local and global features. Convolutional layers extract spatial patterns, including localized variations in grey matter, while pooling layers compress data dimensionality, facilitating computations. Fully connected layers combine these features to make predictions, such as classifying a subject into diagnostic classes like "healthy" or "schizophrenic."

Moreover, CNNs are conducive to transfer learning, whereby pre-trained models (such as on the large dataset ImageNet) can be fine-tuned for target neuroimaging tasks, confronting the usual obstacle of scarce labelled medical datasets. This versatility has further expedited their application to mental health diagnostics.

Although they have merits, CNNs are challenging to apply in clinical settings. The "black-box" character of CNN predictions makes it challenging for clinicians to interpret decision-making, which is a problem from the perspective of trust and transparency.

Additionally, CNNs need a large and diverse dataset to prevent overfitting and perform generalizable across populations. These obstacles highlight the necessity of explainable AI (XAI) methods and standardized, high-quality datasets in the field.

B. Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs)

Recurrent Neural Networks (RNNs) and their advanced form, Long Short-Term Memory Networks (LSTMs), are commonly applied for sequential data analysis like speech, EEG signals, and texts. RNNs, unlike feedforward networks, possess an inbuilt mechanism of remembering previous inputs, and therefore they are highly suitable for handling time-series data. LSTMs improve further over RNNs by resolving the vanishing gradients issue, which allows them to learn long-term dependencies within data, important in the context of mental health diagnostics [10].

Speech is a dense modality for mental health diagnosis, as it holds acoustic and linguistic patterns that are frequently impacted by disorders like depression, bipolar disorder, and schizophrenia. RNNs and LSTMs are used to process time-series speech features such as pitch, energy, and spectral content to identify subtle variations in vocal tone and rhythm. For example, an LSTM model could monitor changes in speech patterns over time to detect symptoms of depression, including lower vocal energy and slower rates of speech. These models can process raw audio or extracted features and thus are useful tools for speech analysis in clinical or real-world environments [9].

EEG signals, which reflect brain activity, are sequential in nature and need models that can capture temporal dynamics. LSTMs are especially suited for this purpose, as they can examine the temporal evolution of brainwave patterns in various frequency bands like delta, theta, alpha, and beta. For instance, an LSTM can identify anomalies in the sequence of EEG signals that are related to mental health disorders such as epilepsy or anxiety. These models can further incorporate spatial and temporal properties, providing a more complete picture of brain activity [10].

Text-based information like clinical documents or social media posts frequently capture mental and emotional states, and as a result, it is a valuable resource for diagnosing mental conditions. RNNs and LSTMs are typically applied in natural language processing (NLP) applications to interpret sequential text data. To illustrate, they can monitor the context and sentiment behind written words over time and detect linguistic patterns that are characteristic of mental conditions like depression or suicidal thoughts. LSTMs work best at these tasks because they can carry context over large sequences of text and thus catch subtle emotional hints [9].

RNNs and LSTMs are very good at capturing sequential dependencies and are thus a goldmine for analyzing dynamic data in mental health diagnosis. They do, however, need large amounts of data and considerable computational power to train well. Moreover, their predictions can be less interpretable at times, which is a major hurdle for clinical uptake. In spite of these issues, developments like attention mechanisms and bidirectional LSTMs have further improved their performance, and they have become a mainstay of time-series analysis in the field.

C. Transformer (BERT, GPT, ViT)

Transformers have revolutionized natural language processing (NLP) by making it possible to analyze large-scale text data with unprecedented accuracy and contextual insight. Models like Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT), and Vision Transformer (ViT) are increasingly being used in text-based mental health diagnosis, examining data from social media, clinical notes, and patient records.

Social media sites like Twitter, Reddit, and Facebook tend to be places where people vent their emotions and states of mind. Models such as BERT and GPT are employed to scan these posts for indicators of depression, anxiety, and other mental illnesses. For instance, by scanning the linguistic style and emotional content of posts, these models can identify early warning signs of mental illness such as suicidal thoughts or social isolation [12], [13]. Pre-trained transformer models are fine-tuned on mental health datasets so that they can identify subtle language cues, including negative sentiment, disorganized thinking, or repetitive statements, which are characteristic of particular disorders.

Transformers are also used to analyze unstructured text in electronic health records (EHRs) and clinical notes. BERT models, for example, can detect mentions of symptoms, diagnoses, or treatments in patient histories to support the early detection of conditions such as PTSD or bipolar disorder. This use case is especially valuable for aggregating large amounts of clinical data to support healthcare professionals in diagnosis [13].

Such models as GPT-3 are multilingual and thus efficient in processing mental health data from a variety of populations. They can perform cross-lingual sentiment analysis and combine text with other modalities, including speech transcripts or neuroimaging reports, to gain a better insight into a patient's condition [12].

One of the main strengths of transformers for mental health diagnosis is that they can better comprehend the context of words than other forms of NLP. Text is processed bidirectionally by transformers, i.e., transformers examine the relationships between words in both previous and subsequent contexts.

This bidirectional text processing enables transformers to accommodate complex sentence structures and long-range dependencies and thus are best suited to picking up subtle expressions of mental health disorders. For instance, they are able to recognize subtle patterns of language that point to depression, like repeated negative thoughts or tone changes, that more basic models might miss [12], [13].

While having numerous benefits, transformers are plagued by major drawbacks in their application to mental health diagnosis. A major concern is the risk of bias in language models. Transformers tend to be trained on general-purpose text datasets, which may reflect societal prejudices. These biases can unconsciously influence the diagnostic performance of the models, especially for marginalized communities or groups. For instance, members of minority cultures or non-dominant language groups can be poorly represented in training data, which means predictions for such population are less accurate [13]. Solving this problem involves having more diverse datasets and using methods to reduce bias while training the model.

D. Autoencoders and GANs

Autoencoders and GANs (Generative Adversarial Networks) are also deep learning's strongest tools, which are often utilized for feature learning and data augmentation in mental health diagnosis. These models solve problems such as data sparsity, noise, and high-dimensionality in the dataset of mental health, hence are critical to drive AI-enabled mental health solutions forward.

Autoencoders are neural networks that learn low-dimensional and compact representations of data by compressing input data into a more compressed space and then reconstructing it. Autoencoders in mental health diagnosis are applied to extract significant features from intricate data such as neuroimaging (MRI, fMRI, EEG), text, or speech data. For instance, in EEG analysis, autoencoders are capable of denoising raw signals, eliminating muscle movement or extraneous noise-induced artifacts while maintaining the patterns of brain activity necessary for the diagnosis of disorders like depression or anxiety [14].

Additionally, autoencoders are especially well-suited to discover latent features that are not overtly visible in the unprocessed data. These representations can enhance the performance of downstream classifiers, including those that predict mental health states. In social media data or clinical note applications, autoencoders can discover hidden linguistic or semantic structure, enabling models to better pick up on subtle indicators of mental health conditions such as depressive language or emotional instability [14].

Generative Adversarial Networks (GANs) are commonly utilized to create artificial data, given the widespread issue of sparse and unbalanced data in mental health studies. Two parts make up GANs: first is a generator that produces artificial data and second is a discriminator that determines its validity. Through the successive enhancement of both, GANs are capable of generating quality, realistic data that closely represents the distribution of the original dataset. This is specifically helpful in neuroimaging since data collection here is costly and time-consuming. In [14], it has been shown that GANs have been utilized to generate simulated MRI or EEG samples, and these can complement current datasets with new samples that can enhance generalization of model-based diagnostics.

Besides data augmentation, GANs can also mimic rare conditions or underrepresented populations, making models more robust and inclusive. GANs, for example, can create speech or text samples with the linguistic patterns of people with certain mental health conditions, such that models can better detect conditions such as schizophrenia or PTSD. In addition, conditional GANs (cGANs) enable scientists to create data with precise attributes, for example, depressive or anxious text tones, making them precious in developing balanced training datasets [13].

Autoencoders and GANs augment other deep learning methods by facilitating improved feature extraction and dataset augmentation. Autoencoders enhance data quality and representation, whereas GANs solve data sparsity and balance problems, making them essential in constructing effective mental health diagnostic systems. Challenges still exist, such as the computational complexity of these models and the possibility of producing biased or low-quality synthetic data if the original dataset is not representative or diverse [13], [14].

IV. Future Directions

A. Explainable AI (XAI) for Trustworthy Models

Explainable AI (XAI) is increasingly becoming a keystone in the use of deep learning models for diagnosing mental illness, especially in clinical environments where trust and explainability are most important. Although deep learning models like convolutional neural networks (CNNs) and transformers are excellent at identifying mental illness disorders, their inherent “black-box” nature tends to make it difficult for clinicians to comprehend the basis of their predictions. This absence of transparency is likely to delay their use in high-risk applications like mental health, where decision-making demands a system that is interpretable, ethical, and accountable.

Explainable Artificial Intelligence (XAI) methods are intended to make deep learning models explainable or interpretable by providing reasons for predictions. For instance, techniques like Local Interpretable Model-Agnostic Explanations (LIME)

and SHapley Additive exPlanations (SHAP) are typically used to determine the features (e.g., specific words in text, speech, or brain areas identified in MRI scans) most impacting a model's prediction [4], [13]. For example, in the case of a text-based model for depression diagnosis, XAI may indicate what words of sadness or despair had been used predominantly and had led to the diagnosis. Similarly, through neuroimaging analysis, interpretable saliency maps generated using Grad-CAM (Gradient-weighted Class Activation Mapping) assist in identifying what regions of the brain were responsible for the prediction of an illness like schizophrenia or bipolar.

B. Federated Learning for Privacy-Preserving Analysis

Federated learning (FL) is a novel method of training machine learning models that maintains data privacy, hence its suitability in sensitive domains such as mental health diagnosis. In the conventional centralized training systems, data from various sources is brought together and housed at a central point, thereby posing substantial privacy issues, particularly on the use of personal data such as social media activity, clinical notes, or neuroimaging data. Federated learning solves these issues by allowing the models to be trained on local devices or within data silos, without necessarily transmitting sensitive data to a central server.

In psychiatric diagnosis, FL allows joint model training across institutions without violating data privacy laws such as GDPR and HIPAA. For example, research institutions and hospitals can train models on EEG or MRI data from various locations without exchanging identifiable patient information. This decentralization strengthens data security while enabling the development of more generalizable, robust diagnostic tools [5].

Federated Learning (FL) also addresses diversity of mental disorders by including data sets of under-represented communities, thereby providing greater inclusivity to artificial intelligence models. Despite all its strengths, communication overhead, heterogeneity of data, as well as aggregation bias in updates, still exist. Future studies may concentrate on integrating FL into more sophisticated encryption algorithms, such as differential privacy, as well as providing improved explainability to build greater clinician trust and facilitate easier use in clinical settings [5].

C. Multimodal Fusion Strategies

Multimodal fusion methods are a great step ahead for the field of mental health diagnoses as they seek to enhance accuracy in combining different modalities of information, from text and speech to brain imaging. Each mode presents unique representations of mental state disorders, e.g., depression, anxiety, and schizophrenia, which, when integrated, give more information about complexities of these mental illnesses. The future will be about fine-tuning the multimodal protocols to achieve an optimal balance for diagnostic validity and accuracy.

In multimodal fusion, early fusion methods, where concatenation of raw features extracted from modalities like text and EEG before model training, and feature-level fusion methods that combine intermediate features extracted from different modalities, have been demonstrated to yield encouraging results. These methods enable deep learning models to learn complementary patterns and, in the process, improve their diagnostic performance. For instance, text data can reveal emotional or cognitive states, while speech analysis can reveal vocal biomarkers, and neuroimaging data can reveal structural abnormalities in the brain. Combining these sources using state-of-the-art architectures, like multitask deep learning models or attention models, has the potential to yield more accurate and robust predictions [9], [11].

Future advancements could focus on overcoming challenges like data synchronization and variability across modalities. Additionally, developing adaptive multimodal frameworks that prioritize the most relevant modalities for each task and patient could further enhance the applicability of these systems. Integrating explainable AI (XAI) into multimodal systems will also be critical to gaining clinician trust and ensuring widespread adoption in real-world mental health diagnostics [9].

V. Conclusion

This survey is focused on the revolutionary potential of deep learning in the diagnosis of mental health. Key findings illustrate the way in which deep learning models like CNNs, RNNs, transformers, and autoencoders have shown extraordinary potential in handling various modalities of data like neuroimaging, speech, text, and multimodal data. These technologies have provided avenues for automated, scalable, and objective diagnostic platforms with high accuracy in identifying complex mental disorders like depression, anxiety, schizophrenia, and bipolar disorder. Novel applications like real-time monitoring via wearable devices and sentiment analysis of social media further highlight the deep learning's pragmatic potential in the provision of early interventions and personalized care.

The integration of deep learning technologies in mental disorder detection holds the potential to transform this industry fundamentally by providing precision medicine options, making mental disorder care more accessible, and minimizing mental disorder stigma through digital normalization of assessment. However, hurdles such as lack of data, ethical issues, and the lack of interpretability of deep learning models are still key challenges. Future work

needs to be oriented toward the construction of more sophisticated multimodal fusion approaches, the creation of explainable artificial intelligence (XAI) systems for developing clinician trust, the use of privacy-preserving mechanisms such as federated learning to maintain data safety, and more complex personalized AI models for more tailored predictions. As technology continues to improve, deep learning has the potential not just to improve the accuracy of mental health diagnosis but to transform the entire structure of mental health care, thereby driving improved outcomes for individuals and populations.

REFERENCES

[1] Institute of Health Metrics and Evaluation. Global Health Data Exchange (GHDx), (https://vizhub.healthdata.org/gbd-results/, accessed 14 May 2022).

[2] McGrath, J. J., Al-Hamzawi, A., Alonso, J., Altwaijri, Y., Andrade, L. H., Bromet, E. J., ... & Zaslavsky, A. M. (2023). Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries. *The Lancet Psychiatry*, *10*(9), 668-681.

[3] Ebert T, Hamuda N, City-Elifaz E, Kobo O, Roguin A. Trends in CV mortality among patients with known mental and behavioral disorders in the US between 1999 and 2020. Front Psychiatry. 2023 Nov 1;14:1255323. doi: 10.3389/fpsyt.2023.1255323. PMID: 38025453; PMCID: PMC10646424.

[4] Abd-Alrazaq, A., Alhuwail, D., Schneider, J., Toro, C. T., Ahmed, A., Alzubaidi, M., ... & Househ, M. (2022). The performance of artificial intelligence-driven technologies in diagnosing mental disorders: an umbrella review. *Npj Digital Medicine*, *5*(1), 87.

[5] Iyortsuun, N. K., Kim, S. H., Jhon, M., Yang, H. J., & Pant, S. (2023, January). A review of machine learning and deep learning approaches on mental health diagnosis. In *Healthcare* (Vol. 11, No. 3, p. 285). MDPI.

[6] Razavi, M., Ziyadidegan, S., Jahromi, R., Kazeminasab, S., Janfaza, V., Mahmoudzadeh, A., ... & Sasangohar, F. (2023). Machine learning, deep learning and data preprocessing techniques for detection, prediction, and monitoring of stress and stress-related mental disorders: a scoping review. *arXiv preprint arXiv:2308.04616*.

[7] Qiao, J. (2020, August). A systematic review of machine learning approaches for mental disorder prediction on social media. In *2020 International Conference on Computing and Data Science (CDS)* (pp. 433-438). IEEE.

[8] Li, Z., Li, W., Wei, Y., Gui, G., Zhang, R., Liu, H., ... & Jiang, Y. (2021). Deep learning based automatic diagnosis of first-episode psychosis, bipolar disorder and healthy controls. *Computerized Medical Imaging and Graphics*, *89*, 101882.

[9] Zhang, Z., Lin, W., Liu, M., & Mahmoud, M. (2020, November). Multimodal deep learning framework for mental disorder recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (pp. 344-350). IEEE.

[10] Rivera, M. J., Teruel, M. A., Mate, A., & Trujillo, J. (2022). Diagnosis and prognosis of mental disorders by means of EEG and deep learning: a systematic mapping study. *Artificial Intelligence Review*, 1-43.

[11] Ma, D., Zhang, H., & Wang, L. (2024). Deep learning methods and applications in brain imaging for the diagnosis of neurological and psychiatric disorders. *Frontiers in Neuroscience*, *18*, 1497417.

[12] Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, *10*(1), 11846.

[13] Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*, *34*(1), 721-744.

[14] Bozhkov, L., & Georgieva, P. (2018, July). Overview of deep learning architectures for EEG-based brain imaging. In *2018 International joint conference on neural networks (IJCNN)* (pp. 1-7). IEEE.