A Study of Artificial Intelligence Frameworks and Their Capability to Diagnose Major Depressive Disorder

Oluwafeyisayo Oyeniyi[1], Shreyansh Sandip Dhandhukia[2], Amartya Sen[1](✉), and Kenneth K. Fletcher[2]

[1] Oakland University, Rochester, MI 48309, USA
{ooyeniyi,sen}@oakland.edu
[2] University of Massachusetts Boston, Boston, MA 02125, USA
{S.Dhandhukia001,kenneth.fletcher}@umb.edu

Abstract. The accurate diagnosis of mental illness is challenging because mental illness does not result in evident physical symptoms as compared to physical illness like the common cold. As a result, no definitive medical tests exist for mental illnesses. This situation is further aggravated by the fact that many of the symptoms of various mental illnesses overlap. Further, traditional means of mental care and therapy are not easily accessible to a majority of the population in developed and developing countries alike. In addition, openly discussing mental illness in major parts of society is still considered taboo. Therefore, a plausible way to improve mental illness diagnosis and address the aforementioned challenges is by using Artificial Intelligence (AI). This paper presents a comprehensive survey of AI-enabled approaches to Major Depressive Disorder (MDD) diagnosis and outlines some future research directions and challenges in this field. The paper also presents a preliminary system architecture of an AI-enabled approach to diagnose mental health with the objectives of making the underlying system more user-centric, scalable and accessible.

Keywords: Artificial intelligence · Explainable AI · Machine learning · Mental health

1 Introduction

A human being's good mental health is vital to ensure emotional, psychological, and social well-being, and any condition that affects this well-being is known as a mental disorder. One such disorder is Major Depressive Disorder (MDD), or simply depression. According to the World Health Organization (WHO)[1], symptoms of depression are characterized by an individual's lack of interest in previously enjoyable activities and persistent sadness.

Motivation: Depression is a leading cause of disability worldwide, with almost 75% of individuals suffering from MDD in developing countries who remain

[1] https://www.who.int/health-topics/depression.

H. Hacid et al. (Eds.): ICSOC 2021 Workshops, LNCS 13236, pp. 3–17, 2022.
https://doi.org/10.1007/978-3-031-14135-5_1

untreated and approximately 1 million susceptible cases lead to suicide [3]. In the United States (U.S.) alone, MDD affects more than 16.1 million American adults each year, which constitutes about 6.7% of the U.S. population aged 18 and older [3]. Additionally, due to the global pandemic from SARS-CoV-2 (COVID-19) in the year 2020 itself, depression diagnoses were up by 873% [17].

The demand for mental health services is rapidly increasing, as shown in a 2018 survey by National Council on Behavioral Health (NCHB)[2]. The survey concluded that at least 56% of people want access to mental health services but face many barriers. These barriers can be attributed to a lack of resources and/or trained healthcare providers, the social stigma associated with mental disorders, and inaccurate assessment[3]. Another survey in 2018 [14], showed that there is a shortage of mental health care professionals in every state across the U.S. This situation is further aggravated due to the high cost and insufficient insurance coverage for mental health conditions, leading to the difficulty of accessing mental health services by economically challenged population.

Nonetheless, most people suffering from mental illnesses have a prevalent behavior of wanting to be alone. Due to this isolation, mental disorder patients seek online venues like Twitter, Facebook, or Reddit to openly or anonymously share about discomforts and anxieties. This gives rise to data repositories comprising a variety of user-curated contents on social media platforms such as personal status, user's pictures, and geo-location, which can be mined for information. While humans have limited capacity to learn, an AI actuated system can easily access thousands of medical information sources and help in the early detection of chronic mental health diseases in patients.

Background: The Need for AI. Traditional means of providing screening (diagnosis) and monitoring, although good, are not sufficient today due to two different reasons. First is the Big Data Connection. The information generated from pervasive devices with Internet connectivity and social media platforms can aid in detecting and monitoring depressive behavior. However, analyzing Big Data generated from online platforms is not supported by traditional care and therapy. Second is the ease of accessible care and constant monitoring of MDD patients. Traditional means cannot address the accessibility issue and provide constant monitoring in contrast to AI-enabled frameworks to address mental well-being. So it is necessary to support and extend the methods of traditional care by using AI-based frameworks to leverage the advancement in computer technology for social good.

A diverse technical solution has been adopted and implemented to help solve the limited access to a medical professional. The solutions range from chatbots, IoT/Wearable devices, mobile and web applications to behavioral technologies [1,7,8,11,13,18]. These solutions have helped to significantly increase the access to the professional help needed for individuals suffering from MDD. The nature of these solutions makes them easily accessible, thereby reducing the stigma that is associated with MDD. Further, The type of study conducted in this paper

[2] https://rb.gy/8hpftt/.

[3] https://rb.gy/zgznyj.

will also benefit other domains like Analysis of Behavioral Disorders in business processes such as banking recommender systems and cognitive recommender systems [4,5].

Objectives and Contributions: Given the importance of AI-enabled frameworks towards diagnosing and monitoring MDD, this work surveys and discusses the different AI-enabled approaches to MDD that have been proposed in the recent literature. The contributions of this work are as follows:

- Review existing literature on AI-enabled frameworks for diagnosing mental health illnesses and summarize existing research challenges and some future directions.
- Propose a decentralized system design of an accessible and scalable AI-enabled approach for diagnosing mental health illnesses.

2 AI-Enabled Approaches

The **systematic search strategy** along with some exclusion criteria were as follows. The paper identified related works based on their solutions that addressed features like reliability, accuracy, accessibility, and explainability. The works prior to 2015 was excluded along with any work that had no testing or implementation, or if the data was not gathered from a valid source like health agencies or social media. For recent literature surveyed belonging to each domain, the paper presents a discussion on their algorithmic description, and the input and output type.

2.1 Expert System and Fuzzy Logic

In [2], the authors proposed an expert system that can be used to diagnose depression in an individual. Due to the nature of depression diagnosis, the expert system would help the psychologist to appropriately diagnose an individual. The expert system was created using Simpler Level 5 (SL5) object language, with its engine implemented in Delphi Embarcadero RAD Studio XE6. The proposed expert system interacted with the human subjects by asking them a set of Boolean questions. Based on the answers chosen, the expert system outlined a diagnosis and recommendation to the user. The knowledge base for the expert system was created with the information documented from experienced psychologists and specialized websites for depression. The proposed expert system thereafter was evaluated by an experienced psychologist, who verified the correctness of the system output. The benefits of the expert system can be categorized by its ease of use. Further, the system can be deployed as a standalone system, without the need for an intervention from a medical professional.

The authors in [18] designed a web-based expert system using fuzzy logic to diagnose individuals suffering from depression and determine the level of severity. The system was designed to be user-centric; it could be implemented in specialized settings such as in a psychiatric office as well as being used by the user solely.

Fuzzy logic was used to address the uncertainty or ambiguity present in human knowledge and the decision-making process. Knowledge was acquired from several resources and professionals, with the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) being the main source. The scale for the severity of depression disorder used was "very low, low, medium, high and very high". The expert system was implemented in Jess. The system's predictive results were evaluated by psychological consultants who verified the accuracy of the result. As reported by the authors, their proposed system also achieved a 79% and 98% outcome for the metrics of sensitivity and specificity, respectively.

Using Fuzzy logic gives allowance for uncertainty in decision making which makes it flexible and easily adaptable to different scenarios. The level of uncertainty or vagueness present in the decision-making process makes the preciseness and accuracy of the results obtained questionable.

2.2 Machine Learning Approaches

The current trends in digitization have penetrated many industries including healthcare. The concurrent growth in medical demands and improved efficacy of machine/deep learning algorithms can help medical institutions and their professionals to detect early signs of chronic diseases, which helps to reduce the time and lower the costs of treatment. This section summarizes the various Machine Learning (ML) based frameworks that have been proposed. The literature within this category is further analyzed based on sub-categories like the type of ML algorithms ((un)supervised, semi-supervised, and reinforcement learning), along with the proposed model's respective algorithmic input and output. Although, a net total of about 15 works have been reviewed in this category (Sect. 4), owning to page limitations, in this section, the description of only a select few notable papers have been presented. For a complete description of the reviews, readers are kindly referred to the following document - https://rb.gy/ajeq5q.

Supervised Learning. Machine learning approaches that belong to this algorithmic category utilizes historical and classified input and output to train themselves and facilitate the intended functionalities.

Summary: To begin with, one such literature that belongs to this algorithmic category of supervised learning is [19]. It discusses the prevalence of Major Depressive Disorder (MDD) and Generalized Anxiety Disorder (GAD) in youth attending universities. The authors state that if such issues are not diagnosed at an early stage then it leads to drinking-related harm and alcohol abuse. The paper [19] proposes a machine learning model that predicts if an individual suffers from MDD and GAD by creating a so-called Electronic Health Records (EHR) data set. The EHR comprised of an undergraduate student's biometric and demographic data with the exclusion of all psychiatric features to preserve privacy. An ensemble algorithm consisting of Support Vector Machine (SVM), XG Boost, K-Nearest Neighbor (KNN), Random Forest, Logistic Regression, and Neural Network was deployed using Bayesian hyperparameter optimization.

Algorithmic Input: The input to the machine learning model consisted of an individual's physiological data like blood pressure, Body Mass Index (BMI), heart rate, and demographic data like housing status, health insurance, and age. The authors used non-psychological factors such as age, housing zip code, and health insurance from Electronic Health Record (EHR) data set to predict if a college student is suffering from Major Depressive Disorder (MDD) or General Anxiety Disorder (GAD), and explain the model's result.

Algorithmic Output: As per the author's reporting, the ensemble model achieved 70% accuracy and some of the top features were satisfaction with living conditions, public health insurance, and parental home. XG Boost classifier was used to predict an individual status if they are suffering from depression or not. Overall, the Area Under Curve (AUC) scores obtained from each model in the ensemble model indicated the model accurately predicted an individual's state, and also Shapely Additive Explanations (SHAP) was used to explain how the importance of each feature affects the overall result of a model's performance. The obtained results allowed the authors to validate that depression can be diagnosed using non-psychiatric features.

Summary: In [12], the authors have proposed the incorporation of digital intervention in predicting depression and other forms of anxiety symptoms. The authors aimed to evaluate the improvement experienced by a patient who received care via digital intervention after hospital discharge from illness around work-related stress. This research was conducted in a randomized controlled trial of 632 people who received the digital intervention. This group of users was split into two groups - a group using the digital interventions for their follow-up and the other group receiving information from a professional. This study was done for 9 months after the hospital's discharge.

Algorithmic Input: The authors implemented an ensemble model which had 2 layers to analyze the data. The first layer, known as the base model consisted of ridge regression, random forests, general linear models, Gaussian process, support vector machines, K-nearest neighbors. The second layer was the averaging layer, which took as input the mean prediction for a given subject across the different models and validation folds.

Algorithmic Output: Based on the output of the results, the authors stated that their ML model predicted a change in the depression of a person. The authors also stated that the ML model could be capable of guiding a clinician to know the best way to allocate resources in caring for a user, either using a higher level of traditional care or a low-level resource consisting of digital intervention.

Summary: The authors in [8] aimed to evaluate chatbots utilization by users. Tess is the chatbot that was used as a case study in this paper. Tess deciphers a user's emotional needs through the content of their conversation. Tess has 12 modules and its questionnaires or content spans across different areas of interest such as self-compassion, cognitive distortion, coping statements, and so on. These areas make Tess a robust chatbot but also present a herculean task for the users to go through the modules in good time.

Algorithmic Input: Tess was deployed on Facebook messenger with anonymous data collection. The authors aimed to analyze the usage of Tess, understand user's flows between the various modules and the usage of each module.

Algorithmic Output: From the analysis carried out, the authors showed that a chatbot is an effective tool for mental health, just the major modules required to assist a user should be implemented in a chatbot, and the overall usage across the different modules was based on some factors such as the questions asked and the time required to complete a module.

Summary: In this work [1], the authors developed a depression diagnosis algorithm using an individual's sequence of responses to a virtual agent and determined their depression severity.

Algorithmic Input: This research involved diagnosing depression through sequencing of responses obtained from a user using audio and text features. The responses were categorized as responses to specific questions asked from the Patient Health Questionnaire (PHQ) and responses that are not dependent on the questions asked. The authors carried out this experiment using three different scenarios to evaluate the effectiveness of their model. In the first scenario called the context-free modeling, a regularized logistic regression was deployed and the time a question was asked was the key factor and not the question type. In the second scenario called weighted modeling, a regularized logistic regression was deployed and the question type, not the timing of the question was the key factor. In the third scenario called sequence modeling, a bi-directional Long Short-Term Memory (LSTM) was implemented.

Algorithmic Output: The authors discovered that in the cases of context-free modeling and sequence modeling, the text features gave a better result. Whereas, in the case of weighted modeling, the audio features performed better. Overall, the best result was obtained from the weighted modeling for both the text and audio data.

Unsupervised Learning is a type of machine learning technique where the label is not pre-defined but determined by the clustering of a set of data points. The labels are defined by the patterns discovered in the data set. In this subsection, the paper presents a discussion on existing works that fall in the category of unsupervised learning to address Major Depressive Disorder.

Summary: In [21], the authors implemented an unsupervised machine learning algorithm using K-Means Clustering to predict if an individual is suffering from depression along with predicting their depression severity. K-Means clustering uses the position of each data point instead of the summary of data points, thereby making its prediction output more holistic in nature. Moreover, the model's results were compared with the results from a traditional norm-based classification to evaluate the model's performance. Middle and High school Chinese students were the focus of this study.

Algorithmic Input: The K-Means clustering was implemented in a 13-dimensional space with 13 features obtained from Beck Depression Inventory

(BDI) questionnaire. The clusters centers were determined using the maximum and minimum points. The authors classified the severity of depression as none, mild, moderate, and severe.

Algorithmic Output: The model was able to correctly predict if an individual was suffering from depression, and when compared to the traditional norm-based classification models, the model had higher accuracy and AUC score.

Semi-supervised Learning is a combination of supervised and unsupervised learning. In semi-supervised learning, a small amount of labeled data is used along with a large amount of unlabeled data, which can be helpful when there is an unavailability of a large amount of labeled data.

The research work presented in [22] is an example of semi-supervised learning wherein the research objective was to monitor the clinical depressive symptoms from Twitter data that imitates the PHQ-9 survey used by clinicians. The authors proposed two approaches to detect the possibility of users suffering from depression. The former was a bottom-up approach where authors used distributional semantics to unwrap the symptoms of depression. The first approach is based on Latent Dirichlet Allocation (LDA), which is specifically unsupervised learning, views tweets as a mixture of latent topics, where a topic is a distribution of co-occurring words. However, the topics learned by LDA are not specific enough to correspond to depressive symptoms. The latter approach was based on a top-down approach where the authors added supervision to LDA by using a probabilistic topic modeling approach, named semi-supervised topic modeling over time (ssToT).

Algorithmic Input: The authors collected data of 45000 Twitter users with self-declared depression and another 2000 tweets of undeclared users. After the examination, it was seen that most users talked about their family and companion issues and the requirement for their help. The authors compared the ssToT learned topics with existing semi-supervised and unsupervised learning approaches such as k-means clustering, LSA, LDA, BTM, Partially Labeled LDA. These experiments suggested that ssToT outperformed all the state-of-art models paying little heed to the corpus that probabilities are acquired from. ssToT model was also tested as a multi-label classifier, using 10400 tweet data-set in 192 buckets. Each bucket contains tweets that are posted by the client inside a range of 14 days. The ssToT model was trained on an unlabeled data set and the performance was estimated by utilizing the labeled data-set.

Algorithmic Output: Accuracy and precision were measured for the ability to predict the presence of 9 depressive symptoms (in compliance with PHQ9) and they were found to be 0.68 and 0.72 on average. The obtained results suggest that semi-supervised topic modeling overtime was successfully able to capture depression symptoms from the Twitter data-set which was competitive with a fully supervised approach.

Reinforcement Learning enables agents to learn by their interaction with the environment by trial and error using feedback from past actions and experiences.

This technique is based on rewards and punishment based on the agent's set of actions to perform a task.

Summary: In [6], the authors aim to understand the relations between model-derived reinforcement learning parameters with the depression symptoms and symptom change after the treatment. Studies were conducted on 101 adults of which 68.3% were females. The authors also assessed the changes in model-learning parameters and symptoms after some of the participants received Cognitive Behavioral Therapy (CBT).

Algorithmic Input: The participants responded to the Mood and Anxiety Symptom Questionnaire (MASQ), a validated self-report measure of the symptom of anhedonia, negative affect, and arousal, as well as the Beck Depression Inventory to assess overall depression severity, the Wechsler Test of Adult Reading to estimate verbal IQ, and a demographic questionnaire. Out of 101 participants in the study, a total of 69 participants suffered from MDD and 32 participants had no history of MDD. The computational model analysis of behavior choices and neural data identified contemporary learning with symptoms during reward and loss learning.

Algorithmic Output: The results showed that during reward learning, the reward values increased slowly with increased anhedonia. The anhedonia was associated with model-derived parameters such as learning rate, outcome sensitivity, and neural signals. The results during the loss function manifested that learning parameter was associated with negative affect were found to be outcome shift, and the model-learning parameter associated with disrupted neural encoding of learning signals. After mapping the reinforcement learning model parameters with the symptoms of MDD it was observed that after CBT the features associated with the symptoms had shown possible learning-based therapeutic processes.

3 Explainable AI (XAI)

The recent advances in the utilization of AI for the medical field have been restricted due to the absence of interpretation for complicated models. This decreases the trust of clients when it comes to the output generated by AI-enabled models. The AI models that explain its outcomes are better known as explainable AI (XAI) and can be classified into two major categories [15]. First, as Ante-hoc, where the framework incorporates logic in the model. Second, as Post-hoc, where the frameworks utilize a more simple model to clarify a black-box model. In the following section, the paper discusses some of the notable literature from the mental health domain that proposes the usage of XAI.

Summary: In [24], authors propose to detect the early signs of depression from users' social media posts. They combined XAI with natural language processing (NLP). The XAI model used Local Interpretable Model-Agnostic Explanations (LIME) [9] for interpretation. LIME is a local model interpretation method utilizing local surrogate models to surmise forecasts of black-box models. The authors observed that LIME interpretation was sensitive to pre-processing of the data set, especially the choice of whether to remove stop-words from the data set. The authors removed stop-words from the data set to study the behavior of occurrence of personal pronouns in the depression data set.

Algorithm Input: consisted of Urdu and English text data from the social media platform Reddit and applied NLP algorithms to predict if a user is suffering from depression. The author used bag-of-Words (BoW) and term-frequency times inverse document-frequency (TF-IDF) features and used them in machine learning classifiers like Logistic Regression and Random Forest classifiers.

Algorithm Output: for the author's experimental setup, the Logistic regression classifiers performed better with an F1 score of 0.89 for TF-IDF features as compared to the Random Forest classifier (F1 score of 0.84).

Summary: In [23], the framework proposed by the authors incorporated the SHapely Additive exPlanations (SHAP) [16] for the purpose of model interpretation. SHAP uses values that are an average of marginal contributions of a single feature value across all possible partnerships of the features. It helps to represent the impact certain features have on the model outcome with which these shapely values are associated. The authors in [23], used these shapely values to assess and foster a novel AI approach for predicting mental health risk in individuals with diabetes mellitus.

Algorithm Input: Data was collected from 142,432 people suffering from diabetes, and their mental health status was verified using two sources - claims data and medication prescription data. The participant's behavior was classified into four categories including demographics and glucometer, coaching, and event data. Data sets were then collected to make member period occurrences, and descriptive analyses were performed to comprehend the connection between psychological well-being status and passive sensing signals. The model used for training the data set was an ensemble of ten LightGBM models and it was evaluated using sensitivity, specificity, precision, the area under the curve, F1 score, accuracy, and confusion matrix.

Algorithm Output: The outcome as reported by the authors showed that the proposed model performed with a score of more than 0.5 for sensitivity, specificity, AUC, and accuracy. The SHAP values determined the elements that offer more towards the classification output, such as demographics (race and gender), participant's emotional state during blood glucose checks, time of day of blood glucose checks, and blood glucose values. The authors were able to effectively anticipate the mental health risk in individuals experiencing diabetes and showed the component significance of elements that contributed most towards the classification output of the model.

Summary: The authors in [20] introduced a framework (What-If tool or WIT) that can visually analyze the ML systems. The WIT tool supports both local and global interpretation, that is it can analyze a local data point as well as analyze model behavior across the whole data set. This framework allows the users to find the nearest counterfactual to better understand the model's behavior. The tool also enables users to analyze the relationship between the features and the predicted output using partial dependence plots, which helps to better understand what features contribute more to the prediction outcome.

Algorithm Input: The authors gave three contextual analyses to show the utilization of WIT to analyze the model performance. The first study, directed by ML specialists, utilized the WIT to analyze regression models. A sample of 2000 data-set was stacked in the regression model. In the second study, a programmer of a large technology company used WIT to predict the health metric for clinical patients. WIT coupled with two regression models tracked down that the inference results and the data-point visualization results were bigger for the first regression model than the second. More interestingly, it was observed that there was a bug in the calculation that caused the first and the last value of the input feature to trade. The error matrix was sensible for the model, but the model was tackling an alternate issue because of this bug. Finally, the third study was led by computer science students from M.I.T examining the legitimacy of the stop-and-frisk practices of the Boston police department. They used a data set of 150,000 records and prepared a linear classifier model. The data-set had features like age, race, sex, and criminal record of an individual being halted by police alongside the date and location of the offender.

Algorithm Output: The first study observed that the partial dependence plot was flat for some features of every data point. In the second case study, the WIT tool found a bug in the software that prevented a certain feature to be fed to the model, which resulted in a wrong decision. The output of the classifier in the third study was either positive or negative i.e. if the offender was searched or not searched during the confrontation. This outcome was coupled with WIT that identified features like age, gender, and race played a key role in determining whether the offender was frisked or not.

4 Discussions

In this section, Fig. 1 and 2 summarizes the literature classified using different categories like machine learning algorithm types, input data type, and so forth. The numerical values in these illustrations represent the number of reviewed papers that belong to a respective category. Further discussions presented in this section illustrate some of the design challenges identified in this field, the scope of future directions, and the tentative design of an AI system to address mental health illnesses keeping in mind computational design parameters like accessibility, efficiency, and effectiveness.

4.1 Lessons Learned: Current Challenges and Future Directions

An important finding of this survey is that majority of the AI-enabled approaches belong to the category of Machine or Deep Learning frameworks (ML). This is

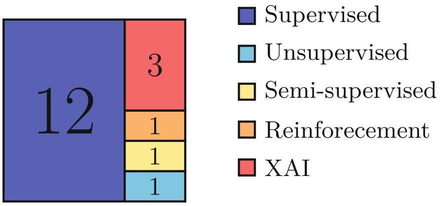

Fig. 1. Literature classification - machine learning algorithm types

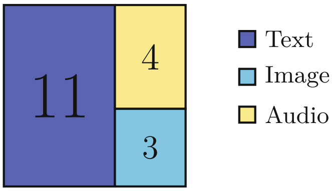

Fig. 2. Literature classification - algorithm input types

owing to the dynamic nature of the information associated with depression and the growing number of patients, which makes fuzzy and expert systems an ill-fit in terms of computational design criteria. The process of developing expert systems or fuzzy logic-based systems requires extensive information gathering, which can inherently be tedious. Thereafter, these systems require a knowledge base, to be developed from the gathered information. These knowledge bases can quickly become outdated and in addition, are challenged with issues of scalability and accuracy. Therefore, the majority of the existing works use ML frameworks since a ML model can be built using a lot of data making it robust and can be retrained with new data, making it accurate, scalable, and up-to-date. A model can be easily trained and deployed quickly as the professional input needed can be obtained from a variety of sources in a short time.

The second finding from the survey is that, within the domain of ML frameworks, the majority of the existing works incorporate the supervised learning methodology compared to unsupervised, semi-supervised, or reinforcement learning methods. The reason behind the popularity of supervised learning can be attributed to a few cases. First, in the mental health research domain, the majority of the AI-enabled frameworks consider the problem (e.g. MDD prediction) as binary classification - depressed or not depressed. This makes labeling easier when compared to say unsupervised learning, which tries to infer its clusters based on the data set which can be challenging at times. Hence, there is a preference for supervised learning frameworks, where the classes are already defined and the models know what they are looking for. Secondly, in these domains, the availability of a reliable, accurate, and a reasonably substantial amount of data is a bottleneck. This bottleneck proportionately impacts the performance of any ML-based AI framework. Supervised learning when compared to unsupervised learning does not require too much data. Third, given the current state of the art of AI systems, in the medical field, it is still favorable to have a human expert intervention. Supervised learning allows this process at an early stage i.e. when the models are being trained and thereby allowing for more accurate models when compared to unsupervised or semi-supervised learning methods.

Additionally, the reinforcement learning framework is not best suited for this nature of problem i.e. depression diagnosis. In reinforcement learning, the agent goes through different states, and depending on the outcome, they either get rewarded or punished. Due to the dynamic nature of how depression can be diagnosed in an individual, there might be a chance of the agent not prioritizing a particular state or discounting a state because it does not seem like a 'norm' when it can be useful in diagnosing depression in the individual. Also, there is a question of how long would the agent run before it arrives at a decision and can prove the result obtained is valid.

The functional benefits of supervised learning algorithms over their counterparts currently make it a more preferable choice. However, looking ahead, researchers in this domain need to assess an important question - if there is a paradigm shift towards unsupervised (or semi-supervised) learning algorithms, what kind of benefits and challenges will be encountered? To begin with, the benefits of such a paradigm shift will be in the ability to identify patterns that exist in diagnosing depressive disorders that are currently unknown even within medical professionals. With this in mind, the outcome of the AI models no longer needs to be a simple yes or no classification. The ranking of depression severity would become more flexible as these algorithms will be able to analyze patterns, relationships, and causality that exist in the data points. This will help to identify features that are easily susceptible to a depression level in an individual.

However, the primary challenge will be in terms of data gathering. The unsupervised and semi-supervised models require a lot of data. Thereafter, functionally designing these frameworks will also be challenging. An imperative question in this regard is - as the patterns are being established across the data points, if the first data set classifies individuals suffering from depression, would the second iteration from the result obtained from the first iteration be used to evaluate depression severity? Researchers will also need to address the issues in a model's bias. In case a data set is obtained from a set of the population (say, individuals attending college), it is possible the features used to predict depression would be different from another set of the population (say, middle-aged career individuals). As such, one will not be able to design a one-size-fits-all framework.

In terms of data collection, currently, the common means through which most AI-enabled technologies gather data is through text. This text could be through social media sites or users answering questionnaires delivered through chatbots. According to the investigations done in this field, a person's social media activity via content posted such as the kind of words used and the frequency of the words used can serve as an indicator for recognizing if a person is suffering from depression. Additionally, the use of biological data is an effective way of diagnosing individuals who are suffering from depression due to the information that can be easily seen when the data is being read. Moreover, using non-psychiatric features to predict depression is also a helpful way of predicting if a user is suffering as it helps to minimize the stigma that is attached to a depression diagnosis. It also helps in estimating or predicting the likelihood of a certain class of individuals suffering from depression if certain features are identified.

Finally, the researchers in this domain need to keep in mind that the acceptance of their proposed framework by the end-users can only come through model outcomes that are trustworthy and reliable. This can be achieved by making the proposed frameworks more transparent to end-users. These requirements necessitate the integration of explainable AI (XAI) frameworks to some degree. The incorporation of XAI frameworks will allow end-users with or without prior experience in the domain to be able to accept (trust) or reject a prediction if they understand the reasoning behind it. For example, if a system predicts if a person suffers from a disease, the expert (doctor) can see the symptoms that contribute to the prediction to either accept or reject a prediction.

4.2 Proposed AI System Design

In this section, based on the challenges discussed in the previous section, the paper presents a system architecture of an AI-enabled approach for the diagnosis of mental health disorders like MDD. The goal of this work in progress is to design a decentralized system that will be capable of leveraging different AI-enabled approaches and work with different input data types with an explanation of the generated output. The motivation behind such a computational design is driven by the need for increased scalability and accessibility by improving framework flexibility, which has not been addressed in the existing frameworks. The system design for the proposed AI framework is illustrated in Fig. 3.

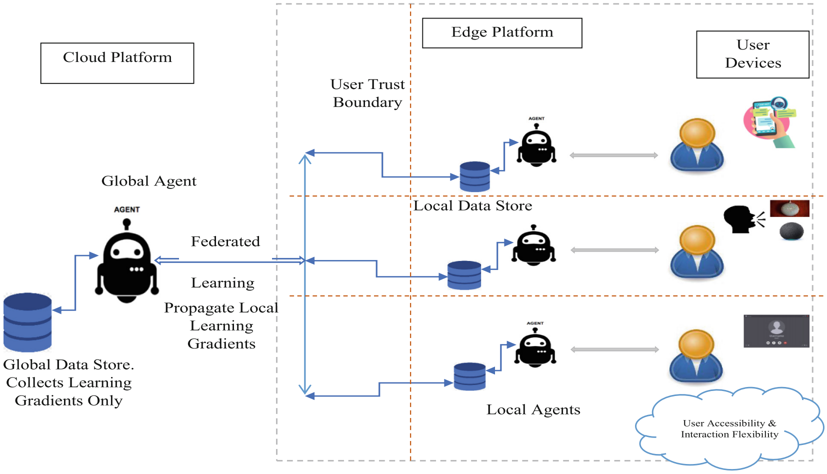
Cloud Platform
Edge Platform
User Devices
User Trust Boundary
Global Agent
AGENT
Federated
Learning
Propagate Local Learning Gradients
Local Data Store
Global Data Store. Collects Learning Gradients Only
Local Agents
User Accessibility & Interaction Flexibility

Fig. 3. Proposed system architecture for AI framework

As shown in Fig. 3, the system is divided into three distinct platforms. First, the user-centric platform consisting of various user devices like cellphones, tablets, or desktops. Second, the Edge platform will host a local agent (or LAI), which will improve the quality of service and preserve the privacy of user data. These local agents will learn from the data collected through the interactions with the users. A user will have the flexibility of interacting in any way that they want - either through text chats, audio interactions, or audio-video interactions.

Depending on the user's choice, a suitable local agent will be deployed ad-hoc to their Edge platform. The Cloud platform will host the global agent (or GAI), whose objective will be to synchronize the activities between different local AI agents hosted on various Edge Platforms. The learning on the Cloud platform will be done by receiving the transmitted learning gradients from the local agents instead of actual user data thereby improving user privacy and incorporating the concepts of federated machine learning. Given the decentralized and distributed nature of the proposed framework, it can then be dispensed as-a-service on an ad-hoc basis with applications similar to IBM Watson's Personality Insights [10] service but more geared toward diagnosing depression.

5 Conclusion

In this paper, a comprehensive survey was presented for AI-enabled approaches proposed in the literature for diagnosing mental illnesses like depression. The survey was carried out by categorizing the various studies by their incorporated methodology like Expert Systems, Fuzzy Logic, or Machine Learning ((un-)supervised, semi-supervised, reinforcement). The paper also summarized the survey by presenting some of the existing challenges in this research domain, the scope of future work, and a design schema of an AI system that can address some of the required computational design requirements for this research area along the lines of accessibility, scalability, privacy, and quality of service.

References

1. Alhanai, T., Ghassemi, M., Glass, J.: Detecting depression with audio/text sequence modeling of interviews. Interspeech, September 2018
2. Alshawwa, I.A., Elkahlout, M., El-Mashharawi, H.Q., Abu-Naser, S.S.: An expert system for depression diagnosis. Int. J. Acad. Health Med. Res. (IJAHMR) **3**, 20–27 (2019)
3. Anxiety and Depression Association of America: Understand anxiety and depression, facts and statistics. https://adaa.org/understanding-anxiety/facts-statistics. Accessed 10 Sept 2021
4. Beheshti, A., Moraveji-Hashemi, V., Yakhchi, S., Motahari-Nezhad, H.R., Ghafari, S.M., Yang, J.: personality2vec: enabling the analysis of behavioral disorders in social networks. In: The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM 2020) (2020). https://doi.org/10.1145/3336191.3371865
5. Beheshti, A., Yakhchi, S., Mousaeirad, S., Ghafari, S.M., Goluguri, S.R., Edrisi, M.A.: Towards cognitive recommender systems. Algorithms **13**(8), 176 (2020). https://doi.org/10.3390/a13080176. http://www.mdpi.com/journal/algorithms
6. Brown, V.M., et al.: Reinforcement learning disruptions in individuals with depression and sensitivity to symptom change following cognitive behavioral therapy. JAMA Psychiatry **78**(10), 1113–1122 (2021)
7. Cupkova, D., Kajati, E., Mocnej, J., Papcun, P., Koziorek, J., Zolotova, I.: Intelligent human-centric lighting for mental wellbeing improvement. Int. J. Distrib. Sens. Netw. **15**(9), 1550147719875878 (2019)

8. Dosovitsky, G., Pineda, B.S., Jacobson, N.C., Chang, C., Escoredo, M., Bunge, E.L.: Artificial intelligence chatbot for depression: descriptive study of usage. JMIR Form. Res. **4**(11), e17065 (2020). https://doi.org/10.2196/17065
9. Guestrin, C., Singh, S., Ribeiro, M.T.: Why should i trust you? Explaining the predictions of any classifier. ACM, August 2016
10. IBM Watson Labs. https://cloud.ibm.com/docs/personality-insights?topic=personality-insights-about. Accessed 26 Aug 2021
11. Inkster, B., Sarda, S., Subramanian, V.: An empathy-driven, conversational artificial intelligence agent(Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth **6**(11), e12106 (2018)
12. Jacobson, N.C., Nemesure, M.D.: Using artificial intelligence to predict change in depression and anxiety symptoms in a digital intervention: evidence from a transdiagnostic randomized controlled trial. Psychiatry Res. **295**, 113618 (2021). https://doi.org/10.1016/j.psychres.2020.113618
13. Johansson, R., Andersson, G.: Internet-based psychological treatments for depression. Expert Rev. Neurother. **12**(7), 861–870 (2012). https://doi.org/10.1586/ern.12.63
14. Kaiser Family Foundation: Mental health care health professional shortage areas (HPSAs). https://www.kff.org/other/state-indicator/mental-health-care-health-professional-shortage-areas. Accessed 10 Sept 2021
15. Khedkar, S., Subramanian, V., Shinde, G., Gandhi, P.: Explainable AI in healthcare. In: ICAST (2019)
16. Lundberg, S.M., Fischer, A., Holt-Gosselin, B., and L.W.: A unified approach to interpreting model predictions. In: NIPS, November 2017
17. Mental Health America: 2021: Covid-19 and mental health: A growing crisis. https://mhanational.org/sites/default/files/Spotlight2021-COVID-19andMentalHealth.pdf. Accessed 16 Feb 2021
18. Mohammadi Motlagh, H.A., Minaei Bidgoli, B., Parvizi Fard, A.A.: Design and implementation of a web-based fuzzy expert system for diagnosing depressive disorder. Appl. Intell. **48**(5), 1302–1313 (2017). https://doi.org/10.1007/s10489-017-1068-z
19. Nemesure, M.D., Heinz, M.V., Huang, R., Jacobson, N.C.: Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci. Rep. **11**, 1980 (2021). https://doi.org/10.1038/s41598-021-81368-4
20. Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., Wilson, J.: The what-if tool: interactive probing of machine learning models. IEEE Trans. Vis. Comput. Graph. **26**(1), 56–65 (2019)
21. Yang, Z., Chen, C., Li, H., Yao, L., Zhao, X.: Unsupervised classifications of depression levels based on machine learning algorithms perform well as compared to traditional norm-based classifications. Front. Psychiatry **11**, 45 (2020). https://doi.org/10.3389/fpsyt.2020.00045
22. Yazdavar, A.H., et al.: Semi-supervised approach to monitoring clinical depressive symptoms in social media. In: IEEE/ACM Advances in Social Networks Analysis and Mining, July 2017
23. Yu, J., Chiu, C., Wang, Y., Dzubur, E., Lu, W., Hoffman, J.: A machine learning approach to passively informed prediction of mental health risk in people with diabetes: retrospective case-control analysis. JIMIR (2021)
24. Zainab, R., Chandramouli, R.: Detecting and explaining depression in social media text with machine learning. In: KDD 2020, August 2020