Applied Intelligence (2024) 54:1619–1636
https://doi.org/10.1007/s10489-023-05095-1

Exploring the potential of federated learning in mental health research: a systematic literature review

Samar Samir Khalil[1,2] · **Noha S. Tawfik**[1] · **Marco Spruit**[2,3]

Accepted: 8 October 2023 / Published online: 12 January 2024

Abstract
The rapid advancement of technology has created new opportunities to improve the accuracy and efficiency of medical diagnoses, treatments, and overall patient care in several medical domains, including mental health. One promising novel approach is federated learning, a machine learning approach that allows multiple devices to train a shared model without exchanging raw data. Instead of centralizing the data in one location, each device or machine holds a portion of the data and collaborates with other devices to update the shared model. In this way, federated learning enables training on more extensive and diverse datasets than would be possible with centralized training while preserving the privacy and security of individual data. In the mental health domain, federated learning has the potential to improve mental disorders' detection, diagnosis, and treatment. By pooling data from multiple sources while maintaining patient privacy by keeping data secure and ensuring that they are not used for unauthorized purposes. This literature survey reviews recent studies that have exploited federated learning in the psychiatric domain, covering multiple data resources and different machine-learning techniques. Furthermore, we formulate the gap in the current methodologies and propose new research directions.

Keywords Federated learning · Mental illness · Psychiatry · Machine learning

1 Introduction

The recent increases in mental health conditions worldwide have made the prevention and treatment of mental disorders a global health priority. Multiple factors have contributed to the dramatic rise in mental illness, such as social media pressure, increased adoption of electronic media (Electronic Screen Syndrome), increased divisive news, increased performance pressures (education, career, financial, etc.), household breakdown, and recently the COVID-19 pandemic [1, 2]. According to the World Health Organization (WHO), depression is the leading cause of disability worldwide, affecting more than 264 million people [3]. Similarly, recently published statistics from the National Institute of Mental Illness (NIMH) in 2020 indicate that an estimated 52.9 million adults aged 18 or older in the United States were diagnosed with a mental illness condition, accounting for 21% of the population [4]. Moreover, suicide is the world's second leading cause of death among people aged 15 to 24 [5]. It is estimated that nearly 800,000 people commit suicide yearly, which translates to one death every 40 seconds. People affected with mental health disorders frequently face significant human rights violations, discrimination, and stigma.

Mental health promotion is rising as societies become more aware of the risks associated with mental illnesses. Many technology-based applications and techniques surfaced in response to the need for more effective mental health prevention, awareness, patient monitoring, and disease diagnosis. Digital mental health platforms powered by artificial intelligence algorithms are also growing popular for diagnosing and treating various psychiatric disorders.

✉ Samar Samir Khalil
samar@aast.edu; s.s.khalil@liacs.leidenuniv.nl

Noha S. Tawfik
noha.abdelsalam@aast.edu

Marco Spruit
m.r.spruit@liacs.leidenuniv.nl

[1] Computer Engineering Department, Arab Academy for Science, Technology and Maritime Transport, 1029 Alexandria, Egypt

[2] Leiden Institute of Advanced Computer Science, Leiden University, 2332CA Leiden, The Netherlands

[3] Public Health & Primary Care, Leiden University Medical Center, 2333CA Leiden, The Netherlands

The main challenge encountered by almost all intelligence (AI)-powered algorithms is the lack of data in general and good quality of data in specific. As sensitive and private as any health problem, data have many privacy policies, problematic data sharing privacy, and ethical constraints. This is specifically relevant in the mental health domain, where patient information is deeply personal and sensitive due to the highly stigmatized nature of the patient's illness. Even when publicly available datasets are published, they are usually limited in size, limiting their performance of current techniques. The data availability will always limit the potential of both machine learning (ML) or deep learning (DL) models if trained in the traditional way, referred to as centralized learning. In centralized learning, only one model is trained on data collected from different sources and compiled together into one dataset. The model is then tested and deployed in a computer program or a web/mobile application to be used by psychiatrists to support their decision-making processes. There is a need for another robust approach that serves clinical psychiatric practice (evaluate, diagnose, and build decision-support systems for patients with mental disorders) but also prioritizes the privacy of the patients and their data collected from multiple sources such as hospitals, clinics, wearable devices, and even social media.

A collaborative learning approach referred to as federated learning (FL) was introduced in 2016 by a team at Google Research [6]. FL follows a client-server approach that trains a centralized model on decentralized data so the data never leave the client side. While federated learning was initially designed for other domains, it quickly gained attention in the healthcare and medical fields through its capability to handle data privacy and governance by training models collaboratively without exchanging data. It provides a consensus solution without moving patient data beyond the firewalls of the healthcare institution in which they reside [7].

Several systematic reviews addressed the use of federated learning in the health domain [8–12]. However, to the best of our knowledge, none of them specifically investigated the use of FL techniques in the mental health domain. This systematic literature review (SLR) aims to bridge the gap and provides in-depth background on using federated learning (FL) and its current state-of-the-art techniques applied in the mental health field. The main research question of this work therefore is:

MRQ: To what extent has federated learning been exploited in mental health state detection?

Through our systematic reviews, we also answer the following sub-research questions.

RQ1: What mental disorders were explored?

RQ2: What data types were most used with FL and mental illness?

RQ3: What countries contributed in this direction?

RQ4: What is the most commonly used FL algorithm in the context of mental illness?

The rest of the paper is organized as follows: Section 2 provides a background on the federated learning paradigm. A detailed description of the systematic review methodology employed is given in Section 3. The findings gathered from the selected papers are highlighted in Section 4. Section 5 provides a list of the challenges and limitations that researchers are currently facing in this area. Section 6 concludes with a critical discussion and suggestions to pave the way for developing FL-based applications and systems in mental health.

2 Background

Privacy-enhancing technologies (PET) aim to prevent data leaks while balancing privacy and usability. Federated learning is one of the PETs with the primary concept of protecting the privacy of clients' data. The more clients guarantee the security of their data, the more available data the model is trained on, and the more generalized the model can be. Unlike traditional centralized model training, where data are brought to the server stored on one machine or a data center, models are sent to clients' end to be trained on their on-device data. Discussed below are the key factors to having an FL system.

2.1 Aggregation methodologies

A preliminary step of a federated learning system is to aggregate the results of each client's model to realize a more powerful generalized model. This step is done by a coordinating centralized server responsible for any client communications. The first aggregation algorithm was titled FederatedAveraging (FedAvg) [6]. In FedAvg, the coordinating server first sends an identical initial model W_i to each of the participating clients K. Each client trains the model locally on their data for a predetermined number of epochs using the stochastic gradient descent (SGD) optimizer. The encrypted trained model results (weights and parameters) W_{i+1}^k are sent back to the coordinating server, which calculates the new updated model W_{i+1} to be shared once more till the learning phase ends. The server updates the model weights by averaging each model's results based on their share of data (weighted average), as shown in Fig. 1. To ensure privacy, a secure aggregation protocol was developed, allowing the server to decrypt the average update only if a predetermined number of users have participated and sent their results [13]. Sharing model parameters throughout

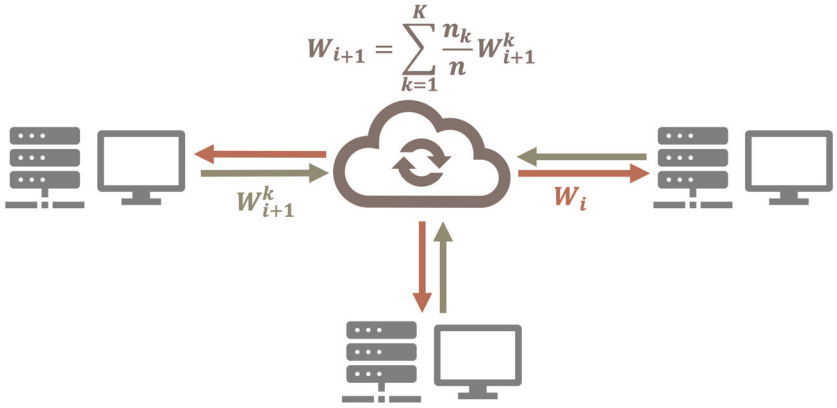


Fig. 1 FedAvg algorithm explanation where W_i is the model shared by the server, W_{i+1}^k is client k update on the shared model, n_k is client k's local data size, n is the total data size and W_{i+1} is the updated model calculated at the server

the network requires ensuring secure communication among clients and the centralized server to avoid problems such as model poisoning. Various techniques, including homomorphic encryption (HE) [14], secure multi-party computation (SMPC) [15], and differential privacy (DP) [16] have been used to compute the defined FL functionality privately.

2.2 Data distribution

Data can have two possible distributions in a federated learning system: independent and identically distributed (IID) and non-IID. Given the nature of the non-IID data in a federated learning system, many statistical challenges can be encountered, such as [10]:

- **Quantity Skew.** Quantity skew is when the class distribution among clients is unequal, referred to as imbalanced data. Imbalanced data are when a client holds far more data records about one class than the others.
- **Label Skew.** Label skew is when data sizes fluctuate among clients when one client holds more data records about a certain class than others. For example, a big hospital has much more data about depression than a small medical center.
- **Feature Skew.** Feature skew is when clients do not have the same set of features. For example, when two hospitals report data about a certain disease, there will be a huge overlap in the reported features owing to the nature of the disease itself. However, some features may differ as the machines used to conduct the results, such as MRI scanners, may not come from the same manufacturer.

Data are considered IID when they are balanced, label distributions are nearly the same at each client, and all clients have the same features. Luckily, challenges such as quantity skew can be overcome by data augmentation and feature skew by data imputation techniques. Label skew is what federated learning is designed for; it can learn from any data source, no matter how small it is.

2.3 Data partitioning

Data partitioning in federated learning has three different types: Horizontal FL (HFL), Vertical FL (VFL), and Federated Transfer Learning (FTL). The three differ in the data each model gets trained on. In **HFL**, each local dataset used to train each client's model has the same features, i.e., each client gets trained on the same set of features for different patients. In **VFL**, each client has a different feature set of the same patients. For example, two different healthcare facilities can have different data (features and labels) for the same patient. Lastly, in **FTL**, the clients don't share the same feature set or the same patients' profiles. It uses a pre-trained model trained on a similar dataset at one client to solve a different problem for another client. HFL is the data partitioning scheme most frequently explored by researchers.

3 Methods

This research employs the *Systematic review Methodology Blending Active Learning and Snowballing* (SYMBALS) [17]. SYMBALS does not only follow authoritative systematic review guidelines but also combines the existing methods into a quick and accessible technique [18–20]. Its stages are explained in the upcoming subsections.

3.1 Database search

Database searching is at the core of all systematic review methodologies. This step constructs a set of all possible relevant publications from different sources. To ensure comprehensive review coverage, we include six databases in our search: Science Direct, Springer, ACM Digital Library, PubMed Central, IEEE Xplore, and Wiley Online Library. The used search query was:

Table 1 Number of papers returned from each database

Database	# of Papers
Science Direct	112
Springer	148
ACM Digital Library	68
PubMed Central	30
IEEE Xplore	26
Wiley Online Library	34
Total	418

Table 2 Number of papers excluded by each criterion

Exclusion Criteria	# of Papers
E1	16
E2	10
E3	74
E4	30
E5	3
E6	251
Total	384

("federated learning" OR "multi-party computing" OR "multiparty computing") AND ("mental health" OR "psychiatry" OR "psychology")

Since FL was proposed in 2016, no FL-based medical research existed until then. The retrieval time range was from 2016 to July 2023. In Springer, Computer Science was chosen as a discipline, and articles and conference papers were selected from the content type. Books were excluded from Wiley Online Library. The query string returned a total of **418** papers; however, after removing the duplicates based on their titles, the final paper set included 402 papers. Table 1 shows the total number of publications returned by each database.

3.2 Screening using active learning

This is a fundamental step in the SYMBALS methodology as it accelerates the screening process without sacrificing accuracy. Machine learning is applied in the title and abstract screening step to spare researchers from manually labeling papers. SYMBALS uses the ASReview tool [21] to achieve this. This is very important when the original paper set is large; however, since the total number of non-duplicated papers retrieved is 402 papers, we decided to perform this step manually to ensure even more validity of the selected papers. The decision to include or exclude a paper was based on the following criteria:

- Inclusion criteria:
  - I1: The paper must describe the use of an FL technique in training an AI model.
  - I2: The paper must address a mental disorder.
- Exclusion criteria:
  - E1: The paper discussed a technique for securely sharing the training parameters during the FL process.
  - E2: The paper discussed a fully decentralized implementation of federated learning such as blockchain.
  - E3: The paper does not address a mental problem.
  - E4: The paper does not explain the FL algorithm used.
  - E5: Local data are shared with the server even if they were encrypted or sent anonymously.
  - E6: Irrelevant papers inaccurately returned by the query.

E6 had the greatest share of set reduction by excluding 251 papers for varying reasons. E3 excluded about 74 articles, such as those that address mood detection, emotion recognition, stress monitoring, and loneliness detection [22, 23, 23, 24]. E4 had a share of 30 papers as the researchers did not mention the federated learning algorithm or how they dealt with the data for federation settings such as [25]. Only three papers were excluded by E5: [26] and [27] as the data left the clients' side, violating the FL concept. After the active learning phase, 384 papers were excluded to end up with 18 papers ready for the next step. Table 2 shows the number of papers excluded by each criterion.

3.3 Backward snowballing

Unlike other SLR techniques that only rely on active learning in their design, SYMBALS complements the output of the previous step with a backward snowballing step. Snowballing ensures the inclusion of relevant papers that could have been missed because its database was not considered or covered by the search query. From a set of selected papers, a researcher can find additional relevant papers by consulting the list of references of each paper, a process called backward snowballing. Other SLRs employ forward snowballing, in which the citations within the papers are inspected to add more relevant papers. However, the authors of SYMBALS argue that older papers will generally constitute the largest group of relevant papers not yet included. It is more efficient to examine the references rather than citations, based on the observation that databases generally have excellent coverage of recent peer-reviewed research. Because the output of the previous step is relatively small, no extra stopping criterion needs to be defined in the current step. One additional paper

Fig. 2 SYMBALS steps

was added from backward snowballing, increasing the total number of papers to 19.

The three subsequent SYMBALS steps are designed to ensure the quality of the included papers, prepare data extraction sheets, and validate the search results.

3.4 Quality assessment

This is an optional step proposed by SYMBALS for a large number of inclusions. Since all the included papers were manually selected and their number is relatively small, this step was skipped in our systematic review process.

3.5 Data extraction and synthesis

Data extraction was performed to give a numerical analysis of the literature reviewed and to describe some approaches in the following section. The following data were extracted from each paper:

- D1: Title and publication year.
- D2: Mental illness type.
- D3: Data type and dataset description.
- D4: Federated learning algorithm.
- D5: Whether the FL model was implemented or simulated.
- D6: Whether the used model was based on traditional machine learning or deep learning (DL)
- D7: Description of the used AI model.
- D8: Performance measures and their results.

It is worth noting that in D5, an actual implementation of FL means working on different client data, sending the model to be trained on their ends, and aggregating the results. On the other hand, an FL simulation is when models are not sent to be trained on users' devices or when data are coming from the same distribution but are divided locally to mimic the FL flow.

During the data extraction phase, we discovered that three pairs of papers were duplicated in terms of their contribution, i.e., in each pair, both papers describe the same model in two different publications [28–33]. To avoid redundancy, only one paper from each pair was considered, leaving 16 papers to be reviewed.

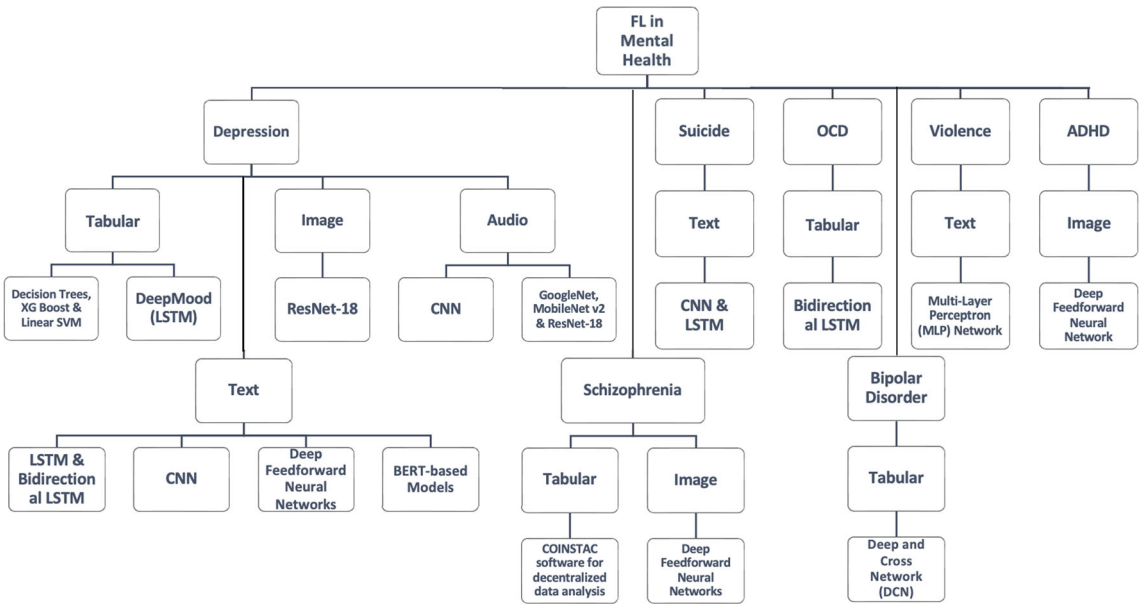
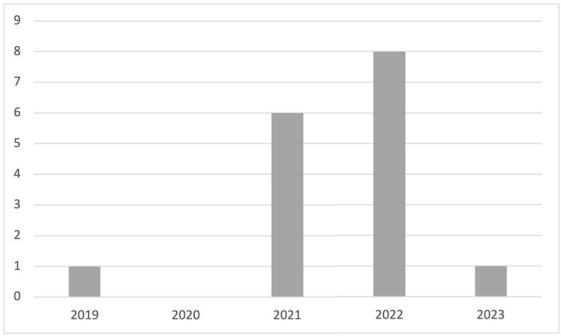
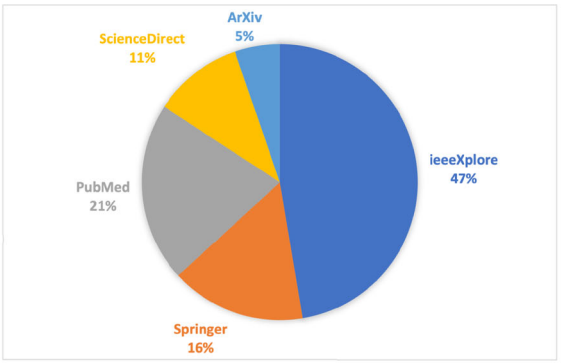


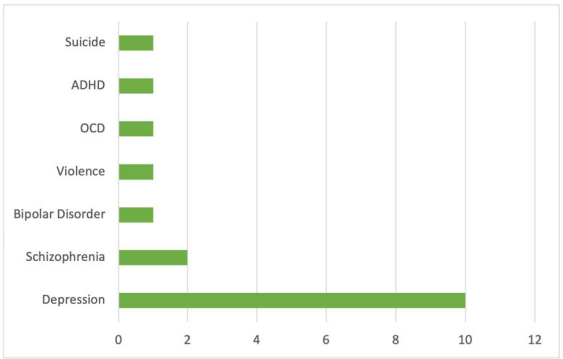
Fig. 3 Visualization of federated learning applications and relevant data types in mental health research



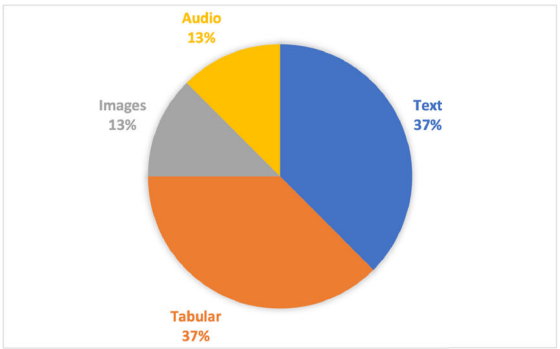
(a) Number of publications across years.



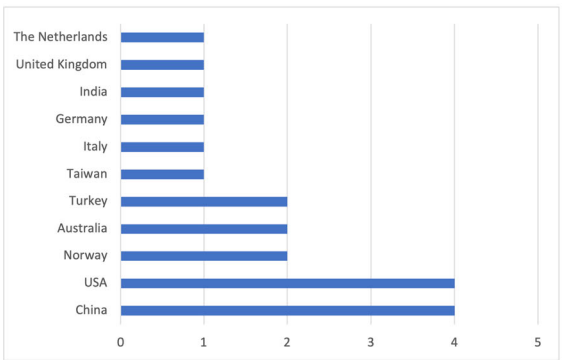
(b) Publishers contribution.



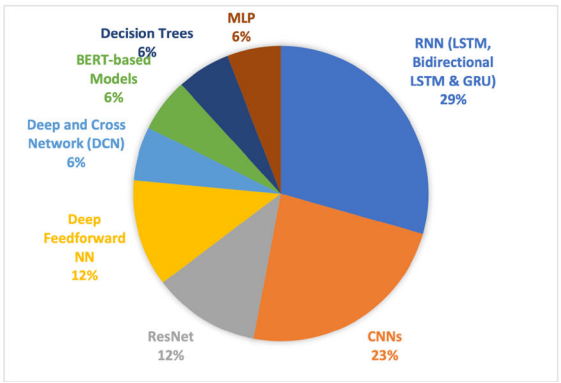
(c) Mental disorders addressed in publications.



(d) Data types used within each publication.



(e) Countries contribution to # of publications.



(f) Machine/Deep learning models used in publications.

Fig. 4 Quantitative analysis

3.6 Validation

This is the last step of the SYMBALS methodology. Its main target is to verify the acquired set of papers. A set of 40 papers resulting from the search query were re-assessed by a different author who did not contribute to the screening process. After viewing the inclusion and exclusion criteria, the author made the same decisions and ended up with the same labeling results as the original author.

A visual representation of the SYMBALS review process applied in our SLR is shown in Fig. 2.

4 Results

In this section, we provide answers to the research question previously introduced after reading and analyzing the research in the selected paper set. Figure 3 gives a summary of the explored mental disorders, data types, and techniques applied in the published research that used FL. Important research insights and quantitative analysis of the reviewed literature are introduced first. A detailed description of each paper in the final selected set is given afterward.

4.1 Quantitative analysis

MRQ: to what extent has federated learning been exploited in mental health state detection? Based on our systematic review and after applying SYMBALS to conduct this SLR, sixteen papers applied the federated learning concept in the mental health domain. While the concept of FL was introduced in 2016, the first published research merging FL and mental health applications appeared in 2019. Since then, there has been an increased rate of publications, specifically in the recent two years, as shown in Fig. 4a. Figure 4b shows the distribution of papers included in this review among the different search engines. Most papers were found in IEEE Xplore. Only four papers were found in PubMed, a medical literature repository; this indicates a lack of exposure to using FL in the mental health domain.

RQ1: what mental disorders were explored? Seven mental disorders were covered: depression, schizophrenia, violence incidents detection, suicidal ideation, obsessive-compulsive disorder (OCD), bipolar disorder, and attention deficit hyperactivity disorder (ADHD), leaving space for many other illnesses to benefit from FL. As illustrated in Fig. 4c, most of the research (10 publications) targeted depression, whereas schizophrenia was the second most addressed mental disorder.

RQ2: what data types were most used with FL and mental illness? All the medical data with the diversity in terms of their type: textual data such as electronic health records, tabular data such as patient information (e.g. age, gender, and sensor readings), images such as scans of patients including ultrasound, CT, MRI, and audio data such as patients' recordings. Textual and tabular sensor data were equally used in the reviewed papers as observed in Fig. 4d. This outcome is not surprising as the nature of mental illness and the spread of social networks made a huge pool of textual data for researchers to work on. Also, tabular sensor data can be obtained from various sources such as smartphones, wristbands, and wearable devices. It is important to emphasize that most of the datasets used in the literature were collected by the authors and not made publicly available, such as the clinical data collected from hospitals and social media posts collected from Twitter, Reddit, and Weibo. Table 3 gives details on the publicly available datasets used to experiment with FL in mental health research.

RQ3: what countries contributed in this direction? Researchers from eleven countries explored the federated learning algorithm to develop a more robust generalized model while keeping the privacy of mental health data. Many researchers from different countries showed interest and contributed to such a beneficial application. Figure 4e lists all these countries by considering every author affiliation in the resulting papers. The United States of America and China each have an equal share of four publications.

RQ4: what is the most commonly used FL algorithm in the context of mental illness? FedAvg was used in more than 75% papers addressing mental illness. This is also expected as the FL is still in its infancy, and FedAvg was the first algorithm introduced and most commonly used in other domains. In Fig. 4f, we provide an overview of the underlying machine learning models that papers employed to evaluate their proposed FL framework. As can be seen, recurrent neural networks (RNNs) are the most commonly used models, followed by convolutional neural networks (CNNs). Less explored are Decision Trees, multi-layer perceptron (MLP), Deep and Cross Network (DCN), and BERT-based models.

4.2 Reviews for federated learning in mental health

In this section, we introduce paper-specific details and findings. For a better comparison among the reviewed literature, papers are segmented by the model employed for learning, i.e., traditional machine learning (ML) or deep learning (DL), followed by a subdivision based on the used data type.

4.2.1 Traditional machine learning based classifiers

Four papers employing traditional machine-learning techniques are discussed in this section.

Tabular data

In [39], depression was detected using sensor data collected from the ActiGraph wristband [34]. In each minute, the quantity, duration, and strength of the movements were recorded for each patient. The authors proposed a new data augmentation approach to tackle the imbalance problem in the collected data. For every minute in the day, if a data sample is missing, then a set of data records representing this patient's data at the same time on other days were extracted from the dataset, and a random one was selected to complete the patient's vector for the day. The data are then fed to a Privacy-Preserving Distributed Extremely Randomized Trees (PPD-ERT) [40] algorithm based on decision trees. PPD-ERT guarantees data privacy by making data holders keep their data. A mediator server initializes and shares a global and personal random seed among data holders and calculates the best candidate node at each step from the aggre-

Table 3 Publicly available datasets used in FL research

Data Type	Dataset Name	Dataset Description
Tabular	Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients [34]	This dataset is used for depression detection. Sensor data collected from ActiGraph wristband. In each minute, the quantity, duration, and strength of the movements were recorded for each patient. The dataset contained recordings of 23 unipolar and bipolar-depressed patients and 32 healthy controls.
	The opportunity challenge: A benchmark database for on-body sensor-based activity recognition [35]	This dataset is used for OCD (Obsessive Compulsive Disorder) detection. It contains a large collection of complex, everyday activities recorded in various settings using various sensors. The dataset includes recordings of 12 individuals, each monitored by 15 sensor systems that used 72 sensors of 10 different types. The sensors were integrated into the environment, objects, and the individuals' bodies.
Image	Center for Biomedical Research Excellence [36]	This dataset is used for schizophrenia detection. It consists of 146 MRI scans for 72 schizophrenic patients and 74 controls.
	ADHD-200 Competition [37]	This dataset is used for ADHD (Attention-Deficit/Hyperactivity Disorder) detection. It consists of 939 samples from 358 ADHD patients and 581 controls.
Audio	The distress analysis interview corpus of human and computer interviews [38]	This dataset is designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. It contains clinical interviews with 224 distressed patients and 397 controls.

gated results to ensure the same tree is built at each client. The proposed augmentation approach led to better classifiers with higher performance measures up to 7.9% higher f1-score, 8.2% higher accuracy, and 0.169 higher Matthews correlation coefficient. The authors continued the work on the same model and introduced [41], an extension of the above-explained work.

In [42], the gender differences in negative symptom severity in schizophrenia were studied using the Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation (COINSTAC) software platform. The authors used data collected by the FBIRN (Function Biomedical Informatics Research Network). COINSTAC [43] is open-source software that enables federated or decentralized data analysis by sharing analysis pipelines and communicating partial results, updated models, and other features. R scripts were written to read clinical and demographic data, calculate five-factor and two-factor model scores from each client's SANS (Schedule for the Assessment of Negative Symptoms) data items, and regress these scores against gender. The five-factor model yields scores for avolition, anhedonia, alogia, blunt affect, and asociality. The two-factor model yields scores for Motivation/Apathy (MAP) and Expressiveness (EXP). The MAP is a weighted combination of avolition, anhedonia, and asociality, while the EXP is a weighted combination of alogia and blunted affect. The SANS and gender data were stored in a standardized CSV file at each site. However, the spreadsheets could be located in any directory on

the local system as the user identifies the required files during data mapping. Data were collected from seven different institutions, and a simulation of seven clients was created, yielding the following results: Males had significantly more severe total negative symptoms than females ($P < 0.05$). On closer inspection, however, men with schizophrenia had a higher EXP factor score than women.

Textual data

In [44], violence risk among psychiatric patients from Dutch clinical notes was predicted using natural language processing and federated learning techniques. Each data point corresponded to a patient's admission period and contained the concatenation of clinical notes from up to 28 days, including the first day of admission. The data points were labeled by whether a violent incident occurred or not (positive/negative outcome) over the next 27 days following the first day of admission. The authors used Doc2Vec [45] to extract a 300-dimension feature vector from the input text. The vector was then fed to a feed-forward neural network with one hidden layer of a ReLU activation function and one output layer with one neuron and a sigmoid activation function to classify the output. Four models were trained and compared: two local, one federated, and one data-centralized model. The collected data were split between two institutions, A and B, where each client was trained only on its share of data. In the data-centralized approach, the model was trained on the full dataset. FedAvg was used to aggregate the models trained by the two local clients. The results indicated that the federated model outperformed the local models and performed similarly to the data-centralized model.

4.2.2 Deep learning based classifiers

Twelve papers are included in this section.

Tabular data

In [46], the problem of depression detection was addressed by using data collected by BiAffect, a mobile application with a special input keyboard. Three types of metadata were collected: alphanumeric characters, special characters, and accelerometer values. To ensure users' privacy, only the duration of the keypress, the duration before the last keypress, and the distance from the last key to the coordinate axis on the horizontal and vertical axes were collected instead of the alphanumeric characters themselves. With its three variations, the DeepMood [47] model was used as a classifier, with the data fusion stage being the main variation. The first used a multi-view machine layer, the second used a factorization layer, and the third used a conventional fully connected layer. Five experiments were conducted on each model: (1) Local training where each client had only their share of the data. (2) Traditional centralized training. (3) Federated model using FedAvg algorithm. (4) Institutional Incremental Training (ILL), where each client sent its model to the next one after it completed its training until all had trained once. (5) Cyclic Institutional Incremental Training (CILL) repeated the ILL training process for a predetermined number of cycles.

Two data distribution scenarios were considered: IID and non-IID. The testing accuracy was reported in each experiment. For IID, Multiple clients were considered (4, 8, 12, 16 & 24), and a different number of data points held by each party was also experimented with (100, 500, 1000, 1500, 2000 & 3000). For both IID and non-IID, the FL model achieved the second-highest accuracy in most experiments after the centralized learning model with the trait of preserving data privacy.

In [48], Obsessive-Compulsive Disorder (OCD) was detected using the OPPORTUNITY Dataset for Human Activity Recognition from Wearable, Object, and Ambient Sensors [35]. The authors used readings from the accelerometer and gyroscope sensors only. To simulate the repeated actions done by OCD patients, a specific set of activities with a particular number of repetitions was assigned to each subject. The baseline-designed model is a two-layer bidirectional Long Short-Term Memory with a fully connected output layer and dropout between each layer. For personalization, the last dropout and fully connected layer were trained individually on the local data, whereas the rest of the model was subject to the FedAvg algorithm. Four experiments were designed to test the model's performance: (1) Traditional centralized model training on the full dataset. (2) Local data training, where each client was trained on its local data only. (3) Federated learning using the FedAvg algorithm on a simulation of 4 clients without personalization. Lastly, (4) FL with personalization using three different personalization schemes. The results showed that FL and federated personalized learning outperformed both centralized and local model training.

In [49], Lee et al. used tabular data extracted from electronic health records of five hospitals in South Korea to apply a real-world horizontal federated learning setting that can detect bipolar transitions in patients with depression. The team tackled the federated real-world environment challenges through four stages: standardized feature extraction, federated feature selection, FL, and cross-site evaluation. For standardizing feature extraction, the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) [50] was used as the data format. The patient data in each hospital's electronic medical records were anonymized and standardized using OMOP CDM, and stored safely within each organization. The extract, transform, and load processes for OMOP CDM were done by a trustworthy broker, who ensured that only data from which personal information had already been taken out were used. The authors added a second stage of federated feature selection due to the lack of powerful computational resources at the contributing hospi-

tals. LightGBM [51] was used for this phase as it quickly trains even on CPU and had proven its feasibility on medical tabular datasets. Only features that were present in all of the internal datasets were selected. An early stopping criterion was defined to make the model stop if its performance did not increase by more than 2% in 3 consecutive searches. Only 100 out of 21,042 features were selected to train the FL model. In the third stage, the FL process, the authors used federated averaging for the weight aggregation algorithm. They applied differential private-stochastic gradient descent (DP-SGD) during the local update to ensure differential privacy. Deep and Cross Network (DCN) [52] was used to train data from four hospitals; the last hospital data were kept for validation. The model was trained for five rounds in FL, with each round training for five epochs. Lastly, in the cross-site evaluation stage, the federated and the four local models were compared with each hospital's data on internal and external validation datasets. Mean AUC was used as an evaluation metric. The reported mean AUC of the federation model was 0.726 across all test datasets, while the local models trained with each hospital's local data had mean AUCs of 0.642, 0.662, 0.707, and 0.692, respectively. This indicates that the federated model has higher generalizability than any local model.

Textual data

In [30], posts collected from both Reddit and Twitter were used to address the problem of detecting suicidal ideation on social media. For text classification, the authors trained two local data-preserving deep learning models: CNN and LSTM. A new optimization algorithm called the average difference descent for learning with data protection (AvgDiffLDP) was proposed for aggregating the locally trained models at the centralized server. AvgDiffLDP used the gradient of the average differences between the server's parameters in the previous time stamp and the updated users' parameters in the current time stamp. The updated model parameters were sent to the local users/clients and trained using stochastic gradient descent. The authors conducted three experiments: SimpleLDP, AvgDiffLDP, and centralized NonLDP. The collected data were distributed among users in the first two experiments. In SimpleLDP, they trained separate local data-preserving models for each user on different devices without sharing data or parameters. In AvgDiffLDP, they trained multiple users locally and used the new proposed optimization algorithm. In the centralized NonLDP, they used the entire dataset to train one centralized model on the server. Average testing accuracy and the average area under the receiver operation curve (AUC) were reported. LSTM model results were slightly better than CNN. Even though the centralized model performance was better than the AvgDiffLDP one, the proposed model kept data privacy which is critical when dealing with such sensitive data.

In [53], Italian text sentences from the ANDROIDS project were used to predict depression. The authors trained a Long-Short Term Memory (LSTM) neural network for text classification. Two experiments were conducted; one was centralized, and the other was federated on three simulated clients. FedAvg was used to aggregate each client's trained model parameters. The architecture of the federated model had four layers: An embedding layer, a bidirectional LSTM, an LSTM, and a Dense layer. The categorical cross-entropy was used as a loss function, while Adam's algorithm was adopted to train the model. The testing accuracy was reported at each experiment and showed that the centralized model outperformed the federated one.

In [33], Li et al. proposed a CNN Asynchronous Federated optimization (CAFed) depression detection system. The system adopted a text-based convolutional neural network model (Text-CNN) for detecting depression from Weibo posts. The team collected data from 900 users throughout an entire year. The proposed model consisted of the following layers: (1) Embedding layer where the Weibo vector was formed using one user's data. (2) Convolution layer where various filter sizes were used with ReLU activation function to get the feature maps. (3) Max pooling layer to get the most important features and create the final feature vector. (4) Dropout layer to avoid overfitting. (5) A fully connected layer and an output layer with a sigmoid activation function to classify the output.

The proposed CAFed algorithm followed the same start as the FedAvg, except when updating the model, CAFed updated the global model instantly after receiving the local updates sent by any client. To ensure the model's privacy, Gaussian white noise with a mean of 0 and a variance of 1 was added to the server process to adjust global values and keep each device's contribution hidden. The authors compared the results of CAFed to FedAvg, revealing that CAFed converged faster than FedAvg. FedAvg waited for all ten user devices in the experiment to respond in each epoch, whereas CAFed required one device's response only to proceed to the next epoch. Furthermore, FedAvg had more communications than CAFed in each global epoch. In general, CAFed converged faster than FedAvg for the same communication overhead.

In [54], Ahmed et al. proposed a hyper-graph attention-based federated learning model for detecting depressive symptoms from text collected from patients using the standard PHQ-9 questionnaire. Data were collected from different internet forums and questionnaire websites. There were two approaches to feature extraction used. The first used an emotional lexicon, while the second used a structure-aware graph model. For vectorization, both models used a 300-dimensional glove vector. The embedding method was used to convert text into node vectors in the lexicon of nine symptoms. The structure embedding model then used the

hyper-graph to extract word-based node patterns. The text was then labeled using trained embedding depending on the question. Two models were built to classify the extracted features. The baseline model was a feed-forward neural network with (30, 20, 10) hidden layers with ReLU activation function and the final layer is a 9-link sigmoid function. The other model was a recurrent neural network with long short-term memory (LSTM) units and an attention position layer. When compared directly to the baseline model, the LSTM network achieved a relatively high level of performance. For applying federated learning, a global initial model was sent to six clients, where it was used to train a local model on the part of the dataset. The FedAvg algorithm was used to update the global model parameters. According to the validation loss, each client can choose whether to use the global updated model or the local model's best iteration. The proposed system achieved a 0.86 ROC score.

In [55], Basu et al. used data scraped from Twitter to address the problem of detecting depression and sexual harassment. The team investigated the effects of differential privacy (DP) on training contextualized language BERT-based [56] models in both a centralized and an FL setting. They used four natural language processing (NLP) models: BERT, ALBERT [57], RoBERTa [58], and DistilBERT [59]. Four experiments were carried out: baseline NLP model, DP NLP model, FL NLP model, and FL+DP NLP model. The team tried both IID and non-IID data distributions for the federated learning setting in an HFL data partitioning scheme. The FedAvg algorithm was used to aggregate the simulation of ten clients. The reported results were as follows: When employing differentially private training, it was observed that smaller networks such as ALBERT and DistillBERT exhibit a more gradual degradation compared to larger models like BERT and RoBERTa. Utility degradation was higher in the Non-IID setting for FL, the typical scenario in medical applications, than in the IID arrangement, indicating the necessity for training methods adapted to such setups. Finally, when the size of the training dataset was limited, the impact of differential privacy on utility was more deleterious than when a larger amount of data were available.

Image data

In [60], ResNet-18 was adjusted to detect the patients with depression using their structure brain MRI (3D-T1). Data were collected from 23 different sites, but as they were limited in size, they were partitioned among five clients where the local models were trained. Encrypted gradients from the clients were weighted and aggregated at the centralized server to produce the updated global gradients at the end of each epoch. The updated model was then re-distributed to the clients to proceed with their training. The average accuracy of five-fold cross-validation was reported. The federated models outperformed the local models by 0.2∼4.33% for each of the five groups.

In [61], Federated Multi-Task Learning for Joint Diagnosis (FMTLJD) used MRI scans to diagnose three mental disorders: schizophrenia (SCZ), attention-deficit/hyperactivity disorder (ADHD), and autism. The used data were aggregated from three publicly available databases: Center for Biomedical Research Excellence (COBRE for SCZ) [36], the ADHD-200 Competition (ADHD-200 for ADHD) [37] and the Autism Brain Imaging Data Exchange I (ABIDE for ASD) [62]. The authors proposed a federated contrastive learning-based feature extractor (FCLFE) for feature extraction that used the Pearson Correlation Coefficient (PCC) to calculate brain functional connective features. A Gaussian noise augmentation step was added to reduce the risk of overfitting. The augmentation output was fed into a multi-layer perceptron (MLP) network with non-linear transformation to extract the higher level of abstraction representation. To train the extracted features of each dataset, a federated multi-gate mixture of expert classifiers (FMMoE) was proposed. Expert networks and gated networks made up the classifier. Given multiple task inputs, the expert network, built using group stacking of neural networks, learned the various feature representations. The gated networks learned to obtain an optimal mixture pattern by assembling these expert networks with different learned weights. An MLP was constructed from each task's MMoE output and acted as a tower network to refine the task-specific representation and make predictions. To simulate the federated learning process, the data were divided among four clients, and FedAvg was used to aggregate the local models and update the shared one. Modifying the minibatch SGD optimization process, differentially private stochastic gradient descent(DP-SGD) [63] was used on private local datasets of client models to ensure the privacy of distributed data processing systems. Four scenarios were created to evaluate the performance of the proposed model: non-federated (centralized) mode and federated mode with multi-task learning and without. The results were not expected as the centralized MTLJD model outperformed the federated one, but the FMTLJD model performance exceeded the centralized model in ABIDE and ADHD-200 databases. This result also demonstrated that, besides lowering the risk of privacy leakage, FMTLJD enabled a reliable diagnostic detection that was competitive with the ideal scenario of gathering all multi-site data for training.

Audio data

In [64], English audio recordings from clinical interviews were used for depression detection. The used data were available online through DAIC-WOZ dataset [38, 65]. A convolutional neural network (CNN) model was proposed to classify the extracted audio features. The authors used the Mel Frequency Cepstral Coefficients (MFCCs) feature and generated 13-dimensional MFCCs from each speech segment by using 26 filters from the Mel filter bank with a window size of 25ms and a step size of 10ms. All MFCC

coefficients were normalized to prevent training from being hampered by their wide variation. The proposed CNN model consisted of 3 convolution layers of 32, 64, and 128 filters and size 3x3. A ReLU activation function followed each convolution layer. A max-pooling layer of size 2x2 was used to reduce the dimensionality of the output feature maps. The output feature was then routed to two fully connected layers with 64 and 32 hidden units, respectively, before being followed by a dropout layer (the dropout rate is set to 0:1). The ReLu function activated each fully connected layer. Finally, a neuron with Sigmoid activation was used to predict whether a person was depressed. An SGD optimizer was used to train the model with binary cross-entropy loss. Three experiments were designed to compare the performance of FedAvg to the baseline centralized one. Centralized learning achieved the best results among the three, with 96.8%, 93.7%, and 92.3% for accuracy, precision, and recall, respectively. The two FL approaches, IID and Non-IID, were trained multiple times, each on a different number of clients (8, 56, and 189). In the IID scenario, results showed that the more clients contributed to the learning process, the lower the model accuracy as the amount of data each held decreased. The non-IID scenario produced lower results than the centralized and IID scenarios. Such performance degradation was expected because data heterogeneity across clients caused computed local model updates to drift in different directions, resulting in suboptimal server updates. A significant number of clients with a more distinct client distribution may make global model convergence more difficult.

Suhas et al. [28] also addressed the problem of depression detection using speech analysis. They used a subset of the clinical audio recordings available online through the DAIC-WOZ dataset [38] to ensure balanced data distribution among classes and genders. Two classification tasks were considered: depression detection and depression severity. The scipy.signal.spectrogram function was used to extract log spectrogram features from an overlapping window with a duration of 1s and a shift of 0.1s. The spectrogram images aided in modeling both the temporal and harmonic structures of audio signals, resulting in better classification performance than existing methods. GoogleNet, MobileNetV2 and ResNet-18 were used to classify the input spectrograms utilizing the concept of transfer learning. Three scenarios were designed for model training: one data-centralized and two federated learning frameworks using the FedAvg algorithm and federated matched averaging (FedMA) [66]. FedMA was designed for modern neural network architecture, such as CNN and LSTM. It updated the global model parameters layer-wise by matching and averaging hidden elements (filters for CNN and neurons for deep feed-forward networks) with similar feature extraction signatures. Five-fold cross-validation accuracy was reported to compare the performance of the models. Across folds, the centralized approach outperformed the federated methods by 6-10%. The centralized approach had the best average five-fold accuracy of 0.934, while the federated scheme had 0.91. The centralized approach was approximately 1.55-2.19x faster than the federated schemes, with ResNet-18 being the fastest for both the centralized (155s) and federated (327 & 340s, respectively). Compared to a centralized approach, the FL models outperformed previous work using the same dataset and allowed for a robust assessment of depression with only a 4-6% accuracy loss. TensorFlow Lite was used for developing the mobile application. The app determined whether or not the speech contained depression symptoms and, if so, how severe they were. FL models were energy-efficient, with low inference latency and a small memory footprint.

5 Challenges and limitations

Applying federated learning in mental health has a number of challenges that can sometimes lead to limitations. In this section, we discuss some limitations that were observed by analyzing the reviewed literature. Deploying a real-world FL setting faces the following challenges:

- **Privacy Leakage and Patient Consent.** In the real world, FL necessitates using personal health data that require regulatory compliance and user acceptance. The latter will not be achieved unless patients have complete confidence that their privacy will be protected through a federated learning application. Not all the reviewed papers considered using a privacy-preserving algorithm such as differential privacy to secure their models [44, 48, 60].
- **Data Heterogeneity.** When working with data collected from different sources, it's common to encounter inconsistencies or discrepancies in the types of data fields available. The variance in the types of data collected across different resources limits the model's ability in terms of the training process. In such cases, models usually rely on the overlapping data across sources, leaving out some important information that could help better identify, diagnose, or treat the mental disorder [49].
- **Data unification.** The nature of clinical data necessitates the creation of a unified process when gathering data from various sources. This ensures a coherent view, so it can be utilized more efficiently and effectively. This process requires time and resources and hence complicates and slows down the research and also limits its transparency, interoperability, reproducibility, and scalability.

Table 4 Summary of the reviewed articles

#	Title	Publication Year	Mental Illness	Data Type	Model	FL Algorithm
1	Monitoring Motor Activity Data [39]	2021	Depression	Tabular Sensor Data collected from ActiGraph wristband	Traditional ML Extremely Randomized Trees, XGBoost and Linear SVM	PPDERT [41]
2	ENIGMA+COINSTAC [42]	2022	Schizophrenia	Tabular	COINSTAC software for decentralized data analysis	
3	Federated learning for violence incident prediction [44]	2022	Violence	Text Dutch Clinical notes	Traditional ML Multi-Layer Perceptron (MLP) Network	FedAvg
4	Privacy-Preserving Federated Depression Detection [46]	2021	Depression	Tabular collected by BiAffect mobile application	DL DeepMood(LSTM) [47]	FedAvg
5	Sensor-Based Obsessive Compulsive Disorder Detection [48]	2021	OCD	Tabular Sensor Data	DL Bidirectional LSTM	FedAvg
6	Privacy-Preserving Federated Model Predicting Bipolar Transition [49]	2023	Bipolar Disorder	Tabular Extracted from Electronic Health Records	DL Deep and Cross Network	FedAvg
7	Detecting Suicidal Ideation [30]	2019	Suicidal Ideation	Text English Reddit & Twitter posts	DL CNN and LSTM	AvgDiffLDP

Table 4 continued

#	Title	Publication Year	Mental Illness	Data Type	Model	FL Algorithm
8	Evaluating Efficiency and Effectiveness [53]	2021	Depression	Text Italian ANDROIDS Project	DL LSTM	FedAvg
9	Intelligent depression detection [33]	2021	Depression	Text Chinese Weibo posts	DL CNN	CAFed FedAvg
10	Hyper-Graph Attention Based Federated Learning Method [54]	2022	Depression	Text English Online forums	DL Deep feed-forward neural network Bidirectional LSTM	FedAvg
11	Benchmarking Differential Privacy [55]	2022	Depression	Text English Twitter	DL BERT-based models	FedAvg
12	Federated Learning on Structural Brain MRI [60]	2021	Depression	3D Images MRI scans	DL ResNet-18	FedAvg
13	Federated Multi-Task Learning [61]	2022	Schizophrenia ADHD	Images MRI scans	DL Deep feed-forward neural network	FedAvg
14	Privacy-preserving Speech-based Depression Diagnosis [64]	2022	Depression	Audio English Clinical interviews DAIC-WOZ Dataset	DL CNN	FedAvg
15	Privacy Sensitive Speech Analysis [28]	2022	Depression	Audio English Clinical interviews DAIC-WOZ Dataset	DL GoogleNet, MobileNet v2 and ResNet-18	FedMA FedAvg

- **Computational Power.** Hospitals/psychiatric clinics do not always have the powerful computational resources needed for the AI model local training step. The speed of training will be limited by the slowest resource that sends its local update. This was one of [49] limitations as they had to use the available hospitals' CPUs.
- **Communication Overhead and Network Stability** Sharing the model between the centralized server and multiple clients for numerous FL rounds results in communication overhead and hence creates a bottleneck for the system. It also needs a stable, secure network connection available for users to upload their updated, locally trained models.

6 Conclusion and discussion

This systematic review highlighted the previous attempts to use federated learning with mental health applications. It followed the SYMBALS methodology to conduct the SLR and answer the main and sub-research questions. Table 4 summarizes the sixteen papers that were selected for review after applying all the inclusion and exclusion criteria.

Besides answering the research questions and providing quantitative analysis, this research explained in detail each included paper in terms of the used learning model, whether a traditional machine learning model or a deep learning model, feature extraction methods, the used federated learning algorithm, data type, and distribution among several clients in the simulation or real-world environment and the addressed mental disorder. Our findings indicate that the provided research shows high potential, but a considerable gap still needs to be filled through the coming research directions.

The first observation is that a relatively low number of published research is found online in this specific research direction. This is surprising, given that mental health is one of the best fields to benefit from the privacy-preserving trait offered by federated learning. FL has been widely adopted for almost five years. Researchers are encouraged to explore and conduct more research in this area.

Secondly, only one of the published papers applied a real-world federated learning scenario where models are sent back to users' devices and trained on their local data. Until now, FL has predominantly been employed in simulated environments only. Technology Readiness Level (TRL) assessment is used to fairly assess the proposed systems. TLR is a widely accepted metrics-based process that evaluates the maturity of technologies under development. It rates technologies on a scale from 1 to 9, where 1 is the lowest level of readiness, and 9 indicates that the actual application of the technology is implemented and in its final form. We map the reviewed papers according to the guidelines and constraints of TRL. Only one paper scored between 7 and 9 on the TRL scale; the rest of the papers received a score between 4 and 6, prototype level, as none of them was demonstrated in an actual operational environment. The main difference among papers was how the simulation was conducted and its corresponding experimental settings; also it differs in the number of clients in each case.

Thirdly, none of the research in the domain of mental health explores vertical FL or Federated Transfer Learning. Current research focuses on horizontal federated learning where each local dataset used to train each client's model has the same features, i.e., each client gets trained on the same features' set for different patients. In some of the reviewed literature, the initial model seed shared by the centralized server to the clients was a pre-trained deep learning model where transfer learning is used. However, the learning setting does not follow the definition of FTL mentioned in Section 2 where one client transfers its trained model to another to fine-tune it to address a similar problem. Rather, it follows horizontal partitioning where all the clients contribute to training one model that addresses one problem with datasets having the same features. From the performance levels and findings of the papers, there is still room for exploring potential enhancements using other FL techniques, i.e., VFL and FTL. Both approaches could benefit mental health applications as each patient could have varying symptoms (features), given a robust global model. FTL can particularly useful when one client has a relatively small set of labeled data that no model can generalize well enough by training solely on the small set. In the FTL, such a small dataset client can exploit the model trained at another client with a larger, somehow similar dataset. On the other hand, VFL is important as it enhances the characterization of samples by incorporating features from different sources to boost the model's capabilities. More research should be addressing the validation of the efficiency of FL with this type of sensitive data.

Federated learning, in general, remains an emerging area of research. The reviewed literature shows a huge potential for the use of FL, specifically for mental health applications across different types of data. For future directions, researchers are encouraged to develop new machine learning and deep learning techniques that follow the FL approach with better efficiency and accuracy, as there is still a huge room for improvement in real-world settings. Similarly, exploring the potential of the different federated learning types. Future research should focus on bridging the gap by deploying robust privacy-preserving algorithms, creating a unified system for data collection from different insti-

tutions, and ensuring that the participating hospitals have wireless resources for networking and powerful hardware. Such improvement in models' performance while preserving patient privacy can be the key to increased accessibility of personalized mental health care.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations

Competing of interest The authors have no competing interests to declare relevant to this article's content.

References

1. Benisek A (2022) Covid-19 and Depression. https://www.webmd.com/lung/covid-19-depression. Accessed 07 Oct 2021
2. Folk J (2022) Why is mental illness on the rise. https://www.anxietycentre.com/faq/why-is-mental-illness-on-the-rise/. Accessed 27 March 2021
3. WHO (2022) Depression. https://www.who.int/news-room/fact-sheets/detail/depression. Accessed 31 March 2023
4. Folk J (2022) Mental illness. https://www.nimh.nih.gov/health/statistics/mental-illness. Accessed March 2023
5. SAVE (2022) Suicide statistics. https://save.org/about-suicide/suicide-statistics/ Accessed 06 Dec 2022
6. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics pp 1273–1282. PMLR
7. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K et al (2020) The future of digital health with federated learning. NPJ digital medicine 3(1):1–7
8. Dasaradharami Reddy K, Gadekallu TR et al (2023) A comprehensive survey on federated learning techniques for healthcare informatics. Comput Intell Neurosci, vol 2023
9. Pfitzner B, Steckhan N, Arnrich B (2021) Federated learning in a medical context: a systematic literature review. ACM Trans Intell Syst Technol (TOIT) 21(2):1–31
10. Shyu C-R, Putra KT, Chen H-C, Tsai Y-Y, Hossain KT, Jiang W, Shae Z-Y (2021) A systematic review of federated learning in the healthcare area: from the perspective of data properties and applications. Appl Sci 11(23):11191
11. Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B (2022) Federated learning for healthcare: systematic review and architecture proposal. ACM Trans Intell Syst Technol (TIST) 13(4):1–23
12. Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RT, Jochems A, Miraglio B, Townend D, Lambin P (2020) Systematic review of privacy-preserving distributed machine learning from federated databases in health care. JCO Clin Cancer Inform 4:184–200
13. Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K (2017) Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security pp 1175–1191
14. Gentry C (2009) Fully homomorphic encryption using ideal lattices. In: Proceedings of the forty-first annual ACM symposium on theory of computing pp 169–178
15. Goldreich O (1998) Secure multi-party computation. Manuscript. Preliminary Version vol 78(110)
16. Dwork C (2008) Differential privacy: a survey of results. In: International conference on theory and applications of models of computation, pp 1–19. Springer
17. van Haastrecht M, Sarhan I, Yigit Ozkan B, Brinkhuis M, Spruit M (2021) Symbals: a systematic review methodology blending active learning and snowballing. Front Res Metr Anal 6:33
18. Keele S et al (2007) Guidelines for performing systematic literature reviews in software engineering. Technical report Technical report ver. 2.3 ebse technical report. ebse
19. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D (2009) The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 62(10):1–34
20. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA (2015) Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. Syst Rev 4(1):1–9
21. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, Kramer B, Huijts M, Hoogerwerf M, Ferdinands G et al (2021) An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 3(2):125–133
22. Chhikara P, Singh P, Tekchandani R, Kumar N, Guizani M (2020) Federated learning meets human emotions: a decentralized framework for human-computer interaction for iot applications. IEEE Internet Things J 8(8):6949–6962
23. Qirtas MM, Pesch D, Zafeiridi E, White EB (2022) Privacy preserving loneliness detection: a federated learning approach. In: 2022 IEEE international conference on digital health (ICDH) pp 157–162. IEEE
24. Meerza SIA, Li Z, Liu L, Zhang J, Liu J (2022) Fair and privacy-preserving alzheimer's disease diagnosis based on spontaneous speech analysis via federated learning. In: 2022 44th annual international conference of the IEEE engineering in medicine & biology society (EMBC) pp 1362–1365. IEEE
25. Salam MA, Badr E, Monier E, Mohamed A (2022) Schizophrenia diagnosis using optimized federated learning models. IJCSNS, vol 829
26. Pranto MAM, Al Asad N (2021) A comprehensive model to monitor mental health based on federated learning and deep learning. In: 2021 IEEE international conference on signal processing information communication & systems (SPICSCON), pp 18–21.IEEE

27. Suruliraj B, Orji R (2022) Federated learning framework for mobile sensing apps in mental health. In: 2022 IEEE 10th international conference on serious games and applications for health (SeGAH) pp 1–7. IEEE
28. Suhas B, Abdullah S (2022) Privacy sensitive speech analysis using federated learning to assess depression. In: 47th IEEE international conference on acoustics speech and signal processing ICASSP 2022 pp 6272–6276. Institute of Electrical and Electronics Engineers Inc
29. Bettapalli Nagaraj S (2021) Privacy-preserving assessment of depression using speech signal processing
30. Ji S, Long G, Pan S, Zhu T, Jiang J, Wang S (2019) Detecting suicidal ideation with data protection in online communities. In: International conference on database systems for advanced applications, pp 225–229. Springer
31. Ji S, Long G, Pan S, Zhu T, Jiang J, Wang S, Li X (2019) Knowledge transferring via model aggregation for online social care. arXiv:1905.07665
32. Li J, Zhang R, Cen M, Wang X, Jiang M (2021) Depression detection using asynchronous federated optimization. In: 2021 IEEE 20th international conference on trust security and privacy in computing and communications (TrustCom) pp 758–765. IEEE
33. Li J, Jiang M, Qin Y, Zhang R, Ling SH (2022) Intelligent depression detection with asynchronous federated optimization. Complex Intell Syst, pp 1–17
34. Garcia-Ceja E, Riegler M, Jakobsen P, Tørresen J, Nordgreen T, Oedegaard KJ, Fasmer OB (2018) Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. In: Proceedings of the 9th ACM multimedia systems conference pp 472–477
35. Chavarriaga R, Sagha H, Calatroni A, Digumarti ST, Tröster G, Millán JdR, Roggen D (2013) The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. Pattern Recognit Lett 34(15):2033–2042
36. Network TMR (2023) COBRE. https://figshare.com/articles/dataset/Cobre_for_machine_learning/1450804. Accessed 05 Aug 2023
37. Network TMR (2023) ADHD200. https://paperswithcode.com/dataset/adhd-200. Accessed 11 April 2011
38. Gratch J, Artstein R, Lucas G.M, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, et al (2014) The distress analysis interview corpus of human and computer interviews. In: LREC pp 3123–3128. Reykjavik
39. Aminifar A, Rabbi F, Pun VKI, Lamo Y (2021) Monitoring motor activity data for detecting patients depression using data augmentation and privacy-preserving distributed learning. In: 2021 43rd Annual international conference of the IEEE engineering in medicine & biology society (EMBC) pp 2163–2169. IEEE
40. Aminifar A, Rabbi F, Pun KI, Lamo Y (2021) Privacy preserving distributed extremely randomized trees. In: Proceedings of the 36th annual acm symposium on applied computing pp 1102–1105
41. Aminifar A, Shokri M, Rabbi F, Pun VKI, Lamo Y (2022) Extremely randomized trees with privacy preservation for distributed structured health data. IEEE Access 10:6010–6027
42. Turner JA, Calhoun VD, Thompson PM, Jahanshad N, Ching CR, Thomopoulos SI, Verner E, Strauss GP, Ahmed AO, Turner MD et al (2022) Enigma+ coinstac: improving findability accessibility interoperability and re-usability. Neuroinformatics 20(1):261–275
43. Plis SM, Sarwate AD, Wood D, Dieringer C, Landis D, Reed C, Panta SR, Turner JA, Shoemaker JM, Carter KW et al (2016) Coinstac: a privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. Front Neurosci 10:365
44. Borger T, Mosteiro P, Kaya H, Rijcken E, Salah AA, Scheepers F, Spruit M (2022) Federated learning for violence incident prediction in a simulated cross-institutional psychiatric setting. Expert Syst Appl 199:116720
45. Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International conference on machine learning pp 1188–1196. PMLR
46. Xu X, Peng H, Bhuiyan MZA, Hao Z, Liu L, Sun L, He L (2021) Privacy-preserving federated depression detection from multisource mobile health data. IEEE Trans Industr Inform 18(7):4788–4797
47. Cao B, Zheng L, Zhang C, Yu PS, Piscitello A, Zulueta J, Ajilore O, Ryan K, Leow AD (2017) Deepmood: modeling mobile phone typing dynamics for mood detection. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining pp 747–755
48. Kirsten K, Pfitzner B, Löper L, Arnrich B.: Sensor-based obsessive-compulsive disorder detection with personalised federated learning. In: 2021 20th IEEE international conference on machine learning and applications (ICMLA) pp 333–339 (2021). IEEE
49. Lee DY, Choi B, Kim C, Fridgeirsson E, Reps J, Kim M, Kim J, Jang J-W, Rhee SY, Seo W-W et al (2023) Privacy-preserving federated model predicting bipolar transition in patients with depression: prediction model development study. J Med Internet Res 25:46165
50. Science OHD (2023) Informatics: standardized data: the OMOP Common Data Model. https://www.ohdsi.org/data-standardization/. Accessed 21 Aug 2023
51. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst , vol 30
52. Wang R, Fu B, Fu G, Wang M (2017) Deep & cross network for ad click predictions. In: Proceedings of the ADKDD'17 pp 1–7
53. Marulli F, Verde L, Marrone S, Barone R, De Biase MS (2021) Evaluating efficiency and effectiveness of federated learning approaches in knowledge extraction tasks. In: 2021 International joint conference on neural networks (IJCNN) pp 1–6. IEEE
54. Ahmed U, Lin JC-W, Srivastava G (2022) Hyper-graph attention based federated learning methods for use in mental health detection. IEEE J Biomed Health Inform 27(2):768–777
55. Basu P, Roy TS, Naidu R, Muftuoglu Z, Singh S, Mireshghallah F (2021) Benchmarking differential privacy and federated learning for bert models. arXiv:2106.13973
56. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
57. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: a lite bert for self-supervised learning of language representations. arXiv:1909.11942
58. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv:1907.11692
59. Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert a distilled version of bert: smaller faster cheaper and lighter. arXiv:1910.01108
60. Fan Z, Su J, Gao K, Peng L, Qin J, Shen H, Hu D, Zeng L-L (2021) Federated learning on structural brain mri scans for the diagnostic classification of major depression. Biol Psychiatry 89(9):183
61. Huang Z-A Hu Y, Liu R, Xue X, Zhu Z, Song L, Tan KC (2022) Federated multi-task learning for joint diagnosis of multiple mental disorders on mri scans. IEEE Trans Biomed Eng
62. Network TMR (2023) ABIDE. https://ieee-dataport.org/documents/autism-brain-imaging-data-exchange-abide. Accessed 27 March 2017
63. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security pp 308–318

64. Cui Y, Li Z, Liu L, Zhang J, Liu J (2022) Privacy-preserving speech-based depression diagnosis via federated learning. In: 2022 44th Annual international conference of the IEEE engineering in medicine & biology society (EMBC) pp 1371–1374. IEEE
65. Ringeval F, Schuller B, Valstar M, Cummins N, Cowie R, Tavabi L, Schmitt M, Alisamir S, Amiriparian S, Messner E-M, et al (2019) Avec 2019 workshop and challenge: state-of-mind detecting depression with ai and cross-cultural affect recognition. In: Proceedings of the 9th international on audio/visual emotion challenge and workshop pp 3–12
66. Wang H, Yurochkin M, Sun Y, Papailiopoulos D, Khazaeni Y (2020) Federated learning with matched averaging. arXiv:2002.06440