Journal of Disability Research

2025 | Volume 4 | Pages: 1–29 | e-location ID: e20240101
DOI: 10.57197/JDR-2024-0101

Responsible Artificial Intelligence for Mental Health Disorders: Current Applications and Future Challenges

Shaker El-Sappagh[1,2,3], Waleed Nazih[4], Meshal Alharbi[4] and Tamer Abuhmed[3,*]

[1]Computer Science Department, Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt
[2]Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Banha 13518, Egypt
[3]Department of Computer Science and Engineering, College of Computing and Informatics, Sungkyunkwan University, Suwon 16419, South Korea
[4]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia

Correspondence to:
Tamer Abuhmed*, e-mail: tamer@skku.edu, Tel.: +82 312907968

Received: February 19 2024; Revised: July 11 2024; Accepted: July 11 2024; Published Online: January 3 2025

ABSTRACT

Mental health disorders (MHDs) have significant medical and financial impacts on patients and society. Despite the potential opportunities for artificial intelligence (AI) in the mental health field, there are no noticeable roles of these systems in real medical environments. The main reason for these limitations is the lack of trust by domain experts in the decisions of AI-based systems. Recently, trustworthy AI (TAI) guidelines have been proposed to support the building of responsible AI (RAI) systems that are robust, fair, and transparent. This review aims to investigate the literature of TAI for machine learning (ML) and deep learning (DL) architectures in the MHD domain. To the best of our knowledge, this is the first study that analyzes the literature of trustworthiness of ML and DL models in the MHD domain. The review identifies the advances in the literature of RAI models in the MHD domain and investigates how this is related to the current limitations of the applicability of these models in real medical environments. We discover that the current literature on AI-based models in MHD has severe limitations compared to other domains regarding TAI standards and implementations. We discuss these limitations and suggest possible future research directions that could handle these challenges.

KEYWORDS

trustworthy AI, mental health disorders, responsible AI, machine learning robustness, model explainability, deep learning

INTRODUCTION

Mental health disorders (MHDs) are conditions that have a clinically significant disturbance, distress, or impairment in a patient's functioning, thinking, personality, cognition, emotional regulation, feeling, behavior, or mood (WHO, 2021). World Health Organization (WHO) estimated that MHDs will become the second leading cause of disability in the future (Alghadeer et al., 2018). In 2010, MHDs were the worldwide leading causes of years lived with disability, especially depression and anxiety (Garcia-Ceja et al., 2018). In 2017, MHDs represented 14.4% of the disabilities worldwide (Christensen et al., 2020). A 2017 US report stated that 46.6 million adults were affected by an MHD, which was nearly 20% of the US population (Thieme et al., 2020). Also, 38.2% of the European population suffers from some form of mental disorder (Rivera et al., 2022). In 2019, 970 million patients, i.e. one in every eight persons, had at least one mental disorder, and the disorders were especially anxiety and depression (Institute for Health Metrics and Evaluation, 2023). In 2019, 301 million patients were living with an anxiety disorder, 280 million patients had depression, 40 million patients had bipolar disorder, 14 million patients had eating disorders, and 40 million patients had conduct-dissocial disorder (Institute for Health Metrics and Evaluation, 2023). In 2020 and because of COVID-19, the number of anxiety and depressive patients increased significantly (WHO, 2022). In 2019 and based on the global burden of diseases, MHDs are considered among the top 10 causes of burden globally (GBD, 2019 Mental Disorder Collaborators, 2022). In 2020, 51.2% of females and 37.4% of males in the United States received mental health services (National Institute of Mental Illness, 2023). The percentage of US adults receiving mental health treatment increased from

19.2% in 2019 to 21.6% in 2021 (National Center for Health Statistics, 2022). Notably, depression is estimated to affect 350 million people worldwide (Abd Rahman et al., 2020). Depression (280 million people) and anxiety (301 million people) have an estimated global economic impact of $1 trillion every year (National Alliance in Mental Illness, 2023). In the Arab region, approximately 200 million Arab youth live with MHDs (Maalouf et al., 2019). Studies reported that some MHDs are highly prevalent in Saudi Arabia (Altwaijri et al., 2023). It is estimated that 12 billion productive workdays are lost yearly because of depression and anxiety alone (Cuijpers et al., 2023). The global spending on poor quality mental health services amounted to an estimated $2.5 trillion in 2010, with projections showing a surge to $6 trillion by 2030 (The Lancet Global Health, 2020).

There are many types of MHD including mental disorders and psychosocial disabilities. These disorders include neurodevelopmental, disruptive behavior and dissocial, schizophrenia, stress, personality (e.g. bipolar), mood (e.g. depression), delusional, and anxiety disorders, to name a few. Due to patient mortality that occurs prior to diagnosis, treatment, and management of illnesses, there is a substantial lag in the identification and management of MHDs (Tan et al., 2022). These MHDs have serious consequences not just for patients, but also for their families, friends, and society. There are effective prevention and treatment options for MHDs; however, patients usually do not have access to effective care, especially in Arab countries (Altwaijri et al., 2023). So, more effective mental healthcare is urgently needed. The WHO's comprehensive mental health action plan 2013-2030 recognizes the critical role of information systems and data analysis in the management of mental disorders.

MHDs are highly prevalent and have enormous burden and huge economic costs; however, globally, the mental health services fail to meet the needs of an average of 29% of patients (70% in high-income and 12% in low-income countries) (Cuijpers et al., 2023). Information and communication technologies including artificial intelligence (AI) technologies have excellent potential to fill the gap in the current MHD management (Thieme et al., 2020). These tools can support the early and accurate diagnosis of MHDs which could improve the quality of life of patients (Su et al., 2020; Rivera et al., 2022). Diagnosis of MHDs is different from that of other chronic diseases because the diagnosis process is mainly based on the patient's self-report to specific questionnaires to detect specific patterns of feelings or social interactions. For handling these issues, Saudi Arabia developed a patient-centered model of care as a part of the Vision 2030 for national health which aims to develop people socially, mentally, and physically. The chronic care system of this model includes mental health. These systems support the availability of big data for individual's mental health status which support the building of AI, machine learning (ML), and deep learning (DL) systems for improving the understanding of MHDs and assisting physicians for improved clinical decision-making. Meanwhile, ML/DL-based systems have been used in various domains.

There is huge literature on ML/DL models for mental disorder diagnosis, prediction, monitoring, and treatment (Garcia-Ceja et al., 2018; Chung and Teo, 2022; Garg, 2023). However, there is very little effect or use of these models in the real medical environments like hospitals and medical centers. The main reason for these limitations is model trustworthiness. AI-based systems were discovered to be vulnerable to attacks, biased against sensitive groups, lacking user privacy, prone to instability, and suffering from stochasticity which degrade user experience and reduce society's trust in these systems. In addition, the recent ML/DL models are black boxes where the user is not able to understand why a model has made specific decisions. ML/DL models learn medical knowledge from data that could be noisy, incomplete, inconsistent, and biased. As David Hume stated, there is no reason to expect the future to resemble the past (Stanford, 2002), and interestingly we use historical data to predict the future. Medical decisions are mainly based on causality, but the principle of ML is to identify correlations, not causality. Medical experts are worried about their decreasing role in AI-based systems, and lack of human control over AI systems is a main concern for the experts. Moreover, ethical issues and accountability of AI decisions raise another significant challenge (Ali et al., 2023). Although AI has great potential to solve real-world problems, applying these models in high-stakes domains like medicine poses significant risks for patients and physicians. Domain experts do not trust AI-based systems to make medical decisions because trust in the AI systems is not only about the model's performance but also about the model's robustness, fairness, transparency, privacy, security, etc. Medical experts need responsible AI (RAI)-based systems to be utilized in real settings. The lack of trustworthiness results in slower adoption of AI-based systems in real life. Trustworthy AI (TAI) is an AI system that is lawful, ethically adherent, and technically robust (Liu et al., 2022; Kaur et al., 2023). The ethical development of AI systems must be established in each stage of the system development lifecycle, i.e. design, development, deployment, and use. Many ethical frameworks, principles, regulations, and guidelines have been published recently to offer a comprehensive approach to support AI ethics and assist companies in designing, developing, deploying, and operating TAI systems (Buruk et al., 2020; Shneiderman, 2020; Sovrano et al., 2020; Wickramasinghe et al., 2020; Crockett et al., 2021; Serban et al., 2021; Han and Choi, 2022). The resulting systems are called RAI systems. RAI is critically designed to get the trust of users including physicians, patients, caregivers, and the government. TAI and RAI are used interchangeably in this study.

Studying the status of TAI in specific domains and highlighting the limitations and future research directions is critical because it is the starting point for building AI-based applications that are applicable and accepted in the real world. We have studied the literature on TAI in the Alzheimer's domain and discovered that TAI is the main reason for the current limitations in applying AI-based systems in this domain (El-Sappagh et al., 2023). Studying the literature of RAI in the mental disorders domain and comparing it with the standards and requirements of RAI and with the literature of RAI in other domains is critical. Currently, there is a growing number of studies that provide ML/DL methods for MHD prediction and management.

However, to the best of our knowledge, no study in the MHD domain has analyzed the literature of TAI. For this reason, this literature review aims to characterize the state-of-the-art of TAI and RAI topics for ML/DL models in the mental disorder domain to clarify its status and highlight the current research gaps that could boost the research in this critical domain. The contributions of the study can be summarized as follows:

- The study provides a review of RAI and TAI concepts and the related dimensions including robustness, fairness, and transparency based on the recent guidelines and protocols.
- The study then concentrated on investigating the literature of RAI in MHD. We review the state-of-the-art ML/DL studies in MHD that focused on model robustness issues including performance, uncertainty quantification and mitigation, validation, and model security and adversarial training.
- The study provides extensive reviews of the recent ML/DL studies in the MHD domain that focused on model fairness issues including data balancing, algorithmic bias, data bias, and the solutions to these issues.
- We studied the literature of explainable AI (XAI) in the MHD domain and highlighted the used XAI techniques to provide different types of explainability for model decisions.
- We further studied the role of multimodal data to improve the robustness of ML models in MHD and the role of data fusion to provide medically intuitive ML models that mimic domain experts.
- We finally explored the limitations of the ML/DL literature in the MHD domain and highlighted the possible research directions that extend this literature to be trustworthy and more responsible. The directions are expected to improve the trust of the MHD domain experts in AI-based applications which boosts the applicability of the ML/DL-based systems in real medical environments.

This paper is structured as follows. The Related Work section presents the surveys about the role of ML and DL in MHD diagnosis, prediction, monitoring, and management. The Methodology section discusses the methods that we followed to complete the survey. The Responsible Artificial Intelligence section 4 introduces the concepts of TAI and RAI. The RAI in the Literature section discusses the role of RAI in the literature. The RAI in Mental Disorder section discusses the role of RAI in MHDs. The Challenges and Future Research Directions section discusses the limitations and the future research directions, and the Conclusion section concludes the paper.

RELATED WORK

RAI is a crucial requirement to produce suitable ML and DL techniques for sensitive domains such as medicine. In this section, we evaluate the existing surveys of ML and DL techniques in the MHD domain. We concentrate on trustworthy and RAI directions such as robustness, fairness, uncertainty quantification and mitigation, and explainability. Table 1 shows a comparison of the 19 existing surveys of ML and DL literature pertaining to MHDs. In this table, we compare the existing survey papers by checking if the survey has handled/provided (i) the ML or DL methods, (ii) the RAI requirements including fairness, robustness, and XAI, (iii) the multimodality, (iv) the datasets, (v) the type of handled task [i.e. disease detection (diagnosis), prediction, monitoring, and medication], (vi) a comparison with the previous surveys, (vii) a systematic research methodology, (viii) the future research directions, and (ix) the type of discussed mental disorders. The table highlighted the critical gap of these surveys regarding the RAI requirements. There is no survey that discusses the current research literature of RAI in mental disorders diagnosis, prediction, monitoring, and management. Khare et al. (2023) concentrated on the detection of nine MHDs using classical ML and DL models and physiological electroencephalogram (EEG), electrocardiogram (ECG), magnetoencephalography, electromyogram (EMG), electrooculogram, heart rate variability, and arterial oxygen saturation signal modalities only. The study concentrated on the detection of diseases in children. In addition, the authors did not discuss the RAI in the study; however, they highlighted the uncertainty quantification and XAI requirement as crucial domains that need further investigation in the future. Moreover, the study noticed that multimodal data fusion improved the performance of ML/DL algorithms, but the study did not concede other modalities like images, text, sound, or structured data. De Bardeci et al. (2021) surveyed the convolutional neural network (CNN) and long short-term memory (LSTM)-based DL models for EEG signal processing for diagnosis and prediction of psychiatric disorders. Sui et al. (2020) provided a survey of classical ML models to predict mental disorders based on the neuroimaging data. The study highlighted the important role of longitudinal multimodal data fusion in improving the performance of classical ML algorithms. Greco et al. (2023) reviewed the transformer-based language models for solving mental health problems based on text data modality only. The study highlighted the challenges of text data processing that need further exploration including the semantic understanding of data using ontologies and the explainability of complex models like transformers. Arji et al. (2023) reviewed the recent DL techniques such as LSTM and CNN for mental disorders including depression and mood recognition analysis. Moura et al. (2020) and Garcia-Ceja et al. (2018) reviewed the ubiquitous monitoring techniques for mental health based on context data collected from ubiquitous devices. Cho et al. (2019) reviewed the role of five classical ML techniques including K-nearest neighbor (KNN), gradient boosting machine, random forest (RF), support vector machines (SVMs), and naïve Bayes for diagnosing mental disorders. Ahmed et al. (2022a) reviewed the ML/DL models [e.g. AdaBoost, CNN, gated recurrent unit, KNN, logistic regression (LR), LSTM, multilayer perceptron (MLP), RF, decision tree (DT), visual geometry group, and XGBoost] to detect anxiety and depression using social media data (Twitter, Facebook, Instagram, Reddit, and Sina Weibo), especially during the COVID-19 pandemic (2019-2020). The study highlighted the critical role of XAI in enhancing the interpretability of deep black-box models.

Table 1: List of related survey/review papers in mental health disorders.

Ref.	Year	Period	# of paper	Research methodology	ML/DL	PRISMA used?	Future challenges	Uncertainty	Fairness	Robustness	XAI	Multi-modalities	Datasets	Mental disorders	Literature compared	Medical task
Ours	2023			√	ML/DL	√	√	√	√	√	√	√	√	Any	√	DPRM
Khare et al. (2023)	2023	2012-2022	117	√	ML/DL	√	√	×	×	×	×	√	√	Nine	√	D
Squires et al. (2023)	2023	×	×	×	ML/DL	×	√	×	×	×	×	√	×	D	×	DM
Greco et al. (2023)	2023	×	16	×	DL	×	√	×	×	×	×	×	×	Any	×	D
Arji et al. (2023)	2023	2000-2023	85	√	DL	√	√	×	×	×	×	×	×	Any	×	D
Iyortsuun et al. (2023)	2023	2013-2022	33	√	ML/DL	√	√	×	×	×	×	×	×	SDABPA	×	D
Ahmed et al. (2022a)	2022	2010-2021	54	√	ML/DL	√	√	×	×	×	×	×	×	AD	×	P
Chung and Teo (2022)	2022	2010-2020	30	√	ML	√	√	×	×	×	×	×	×	Any	×	P
Malhotra and Jindal (2022)	2022	2010-2022	96	√	DL	√	√	×	×	×	×	×	√	SD	√	DP
Rivera et al. (2022)	2022	2017-2020	46	√	DL	×	√	×	×	×	×	×	√	Any	√	DP
de Bardeci et al. (2021)	2021	Up to 2020	30	√	DL	√	×	×	×	×	×	×	×	PD	×	DP
Su et al. (2020)	2020	Up to 2019	57	√	DL	√	√	×	×	×	×	×	×	Any	×	DP
Moura et al. (2020)	2020	2011-2019	20	√	ML	√	√	×	×	×	×	×	×	Any	√	R
Sui et al. (2020)	2020	2010-2019	122	×	ML	×	√	×	×	×	×	×	×	Any	×	P
Thieme et al. (2020)	2020	2000-2019	54	√	ML	√	√	×	×	×	×	×	×	Any	×	DM
Abd Rahman et al. (2020)	2020	2007-2018	22	√	ML/DL	√	√	×	×	×	×	×	×	Any	×	D
Cho et al. (2019)	2019	×	59	×	ML	×	×	×	×	×	×	×	×	SSDAB	√	D
Graham et al. (2019)	2019	2015-2019	28	√	ML	√	√	×	×	×	×	×	×	DSS	×	DP
Garcia-Ceja et al. (2018)	2018	×	×	×	ML	×	√	×	×	×	×	×	×	Any	×	R
Mendelson and Eaton (2018)	2018	2013-2018	×	×	ML	×	×	×	×	×	×	×	×	DAF	×	D

Abbreviations: AD, anxiety and depression; D, depression; DAF, depression, anxiety, and first-episode psychosis; DL, deep learning; DPRM, disease Detection (diagnosis), Prediction, Remote monitoring, or Medication; DSS, depression, schizophrenia or other psychiatric illnesses, and suicide ideation and attempts; ML, machine learning; Nine, autism spectrum disorder, attention-deficit hyperactivity disorder, schizophrenia, anxiety, depression, dyslexia, posttraumatic stress disorder, Tourette syndrome, and obsessive-compulsive disorder; PD, psychiatric disorders; PRISMA, preferred reporting items for systematic reviews and meta-analyses; SD, suicide and depression detection; SDABPA, schizophrenia, depression, anxiety, bipolar disorder, posttraumatic stress disorder (PTSD), anorexia nervosa, and attention-deficit hyperactivity disorder (ADHD); SSDAB, stress, schizophrenia, depression, autism, and bipolar; XAI, explainable AI.

However, the study did not investigate the XAI methods in the mental disorders domain. Similarly, Abd Rahman et al. (2020) and Malhotra and Jindal (2022) surveyed the literature for analyzing social media data to detect mental disorders using text analysis techniques. Thieme et al. (2020) reviewed the role of human–computer interaction and ML techniques to improve the quality of managing MHDs. The study mentioned the critical role of RAI and highlighted the issues of RAI pertaining to fairness, uncertainty, and algorithmic interpretability. Chung and Teo (2022) and Graham et al. (2019) reviewed the role of classical ML in predicting mental disorders such as anxiety and depression, schizophrenia, posttraumatic stress disorder (PTSD), bipolar disorder, and mental health problems among children. The study highlighted the main limitations of the current literature pertaining to the dataset size and quality, the limited role of DL algorithms especially using transfer learning, and model trustworthy issues such as the ones related to XAI and robustness. Iyortsuun et al. (2023) reviewed the recent ML/DL techniques to predict mental disorders including depression, anxiety, bipolar disorder, schizophrenia, attention-deficit hyperactivity disorder (ADHD), PTSD, and anorexia nervosa. Rivera et al. (2022) reviewed the role of EEG data and DL algorithms to detect mental disorders. The study highlighted the limitations of the current literature regarding model reproducibility and explainability and mentioned this as a hot research direction. Squires et al. (2023) demonstrated the critical role of multimodal data fusion to enhance the performance of ML and DL algorithms to diagnose and predict mental disorders. The study concentrated on depression detection and highlighted the limitations of the literature concerning uncertainty quantification, causal inference, data fusion of EEG, functional magnetic resonance imaging (fMRI) or MRI, and automatic ML. Su et al. (2020) conducted a comprehensive literature review regarding the use of DL techniques in mental issues. The study analyzed many types of data, including social media, genetics, clinical, and visual expression data.

As can be noticed in Table 1, there are no studies in the literature that evaluated the literature of RAI in the mental health domain. However, many studies (Thieme et al., 2020; Chung and Teo, 2022) have highlighted the severe limitations of the current literature regarding this critical research direction. It is also noticed that old surveys concentrated much on ML, but recent surveys concentrated more on DL methods. Most surveys concentrated on more than one mental disorder (Garcia-Ceja et al., 2018; Rivera et al., 2022; Greco et al., 2023; Khare et al., 2023), but some studies have concentrated on specific disorders, e.g. depression (Squires et al., 2023) and anxiety and depression (Ahmed et al., 2022a). Even though RAI is a crucial requirement to build TAI models, there are no surveys in the literature that evaluated the current RAI research in the mental health domain. TAI gained the attention of researchers in other domains. For example, in El-Sappagh et al. (2023), the authors studied the literature on Alzheimer's disease and TAI and highlighted the critical research directions to enhance the current literature of this disease. To the best of our knowledge, the use of RAI techniques for specifically addressing MHDs has not been deeply investigated so far. In our study, we target filling this literature gap by providing the first survey of methods using RAI for the detection, prediction, or monitoring of mental disorders.

METHODOLOGY

Search strategy

The selection of relevant studies was conducted by using precise candidate search terms. For the purpose of identifying the most recent studies, the following search terms were used: {responsible, ethics, ethical, responsibility, trust, trusted, trustworthiness, trustworthy, robustness, fairness, explainable, interpretable, transparent, transparency, explainability, reliability, reliable, safety, safe, privacy, private, security, secure, biased, Trustworthy, trustworthiness, user-centric, human-centric, ethical, reproducibility, reliable, XAI, or accountability}, {Multimodal, images, text, time series, structured data}, {machine learning, deep learning, artificial intelligence, AI, model, ensemble, decision support, clinical decision support systems, CDSS, system, algorithm}, {mental health disorder, depression, anxiety, schizophrenia, stress, eating disorders, bipolar, addictive behaviors, disruptive behavior and dissocial disorders, neurodevelopmental disorders, mental disability}, {diagnosis, prediction, detection, monitoring, management}.

For a more effective search strategy, the controlled terms and their synonyms were combined using "AND" and "OR." An initial comprehensive search was conducted across five major electronic databases, namely Nature, ScienceDirect, SpringerLink, PubMed, and IEEE Xplore. Additionally, Google Scholar and Scopus were searched to verify the results. To compile the latest scholarly works, every publication that was uploaded to arXiv and medRxiv was considered. Duplicates have been eliminated from the arXiv and medRxiv publications. The search started by checking the titles, abstracts, and keywords of the papers. These papers were then reviewed manually. In the third step, the entire texts of the articles that satisfied our inclusion criteria were considered. Furthermore, to identify other important papers, the reference lists of the chosen articles were reviewed manually. The final compilation of articles was included in the review procedure. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards were employed to check these articles (Moher et al., 2009); see Figure 1. As shown in Table 1, PRISMA is the most popular methodology for making systematic surveys.

Eligibility criteria

This study focused on reviewing the recent literature on MHDs. We concentrate on reviewing recent ML and DL studies in the MHD domain and investigate the implemented trustworthy guidelines in these studies. The inclusion and exclusion criteria were as follows: inclusion criteria are (i) the study focused on developing, testing, and discussing classical ML, DL, trustworthiness, multimodality, ensemble, or any other hybrid algorithms in the mental disorders

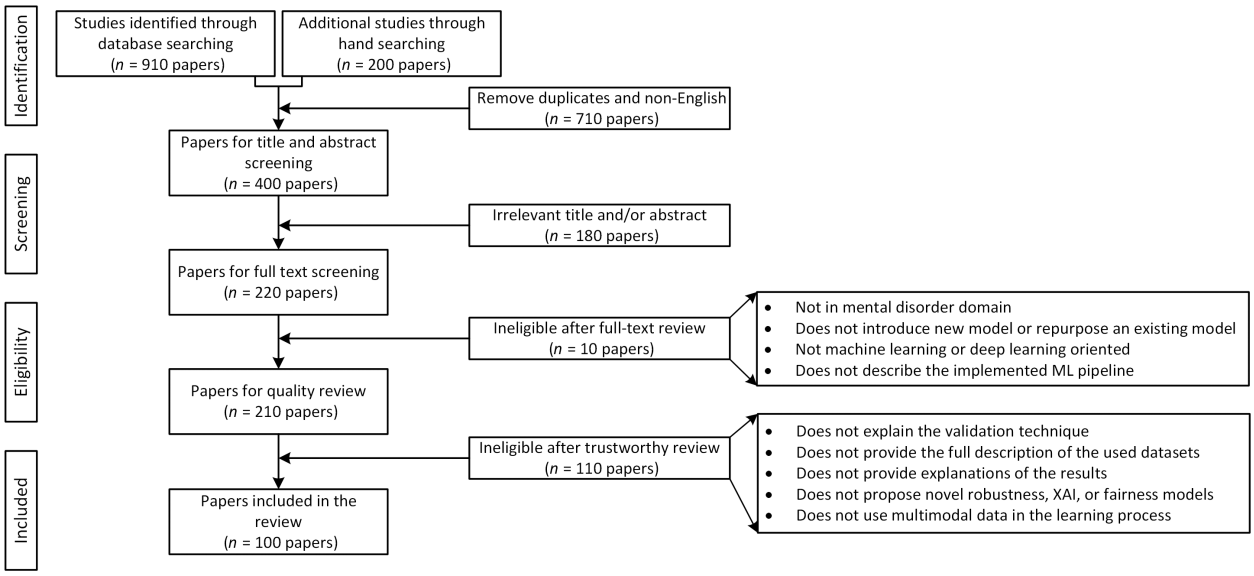

Figure 1: PRISMA flowchart for the study. Abbreviations: ML, machine learning; PRISMA, preferred reporting items for systematic reviews and meta-analyses; XAI, explainable AI.

domain; (ii) only English language studies were considered; (iii) the studies published in the period of 2018-2023 were considered; and (iv) only MHD diagnosis, detection, prediction, monitoring, and management studies were included. Articles that do not meet the included criteria are excluded from further processing. For example, incomplete studies, editorials, opinion papers, or reviews were considered out of scope. Furthermore, articles that could not be accessed in full text are excluded. Finally, the full texts of 100 papers were included in the full-text review.

Study selection

Figure 1 shows that 100 papers were ultimately selected for review and analysis in this study.

RESPONSIBLE ARTIFICIAL INTELLIGENCE

AI has become essential across industries. It helps to boost the human decision-making process. However, in sensitive and high-stakes domains like medicine, we did not see the major effects of AI-based systems on patients and in physicians' daily practice. The benefits of AI-based systems do not outweigh their potential negative impacts on society because of the negative consequences of their underuse, misuse, and abuse. The main reason for this non-outweighing is the level of trustworthiness of the underlying ML models of these systems. TAI, RAI, lawful AI, or ethical AI is a new track in AI and ML research. Liu et al. (2022) discussed the relationship among these concepts. Dignum (2018, 2019) defined RAI as "*Responsible AI is about human responsibility for the development of intelligent systems along with fundamental human principles and values, to ensure human flourishing and wellbeing in a sustainable world*." Trustworthiness has been defined by Ding et al. (2022) as "*the degree of confidence to which the AI solution will behave as expected when encountering real-world problems*." RAI has the following seven main principles, as shown in Figure 2:

- *Interpretability (explainability, transparency, and provability):* where the AI system can explain its model decisions.
- *Robustness (reliability, accuracy, and security):* where AI systems can operate reliably, accurately, and safely over long periods and be able to prevent, detect, and recover from possible attacks using the right models and datasets. All sources of uncertainty must be quantified, and

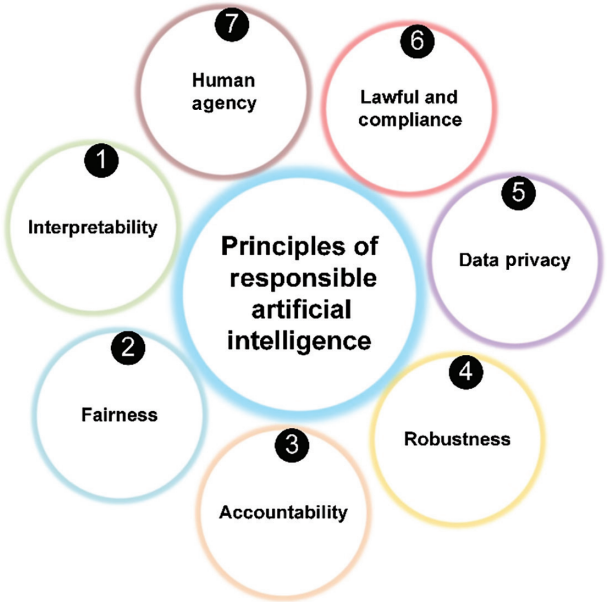

Figure 2: The principles of responsible AI. Abbreviation: AI, artificial intelligence.

suitable countermeasures must be set to mitigate these uncertainties.
- *Fairness:* where AI systems should prevent any possible discrimination or biases (e.g. algorithmic and data) against individuals within similar groups or against social groups based on their sensitive features like gender, color, religion, etc.
- *Accountability:* where the moral implications of the use and misuse of an AI system should be the responsibility of its stakeholders. A clearly identified accountable party of AI decisions must have effective oversight and control over these decisions.
- *Data privacy:* where individuals must have full control of their data when they are used to train and run AI solutions.
- *Lawfulness and compliance:* where the AI system's stakeholders must act according to the law and suitable regulatory regimes.
- *Human agency:* where human intervention in the operations of AI solution should be dictated by the level of perceived ethical risk severity.

RAI principles should be used throughout the whole life cycle of ML models, including problem definition, system deployment, and performance monitoring. To implement these principles, a number of AI ethical frameworks outlining the optimal procedures for the whole lifecycle of AI systems have been released (Jobin et al., 2019). Notably, about 100 fundamental AI ethical principles have been declared by corporations, governments, and organizations (Jobin et al., 2019). In addition, RAI standards have been developed to govern the processes of developing and using AI systems (Shneiderman, 2020). Examples of these standards include (i) ISO/IEC 42001 IT-AI-Management System Standard (https://www.iso.org/standard/77304.html) released by ISO/IEC JTC 1/SC42 AI Technical Committee which facilitated AI systems' certification and WG3 trustworthiness for risk management and bias (https://www.iso.org/committee/6794475.html), (ii) Architectural Framework and Implementation Guide for Federated Learning (https://standards.ieee.org/ieee/3652.1/7453/), (iii) Technical Framework Standard and Trusted Execution Environment Requirements for Shared Machine Learning (https://standards.ieee.org/ieee/2830/10231/), and (iv) IEEE p7000 IEEE Standards for Model Process for Addressing Ethical Concerns During System Design (https://ethicsinaction.ieee.org/p7000/) released by IEEE. Moreover, leading AI organizations implemented codes of ethics which are sets of rules that employees should follow when building AI systems (Shneiderman, 2020), and many industrial companies have committed to the principles of RAI (de Laat, 2021).

Major industry players have provided tools to implement RAI requirements. These tools are mostly focused on the robustness, fairness, and explainability of AI models. For example, Microsoft (https://www.microsoft.com/en-us/ai/responsible-ai) provided Human AI Interaction Toolkit, AI Trust Score, Fairness checklist, Fairlearn, InterpretML, Counterfeit, Conversational AI guidelines, SmartNoise, Presidio, Datasheet for Datasets, Confidential computing for ML, SEAL, and Responsible AI toolkit. Google (https://ai.google/responsibility/responsible-ai-practices/) provided People + AI Guidebook (PAIR), Rules of Machine Learning, Human-Centered Machine Learning, Model cards, Data cards, Fairness indicators, know your data, ML-fairness-gym, Language Interpretability Tool, What-If tool, Explainable AI, Google Tensorflow Privacy, and Google TensorFlow Federated. IBM (https://www.ibm.com/impact/ai-ethics) provided AI Explainability 360, AI Fairness 360, AI Privacy 360, Adversarial Robustness 360, AI FactSheets 360, Uncertainty Quantification 360, and Causal Inference 360. Meta (https://ai.meta.com/responsible-ai/) provided Fairness Flow, AI System Cards, Crypten, and Captum. Amazon (https://aws.amazon.com/machine-learning/responsible-machine-learning/) provided Amazon SageMaker Clarify, Amazon SageMaker Model Monitor, and Amazon Augmented AI. In the medical domain, the most practical ethical principles for AI system building are interpretability (XAI), robustness, and fairness (El-Sappagh et al., 2023). The current study shows how the use of these crucial principles can enhance the domain expert's trust and reliance on an AI to build clinical decision support systems (CDSSs). RAI is discussed in the context of medical CDSS for mental disorders as a case study. These ethical principles have been defined by the High-Level Expert Group on AI, appointed by the European Commission, in the document "*Ethics Guidelines for Trustworthy AI*," published in April 2019 (Smuha, 2019).

Explainable AI

Model explainability (i.e. XAI) is crucial for the model trust from stakeholders (Ding et al., 2022; Ali et al., 2023). Arrieta et al. (2020) defined XAI as "*given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.*" It is crucial for humans to comprehend, properly trust, and successfully control AI algorithms. XAI supports the discovery of deeper knowledge about the task, the justification of model decisions, the control of the AI system and adjusting for possible mistakes, and the debugging of the AI model (Arrieta et al., 2020). According to Miller (2019), the explainability of AI decisions is significant for two main reasons: (i) trust, because stakeholders cannot believe that AI decision is correct just from some statistics about the model performance, and (ii) ethics, because it should be proved that the system will not entertain discrimination of any kind in its functioning. As a result, XAI is very related to model trustworthiness (Albahri et al., 2023).

During the last decade, numerous XAI methods were proposed; see Figure 3. These methods have been divided according to many criteria. For example, model-based XAI divided XAI methods based on the transparency of the model, i.e. transparent models like rule-based, DT, KNN, etc., or opaque models like DL (CNN, recurrent neural network, transformer, and MLP), ensemble (bagging, voting, boosting, and stacking), and SVM models. A variety of approaches have been proposed for each of these groups. To obtain a thorough and up-to-date list of

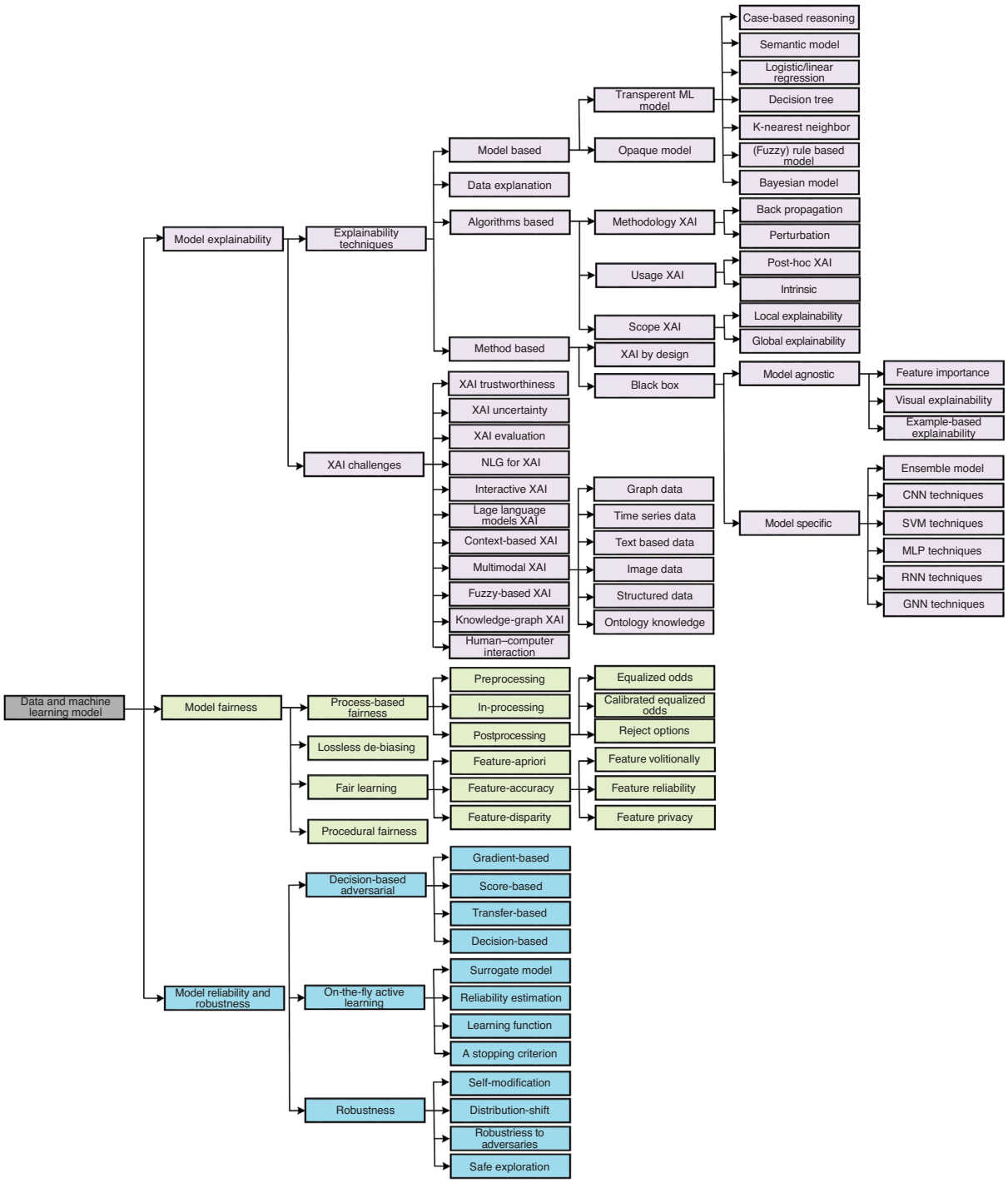

Figure 3: TAI principles and their methods. Abbreviations: CNN, convolutional neural network; GNN, graph neural network; ML, machine learning; NLG, natural language generation; RNN, recurrent neural network; SVM, support vector machine; TAI, trustworthy AI; XAI, explainable AI.

these XAI approaches, readers are encouraged to read the following recent survey papers: Ding et al. (2022) and Saranya and Subhashini (2023). Transparent models are explainable by design because they are white-box models. Their XAI is known to follow ante-hoc approaches. Other posthoc surrogate models, which could be model-specific or model-agnostic, are used to explain black-box models. Model-specific approaches are specific to the model at hand, whereas model-agnostic techniques, such as feature importance techniques such as local interpretable model-agnostic explanations (LIME) and SHapely additive exPlanations (SHAP), and visualization techniques such as saliency map and gradient-weighted class activation mapping, are independent of any model (Ali et al., 2023). More formally, the architecture of XAI is shown in Figure 4 (Vilone and Longo, 2021) for an opaque model M_o and dataset $D = \{X, Y\}_1^N$ of N samples. Explanation of M_o is to find a mapping function f^o: $M_o, D \rightarrow M_w$ with M_o and D as the input, and M_w white-box model as the output, such that M_w behaves similar to M_o and has an XAI function f_w: $M_w, X_i \rightarrow e_i$ which can provide

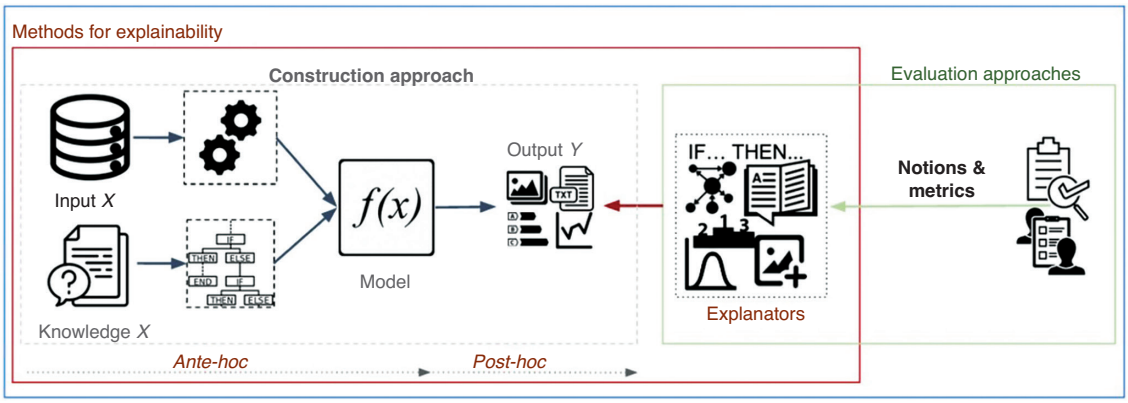

Figure 4: Main architecture of explainable artificial intelligence (Vilone and Longo, 2021).

a human interpretable explanation e_i for the decision made for each sample X_i. Data explainability is to understand the nature of the data through different exploratory data analysis techniques. For space restrictions, we will not get into details of the theory and math behind each of these methods, but interested readers are advised to read our recent survey on XAI techniques (Ali et al., 2023). Other surveys for XAI techniques and tools are the ones like Arrieta et al. (2020) and Saeed and Omlin (2023). XAI research path is still in its early stage (Leichtmann et al., 2023). The XAI evaluation (Ding et al., 2022), XAI stability, trustworthiness of XAI, XAI uncertainty quantification and mitigation (Zhang et al., 2022a; Lofstrom et al., 2023; Mehdiyev et al., 2023), multimodal XAI fusion including time series data (Joshi et al., 2021; Rojat et al., 2021; Lucieri et al., 2022), semantic (ontology-based) XAI (Panigutti et al., 2020; Rožanec and Mladenić, 2021; Adhikari et al., 2022), human-centric XAI (Kim et al., 2023), causality (Chou et al., 2022), context-aware XAI (Jiang et al., 2022), XAI-as-a-service and XAI embedding (Saeed and Omlin, 2023), machine-to-machine explanation (Saeed and Omlin, 2023), knowledge graph-based rich XAI (Rožanec et al., 2022), model security and data privacy, interactive and dynamic XAI, and human–computer interaction-based XAI are crucial topics to consider in the future to enhance model trustworthiness (Schoonderwoerd et al., 2021; Williams, 2021; Nyrup and Robinson, 2022; Rožanec et al., 2022; Moulouel et al., 2023; Panigutti et al., 2023). For example, XAI and security must be discussed from two main viewpoints including model confidentiality and adversarial attacks (Arrieta et al., 2020). The majority of XAI studies are focused on a single modality, such as image, text, structured data, or time series data. Domain specialists, on the other hand, always like to investigate why a model made a given decision from several perspectives.

As a result, it is critical to combine multimodal XAI capabilities and validate their consistency to confirm the stability and sufficiency of the offered XAI features. Since there is no widely acknowledged criterion for assessing the quality of generated explanations, XAI evaluation is a hot research topic. There are many proposed models for XAI evaluation including mental models, functional evaluation, effectiveness and satisfaction, confidence and dependence, and human–AI performance (Vilone and Longo, 2021; Ding et al., 2022). Generally, it is difficult to quantitatively evaluate the XAI features of a model. The trustworthiness, level of stability and consistency, and uncertainty quantification and mitigation of XAI results need further exploration (Ali et al., 2023; Ding et al., 2022). Explainability must be context aware, where the provided explanations are based on the level of experience of the user (Jiang et al., 2022). Many limitations of the current XAI literature can be found in Saeed and Omlin (2023).

Robustness and reliability

The ability of an algorithm to deal with execution failures, incorrect inputs, or unknown data is referred to as robustness (Li et al., 2023); see Figure 3. A lack of robustness may result in unanticipated or detrimental behavior in the system, reducing its safety and trustworthiness. The robustness principle deals with the system's performance, generalizability, security, privacy, uncertainty, and reproducibility. As shown in Figure 3, robustness has been discussed from different perspectives. The four dimensions of resilience defined by Leike et al. (2017) include self-modification, distributional shift, robustness to adversaries, and safe exploration. The ability of an AI system to adapt itself in response to the demands of new settings is referred to as self-modification. The ability of an AI system to adapt to its deployed environment is known as distributional or domain shift. The ability of AI systems to withstand adversarial attacks is referred to as robustness to adversaries. Exploration of AI agent safety in both actual and learning contexts is called safe exploration. Brendel et al. (2017) divided adversarial attacks into three types: gradient based, score based, and transfer based. Moustapha et al. (2022) suggested an active learning-based robustness framework that included (i) a surrogate model, (ii) a reliability estimation algorithm, (iii) a learning function, and (iv) a stopping condition.

In terms of performance, AI systems have to be accurate before even being deployed in the real world. Many metrics, including accuracy, precision, recall, F1-score, balanced accuracy, and area under the curve (AUC), can be used to assess the AI system performance. These metrics could be collected for cross-validation performance, internal testing,

or external validation using a dataset from different distributions (El-Sappagh et al., 2023). The sizes of the training and testing data, the way of preparing these data, and the model's external validation affect the model's generalizability level. Information leakage is a problem where the testing data are like the training data. Designing an accurate ML pipeline could prevent this problem. Distributional shift is another critical problem that affects the performance of deployed systems. Training of AI models must consider the diversity in data distributions. First, AI models need to be trained on sufficient data from the same distribution of real-world scenarios. Second, after deployment, the system needs to be robust against distributional shifts. This could be done by periodically retraining the model with new data that reflect the current distributions.

Regarding security, AI systems are vulnerable to adversarial attacks with malicious intentions for many reasons (Xue et al., 2020). Figure 5 shows the vulnerabilities of the AI pipeline for various adversarial attacks on the stages of the learning pipeline (i.e. training and testing phases). For instance, data poisoning and backdoor attacks are the key vulnerabilities at the data preparation stage. As the AI model is trained and deployed to a real environment, model outputs can be exploited by the adversary to conduct several attacks such as manipulating inputs to the model (i.e. generating an adversarial perturbed sample) to make the model generate bad output. Model theft, membership inference attacks, and training data recovery are among various security and privacy attacks that can be launched against the deployed model at the testing phase. The training technique, for example, can be outsourced, training data can originate from untrustworthy sources, and pre-trained models can come from third parties. However, the motivations behind these attacks are currently unclear. Security threats of AI systems are classified as follows: (i) training set poisoning, (ii) backdoor in the training data, (iii) adversarial example or evasion attacks, (iv) model shift, and (v) training data recovery (Xue et al., 2020). The evasion attack generates perturbed samples to mislead the model, the poisoning attack inserts carefully designed training examples into the training dataset in order to change the model's decision to specific samples or patterns, and the exploratory attack attempts to steal knowledge about the models. Threats of adversarial attack and various defense strategies have been well studied in the literature (Wang et al., 2019; Qayyum et al., 2021). Evasion attacks are divided into white-box and black-box attacks. White-box attacks include the auto projected gradient descent, shadow attack, Wasserstein attack, PE malware attacks, Brendel & Bethge attack, high confidence low uncertainty attack, iterative frame saliency, robust DPatch attacks, ShapeShifter attack, projected gradient descent, NewtonFool, adversarial patch, basic iterative method, Jacobian saliency map, DeepFool, virtual adversarial attack, fast gradient attack, etc. (Liu et al., 2018; Nicolae et al., 2018). Black-box attacks include Square attack, HopSkipJump attack, threshold attack, pixel attack, spatial transformation, query-efficient black-box, zeroth order optimization, decision-based/boundary attack, and geometric decision-based attack (Liu et al., 2018; Nicolae et al., 2018). Poisoning attacks include backdoor attack, clean-label backdoor attack, Bullseye Polytope, and BadDet attacks among others (Liu et al., 2018; Nicolae et al., 2018). Extraction attacks include functionally equivalent extraction, Copycat CNN, and KnockoffNets (Liu et al., 2018; Nicolae et al., 2018). Every attack can target DL models or classical ML models, and each one has its corresponding countermeasure.

Data sanitization is a countermeasure to poisoning or backdoor attacks in which poisoned data are filtered out before the training process. Smoothing model outputs are countermeasures against hostile instances that lower the model's sensitivity to small changes in the input. There are three types of defense techniques for sensitive information leakage: (i) distributed learning framework, (ii) classical cryptographic primitives-based approaches, and (iii) trusted platform-based approaches. Comprehensive information about adversarial attacks and their countermeasures may be found in Xue et al. (2020). Robustness tests like the monkey test (Exforsys, 2011) and security evaluation curves (Biggio and Roli, 2018) can be used to evaluate the security of the AI system. There are situations when the requirements for robustness conflict. For example, the training dataset as well as the code for model optimization and data preparation should be readily available for repeatability. This, however, will jeopardize the model and make it vulnerable to different attacks. Another example of the relationship between adversarial robustness and generalization is that the algorithms that are resilient against small perturbations have higher generalization (Xu and Mannor, 2012). However, recent studies (Raghunathan et al., 2019) showed that improving adversarial robustness through adversarial training could decrease test accuracy and hinder generalization. For recent advances in the robustness of ML models, readers are advised to refer to Xiong et al. (2022).

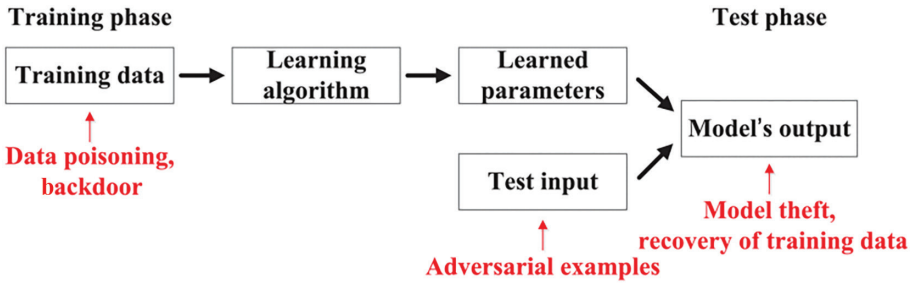

Figure 5: Different attacks on ML models (Xue et al., 2020). Abbreviation: ML, machine learning.

Fairness and diversity

The avoidance or reduction of undesired discriminatory bias impacts on individuals and social groups is referred to as AI system fairness. Bias is defined as the unfair treatment of certain groups of individuals based on sensitive information (e.g. gender, race, ethnicity, etc.). Carey and Wu (2022) defined bias as "*the prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair*." This bias can be found in data (e.g. measurement bias, excluded variable bias, representation bias, accumulation bias, sampling bias, longitudinal data fallacy, and linking bias), algorithms (e.g. algorithmic bias, user interaction bias, popularity bias, emergent bias, and evaluation bias), user experiences (e.g. historical bias, population bias, self-selection bias, social bias, behavioral bias, and temporal bias), and evaluation (i.e. wrong evaluation metrics were used) (Caton and Haas, 2020; Mehrabi et al., 2021). If training data contain biases, the ML model trained on these data learns these biases and reflects them in its decisions. Existing data biases affect the models trained by these data and result in biased predictions. On the other hand, models could amplify the existing data biases. Moreover, algorithms could have biased behavior based on specific design choices, even if the data are unbiased. Decisions of the biased algorithms that are fed into real-world environments affect end users' decisions, and this results in much more biased data that will be used to (re)train future models. There are four types of ML fairness techniques: (i) fairness based on process, (ii) fairness based on lossless de-biasing, (iii) fair learning, and (iv) procedural fairness (Wu et al., 2023). For processing-based fairness, as shown in Figure 6, IBM's AI Fairness 360 package (Bellamy et al., 2019) offers four *preprocessing* algorithms that transform training data to remove any discriminations: (i) re-weighing preprocessing, (ii) optimized preprocessing, (iii) learning fair representations, and (iv) disparate-impact remover. The package also offers three *in-processing* algorithms that modify the algorithm to remove any discriminations: (i) adversarial debiasing, (ii) prejudice remover, and (iii) meta fair classifier. Furthermore, three *postprocessing* algorithms are also offered by the package that reassigns the model's labels based on a function: (i) adversarial debiasing, (ii) prejudice remover, and (iii) meta fair classifier with (i) equalized odds, (ii) calibrated equalized odds, and (iii) classification of reject decisions (Liu et al., 2022). From Figure 6, it can be noticed that fairness can be implemented at many levels. We can apply the fairness preprocessing techniques to the prepared training data to form transformed data. These data can be used directly to train an ML model, as shown in the orange path. We can apply the in-processing techniques to the learning algorithm and train it using the preprocessed data, as shown in the purple path. We can combine both preprocessing and in-processing techniques as shown in the green path. We can combine all techniques as shown in the blue path of Figure 6, and this achieves complete fairness. As shown in Figure 3, Zhou et al. (2021) offered an information lossless approach for fairness. This approach oversampled the underrepresented group to balance the demographic population's majority and minority. Grgić-Hlača et al. (2018) introduced procedural fairness or fair learning, which took into account input attributes and moral judgments. The technique included three metrics of procedural fairness: feature apriori, feature accuracy, and feature disparity. These metrics

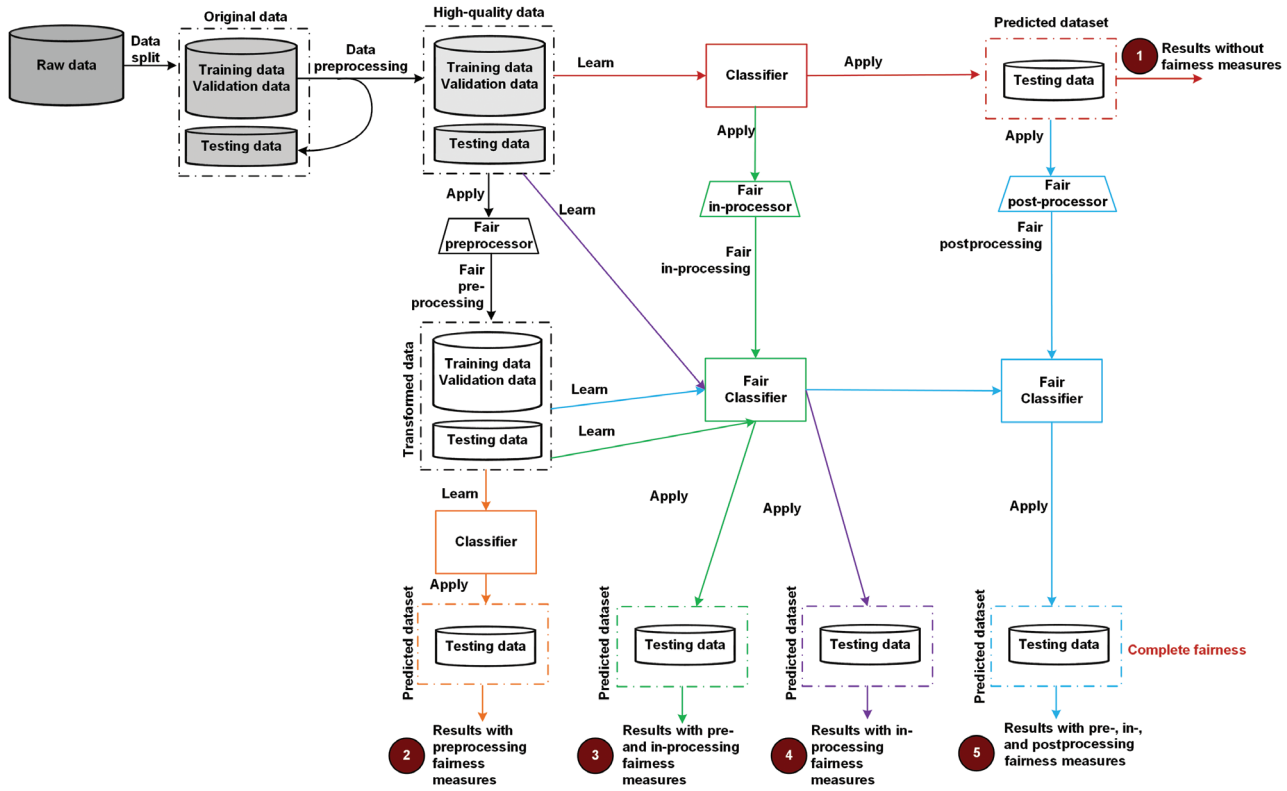

Figure 6: A proposed pipeline for preprocessing, in-processing, and postprocessing fairness.

eliminated binary discrimination. Metrics for evaluating fairness are based on statistical evidence. Positive classification rates are measured using statistical parity. The equalized odds method uses false-positive and false-negative rates. Predictive parity uses true-positive rates. Any fairness metric must satisfy the three fairness criteria of independence, separation, and sufficiency (Carey and Wu, 2022). Fairkit-learn provided a Python-based framework for fairness evaluation and comparison (Johnson and Brun, 2022). The math behind the fairness ML algorithms, fairness evaluation metrics, and the challenges and future research directions of fair ML can be found in Bellamy et al. (2019), Ashokan and Haas (2021), Mehrabi et al. (2021), and Liu et al. (2022).

Building TAI systems requires the application of TAI principles and guidelines at every step of the entire lifecycle of the AI systems. These principles are beyond traditional performance metrics like accuracy and range from data acquisition to model development to model deployment and finally to the model's continuous monitoring and governance. Other dimensions, such as fairness, explainability, adversarial robustness, and distribution shift, must be evaluated to produce RAI models. Note that these are kinds of trade-offs among the dimensions (Singh et al., 2021; Li et al., 2023); for example, improving the explainability of a model needs to decrease its complexity which decreases its performance, and optimizing the fairness affects the model's accuracy (Liu and Vicente, 2022). A formal analysis of the conflict between fairness and robustness can be found in Chang et al. (2020). Zhang et al. (2022b) formulated the best practices for ML model testing regarding fairness, robustness, and correctness. For comprehensive surveys about TAI, readers are requested to refer to Li et al. (2023). As can be noticed from the previous discussion, there are three requirements for deploying a TAI system. For a survey of all TAI requirements, and their status, future research directions, tools, and theory behind, readers are guided to refer to Liu et al. (2022). In El-Sappagh et al. (2023), we proposed a comprehensive trustworthy ML pipeline which considered the measures of fairness, robustness, and explainability in every step of the pipeline. The framework covers the checklist defined by Han and Choi (2022) to achieve TAI systems. The framework was oriented toward Alzheimer's domain, but there is nothing special about Alzheimer's in this framework. The model can be customized for any medical problem including MHDs. Implementing this framework assures the handling of TAI requirements related to the three critical principles of fairness, robustness, and explainability.

RAI IN THE LITERATURE

In this section, we briefly explore the TAI and RAI literature. Albahri et al. (2023) reviewed the role of XAI, data fusion, data quality, and bias analysis to improve the trustworthiness of the medical applications of AI. The study highlighted that respecting laws, ethics, and model robustness are the major challenges in modern AI applications. Improving the quality of the data engineering pipeline (i.e. data design, data sculpting, and data evaluation) to prepare suitable datasets is crucial to improve the trustworthiness of the resulting models (Liang et al., 2022). Ali et al. (2023) considered that the trustworthiness of AI systems can be achieved by model explainability only. They also provided a comprehensive review of XAI techniques in the medical domain. The studies in the review assumed that safe, robust, and TAI models could be implemented by concentrating on XAI features that could remove the lack of transparency of ML and DL models. On the other hand, Rasheed et al. (2022) surveyed the TAI principles and applications in the general healthcare domain and highlighted the role of XAI in improving the model's trustworthiness. They discussed the relationship among the requirements of the model including the trade-off between accuracy, explainability, and robustness (Liu et al., 2022). Many challenges have been discussed by Rasheed et al., to achieve RAI in the healthcare domain, including: (i) the quality of medical data such as imbalanced, biased, and noisy data, (ii) preserving the privacy of patients and the confidentiality of their data to prevent possible missuses and unprotected data sharing, and (iii) obtaining informed consent from patients before exposing them to any medical intervention, which is difficult because of the black-box nature of most ML models. Siala and Wang (2022) reviewed 253 articles about TAI and AI ethics in healthcare. Siala and Wang reached the conclusion that implementing RAI in healthcare settings remains a challenge. The study proposed an RAI framework that encompasses the five themes of sustainability, human-centeredness, inclusiveness, fairness, and transparency. The study also asserted the link among TAI principles.

Meanwhile, the literature on XAI explored the role of multimodal data and knowledge to enhance the level of explainability. Lucieri et al. (2022) provided a multimodal XAI framework that provided multimodal concept-based and visual explanations for detecting malignancy of skin lesions based on dermoscopic image analysis. The study highlighted the importance of multimodal data for providing comprehensive XAI. Díaz-Rodríguez et al. (2022) proposed the eXplainable Neural-symbolic learning which fused DL representations with expert domain knowledge graph for monument façade image classification. Baniecki et al. (2023) claimed that they implemented a TAI model for hospital length of stay prediction based on multimodal data. The study only provided a few XAI features and neglected the other main TAI principles including fairness and robustness. Pessach and Shmueli (2021) concentrated only on the fairness dimension of TAI by studying the algorithmic bias of AI-based fairness-aware hiring semi-supervised association learning system. The study proposed many in-process and preprocess techniques to overcome biases. Holzinger et al. (2022) mentioned that achieving TAI in the medical domain needs the combination of (i) complex DL models, (ii) graph causal models and expert knowledge, and (iii) verification and explainability methods. To achieve this objective, information fusion can be used by integrating data from expert knowledge, AI-based models, and medical databases.

Kovalchuk et al. (2022) proposed a three-stage CDSS for the type 2 diabetes mellitus prediction process. The model integrated the predictive power of rule-based and data-driven approaches. The study extended the model by adding XAI

features using SHAP. González-Gonzalo et al. (2021) studied the literature of AI and TAI in ophthalmic practice. These authors noted that even though AI models achieved close or even superior performance to medical experts, there is a critical gap between the development and the real application of AI systems in this domain. TAI can close this gap. TAI challenges (accuracy, resiliency, robustness, fairness, explainability, safety, and accountability) need to be considered along the AI design pipeline. González-Gonzalo et al. asserted that building TAI results from multi-stakeholder interaction including AI developers, healthcare providers, healthcare institutions, patients, regulatory bodies, etc. Zhou et al. (2023) surveyed the XAI and robustness measures for EEG-based systems, and Ma et al. (2022) reviewed the TAI, including explainability methods, in the dentistry domain.

Many studies in the literature in different domains focused on the fairness principle to improve the resulting model of responsibility and trustworthiness. Liefgreen et al. (2023) discussed the role of fairness and transparency to improve the acceptance of AI systems in the medical domain. The study mentioned that an effective TAI solution requires human engagement. Chiu et al. (2023) built a dermatological disease diagnosis trustworthy model by focusing on fairness. The study discovered that deeper neural network (NN) layers result in higher accuracy, but fairness conditions deteriorate for the extracted features of the deeper layers. The study proposed to use the fairness-oriented concept of multi-exit to enhance model fairness. Drukker et al. (2023) studied TAI in medical image analysis through the fairness principle. The study identified 29 sources of potential bias and mitigation strategies. Venugopal et al. (2023) studied the concept of fairness in radiology by focusing on bias auditing using the Aequitas open-source toolkit. Kozodoi et al. (2021) studied the role of fairness in profit-oriented credit scoring systems. These authors extended the ML pipeline by some algorithmic processors to achieve the fairness goals. The authors also found that the in-processing techniques have achieved a good balance between profit and fairness. In addition, the study reduced algorithmic discrimination to a reasonable level at a relatively low cost. Pfohl et al. (2021) investigated the fairness of ML models for clinical risk prediction. The study investigated the impact of penalizing group fairness violations on a variety of model performance and group fairness measures. According to the findings of the study, medical evaluations of algorithmic fairness lack the contextual grounding and causal understanding required to understand the mechanisms that contribute to health disparities. Yang et al. (2022) studied the algorithmic fairness and bias mitigation techniques of ML models using COVID-19 diagnosis as the case study and based on the equalized odds metric. The study proposed an adversarial training framework for mitigating biases resulting from data collection or magnified during model development. The study improves the model fairness while keeping the performance not suffering much.

Robustness has been discussed from different angles. Most ML/DL studies considered a robust model as one that had high accuracy, and this is not right because robustness is also related to model security, privacy, certainty, generalizability, and reproducibility (Abuhmed et al., 2021; El-Ghany et al., 2023). The second major dimension of robustness is in security and adversarial training (Silva and Najafirad, 2020; Apostolidis and Papakostas, 2021). Meanwhile, privacy-preserving techniques in the medical domain have been studied in Torkzadehmahani et al. (2022). Zou et al. (2023) reviewed the uncertainty quantification techniques in the medical image domain. Model reproducibility for ML and DL models has been studied by McDermott et al. (2019) and Gundersen et al. (2022). As it can be noticed, the current research literature of RAI is still in the developing phase (Kumar et al., 2021). Hryniewska et al. (2021) reviewed the RAI for DL modeling of medical images for COVID-19 detection. All studies of these authors focused on one specific requirement and neglected the other tightly related requirements. As a result, there is no fully trustworthy AI application in the medical domain. This absence could be a suitable justification for the currently limited roles of AI in medical literature. Sivarajah et al. (2023) urge future research to examine RAI in medicine in a recent special issue on RAI in the medical arena. Liu et al. (2021) investigated the impact of RAI on businesses through the analysis of 25 in-depth interviews with healthcare experts. This investigation concluded that RAI principles can enable healthcare businesses to fully utilize AI.

RAI IN MENTAL DISORDER

In the medical domain, Zhang and Zhang (2023) highlighted the following points that affect AI stakeholders: (i) data quality, (ii) data and algorithmic biases, (iii) opacity of algorithms, (iv) algorithmic safety and security, and (v) accountability of medical errors. Meanwhile, recently, RAI has become a hot research topic. Figure 7 shows that RAI research has gotten too much attention in recent years because the number of research papers is increasing exponentially. The research topics of these papers are mainly concentrated on TAI and RAI. In the mental health domain, the same observation can be noticed because the number of papers is increasing in different TAI domains including robustness, fairness, multimodality, and XAI.

Taking into consideration the above-discussed RAI requirements, we investigate the literature on ML and DL in the field of MHDs in this section. We concentrate on the state-of-the-art crucial problems such as multimodal data fusion (Multimodal-based Applications section), model fairness (Fairness Applications section), and model explainability (XAI Applications section). Mental Disorder Datasets

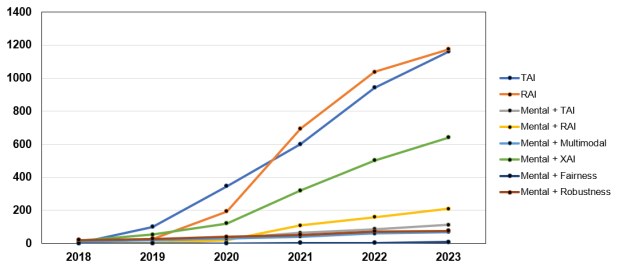

Figure 7: Literature trends in TAI domains. Abbreviations: RAI, responsible AI; TAI, trustworthy AI; XAI, explainable AI.

section also compiles existing datasets in the literature on ML and DL for MHDs.

Multimodal-based applications

In recent years, the merging of ML with mental health research has opened up new avenues for innovative approaches to classifying and diagnosing mental disorders. One particularly promising direction in this field involves the integration of multimodal techniques, where information from various sources, including imaging, textual data, and physiological signals, is combined. This section examines a group of papers on the cutting edge of this transformative landscape, where researchers have used both ML and DL models within a multimodal context. Table 2 contrasts the revised papers based on various aspects, including the specific mental disorder(s) addressed, the datasets employed, and the performance levels attained.

The main goal of Vaz et al. (2023) is to classify anxiety by treating it as a skewed binary classification challenge, using an examination of physiological signals. This research utilizes a dataset that includes data from ECGs, electrodermal activity (EDA), and EMGs, all of which are sourced from the wearable stress and affect detection dataset (Schmidt et al., 2018). What distinguishes this approach from traditional research is its focus on assessing anxiety levels in a natural, uncontrolled setting, thereby characterizing anxiety as a neutral state. This distinctive perspective offers valuable insights with potential implications for the development of improved techniques and strategies for individuals to manage their overall health and well-being. The data analysis process involved feature extraction and selection methods. In the context of feature extraction, each physiological signal was partitioned into segments, each spanning a 5-min duration with a 4-min overlap. Consequently, 15 segments were derived for each subject, resulting in the computation of 109 features for each segment. The subsequent feature selection process encompassed the following three key stages: addressing missing values and conducting variance analysis, evaluating unsupervised correlations to establish an appropriate threshold, and applying a supervised wrapper method. Following these steps, all features underwent normalization through Min-Max scaling. Furthermore, the study tackled the uneven distribution of data by applying four different data balancing strategies. These strategies included random oversampling, synthetic minority oversampling technique (SMOTE), and borderline SMOTE 2. The application of these methods helped in augmenting the data from the underrepresented class. Additionally, various ML algorithms, including adaptive boosting (ADB) and RF, were employed for the purpose of anxiety classification. The most favorable outcome in terms of F1 score, achieving a performance level of 86.4%, was attained through the use of ADB. Safa et al. (2022) embarked on an innovative investigation involving the automated collection of extensive datasets containing depression-related symptoms from the Twitter platform. This study represents the inaugural exploration of biotext, wherein visual elements are associated with pre-established lexicons. Additionally, the study introduces the use of profile headers as a distinctive feature in the prediction of mental disorders. The primary objective was to discern the interplay between depression and linguistic patterns, employing lexicon analysis and natural language processing (NLP) techniques. The proposed architectural framework of the study is structured around three fundamental modules. The initial module encompasses data collection and dataset construction, entailing the aggregation of tweets with diagnosed depression indicators and the implementation of an automated preprocessing pipeline to facilitate subsequent analytical procedures. The dataset underwent refinement through the removal of retweets, emoticons, URLs, special characters, and non-English content. Furthermore, GIF images were converted into the JPG format for profile and header images.

The second module centers on the extraction of pertinent features, fostering cross-examination between textual and visual attributes to identify those that exert the most significant influence. Subsequently, the third module is dedicated to the classification task, aimed at ascertaining the psychological states of users, alongside conducting comparative analyses. To assess the efficacy of the proposed methodology and gauge the performance of various features, a benchmark classifier was established using LR. Furthermore, a set of classifiers was employed for the prediction tasks, encompassing a total of nine classification techniques (i.e. DT, linear SVM, gradient boosting, RF, ridge classifier, AdaBoost, CatBoost, and MLP). Yazdavar et al. (2020) presented a work similar to Safa et al. (2022), while extending their analysis to encompass individual-level demographic attributes. Moreover, they delved into the examination of the attributes associated with posted images, including color palettes, aesthetics, facial expressions, and their correlations with indicators of depressive symptoms. Qureshi et al. (2019) presented a new method that uses a multitask NN to encode different types of data and an attention-based NN for merging various data modalities. These authors developed a pair of models to process audio data, a single model for text, and a trio for video content. These specialized encoders are used for performing tasks connected with depression-level regression and depression-level classification. The proposed architecture comprises three primary components: (i) modality encoders for multitask learning, which takes unimodal features as the input and generate modality embeddings, addressing both regression and classification tasks; (ii) an attention-based fusion network responsible for merging individual modalities; and (iii) a deep neural network (DNN) responsible for producing estimated scores or classifying patients into medically relevant categories, with its output conditioned on the results of the attention fusion network.

The experimental findings of the distress analysis interview corpus - wizard of oz (DAIC-WOZ) dataset (Valstar et al., 2016) demonstrate that multitask representation learning networks exhibit superior performance when contrasted with single-task representation networks. Furthermore, the textual input emerges as a pivotal factor influencing the estimation process. Wei et al. (2023) introduced an ML model designed to facilitate the early detection of autism spectrum disorder (ASD), developmental language disorder (DLD), and global developmental delay (GDD) in children. The study involved the assembly of a dataset comprising 2004

Table 2: Multimodality-based machine learning for mental health disorders.

Ref.	Mental disorder(s)	Dataset	Modalities	Preprocessing steps	Feature engineering	ML/DL model	Feature selection	Validation	Results
Vaz et al. (2023)	Anxiety	WESAD	ECG/EDA/EMG	Butterworth bandpass and Notch filters.	Borderline SMOTE 2.	RF	Unsupervised correlation threshold evaluation	Cross-validation	F1-score: 86.4%
Safa et al. (2022)	Depression	Collected from Twitter	Text/image	Remove retweets, emoticons, URLs, special characters, and non-English language.	N-gram language models, LIWC dictionaries, and bag-of-visual-words.	DT, linear SVM, GB, RF, RC, and AdaBoost	Correlation-based and SVD	None	Acc.: 91% and F1-score: 89%
Qureshi et al. (2019)	Depression	DAIC-WOZ	Audio/Video/Text	None	None	LSTM	None	None	Acc.: 66.66% and F1-score: 0.53
Wei et al. (2023)	Autism spectrum, developmental language, and global developmental delay	Collected from 2004 children	Text	None	Age, sex, and metrics from assessment instruments.	NN, SVM, DT, XGB, and LR	None	Cross-validation	Acc.: 78.3%
Zhang et al. (2020)	Bipolar and depression	BDC and E-DAIC	Audio/video/text	None	Facial landmarks, head pose, eye gaze, and MFCC.	MDDA	PCA and RF	Cross-validation	Bipolar UF1: 0.721 and depression UF1: 0.917
Tang et al. (2020)	Autism spectrum	ABIDE	fMRI/ROI	Slice timing and motion correction, and mean intensity normalization.	Activation map of fMRI and ROI.	ResNet-18	None	None	F1-score: 0.805
Abbas et al. (2023)	Autism spectrum	ABIDE	fMRI/sMRI	ACPC, intensity rescaling, skull stripping, and linear and non-linear image registration.	None	CNN	None	Cross-validation	Acc.: 87.09%
Cai et al. (2020a,b)	Depression	Private	EEG	Notch, finite impulse response, and Kalman filters.	60 linear features and 36 non-linear features.	KNN, DT, and SVM	*T*-test and genetic algorithms	None	Acc.: 86.98%
Mallol-Ragolta et al. (2018)	Posttraumatic stress	EASE	Signal/text	Down sampling, normalizing, and decomposition of signals into tonic and phasic components.	Signal: mean, standard deviation, and number of peaks per second. Text: CSE-T score.	SVR	None	Cross-validation	MSE: 0.79
Sun et al. (2021)	Depression	AVEC 2019 DDS	Audio/video	Split every sequence into segments.	Use the dataset original features, such as MFCC.	Transformers	None	None	CCC: 0.733
Hassan et al. (2023)	Schizophrenia	EEG in schizophrenia	EEG	Butterworth filter and *z*-score normalization.	None	CNN, LR, SVM, RF, and GB	None	Cross-validation	Acc.: 98%
Mellem et al. (2020)	Schizophrenia, bipolar, and attention deficit/hyperactivity	ds000030	MRI/sMRI/clinical scale assessments	None	Use the dataset's original features.	LASSO, elastic net, and RF	Importance-weighted	Cross-validation	Percentage of explained variance: 65-90

Abbreviations: ACPC, Anterior Commissure Posterior Commissure; BDC, bipolar disorder corpus; CCC, concordance correlation coefficient; CNN, convolutional neural network; CSE-T, trauma-focused coping self-efficacy measure; DL, deep learning; DT, decision tree; EASE, engagement arousal self-efficacy; ECG, electrocardiogram; EDA, electrodermal activity; EMG, electromyogram; fMRI, functional magnetic resonance imaging; GB, gradient boosting; KNN, K-nearest neighbor; LIWC, linguistic inquiry and word count; LSTM, long short-term memory; MDDA, multimodal deep denoising autoencoder; ML, machine learning; MSE, mean squared error; NN, neural network; PCA, principle component analysis; RC, ridge classifier; RF, random forest; ROI, regions of interest; SMOTE, synthetic minority over-sampling technique; sMRI, structural MRI; SVD, singular value decomposition; SVM, support vector machine; SVR, support vector regressor; UF1, unweighted F1 score; XGB, eXtreme Gradient Boosting.

children, with data collection spanning the years 2019-2021. Each child underwent assessments such as the ones involving the Gesell developmental scale and the autism behavior checklist. To model development, a total of 14 features were incorporated. These features encompassed variables such as age, gender, and 12 summative metrics derived from assessment instruments. The authors devised an ML-based approach that harnessed easily accessible tools, serving as a decision support system for the early detection of ASD, DLD, or GDD in children. Moreover, they refined the ML model and provided visual representations of its classification process, which improved the model's clinical clarity. This enhancement is aimed at supporting less experienced pediatricians by increasing their diagnostic accuracy.

Lastly, the proposed model was deployed into a web application tailored for clinicians, offering real-time decision support. Five distinct ML algorithms, namely eXtreme Gradient Boosting (XGB), DT, LR, SVM, and NN, were evaluated for their ability to identify children with ASD, DLD, and GDD. Notably, XGB demonstrated the highest accuracy, achieving a rate of 78.3%. Zhang et al. (2020) addressed the disparity and granularity inherent in audiovisual–textual modalities by segregating them into two distinct subgroups. The first subgroup, encompassing the audiovisual modality, operates at the frame level, while the second subgroup, involving the textual modality, functions at the session level. In this system, audio and visual characteristics are processed using a multimodal deep denoising autoencoder on a frame-by-frame basis, then transformed into fixed-length vectors at the session level. To determine the viability of their proposed method, the researchers carried out assessments related to bipolar disorder and depression, making use of the bipolar disorder corpus (Ciftci et al., 2018) and the Extended Distress Analysis Interview Corpus (Ciftci et al., 2018). The latter is an expanded form of the DAIC-WOZ (Valstar et al., 2016), which encompasses semi-clinical interviews.

For the detection of bipolar disorder, the most effective multimodal framework achieved an Unweighted F1 (UF1) score of 0.721, signifying a substantial improvement over unimodal architecture. In the context of depression detection, the UF1, employing 700 units of fused dimension, reached 0.917. These results led to the conclusion that in depression detection, the audiovisual features demonstrate a lesser degree of discriminative capability compared to textual features. Tang et al. (2020) introduced a multimodal framework centered around fMRI to detect ASD. The architecture of this model exhibits the capacity to analyze two distinct forms of activation maps through the amalgamation of various DL networks. The initial input includes a time series activation map for the regions of interest (ROI), created by calculating the correlation matrix between every pair of ROI. The second input is the activation map that integrates fMRI data with the ROI. Using both types of activation maps, which are derived from functional data, improves the classifier's aggregate performance. These features of the data mentioned are used as inputs for two different classifiers: a three-dimensional (3D) ResNet-18 network and a MLP classifier. The feature vectors produced by these classifiers are then combined and used as the input for a series of fully connected layers that lead to the final determination of whether the individual is healthy or shows signs of autism. The research study was conducted using the autism brain imaging data exchange (ABIDE) dataset (Di Martino et al., 2014). The dataset was subjected to a series of preprocessing procedures which included the correction of slice timing, adjustment for motion artifacts, and the normalization of global mean intensity. The proposed framework achieved notable results, demonstrating an F1-score of 0.805 and an impressive recall rate of 95%, which holds considerable significance for the development of computer-assisted diagnostic systems.

Abbas et al. (2023) introduced the deep multimodal neuroimaging framework (DeepMNF) to detect ASD using both fMRI and structural MRI. DeepMNF leverages the integration of spatiotemporal information across different modalities, combining two-dimensional time series data with 3D images. The primary objective is to fuse complementary data, thereby enhancing both group distinctions and homogeneity in the context of ASD diagnosis. DeepMNF incorporates four distinct modalities into a classification system, allowing for the study of spatiotemporal information related to ASD diagnosis, including neuronal activations and morphological features within the brain. The rationale for employing multiple modalities is rooted in the recognition that a multimodal framework can effectively address the heterogeneities inherent in ASD classification by amalgamating complementary information. This addressing, in turn, holds the potential for outperforming single-model frameworks in terms of classification accuracy. Meanwhile, the classification component of the study is anchored in CNN. Initially, the authors explored a configuration comprising a single CNN and a single max-pooling layer for each modality. Subsequently, they expanded the architecture, determining the number of layers based on validation accuracy. Furthermore, the research involved the evaluation of various possible combinations of multimodal options, considering all four modalities. For training and testing, the ABIDE dataset was employed within the proposed framework. The dataset underwent a comprehensive five-stage preprocessing, encompassing Anterior Commissure Posterior Commissure alignment correction, intensity rescaling, skull stripping, and linear and non-linear image registration. Notably, DeepMNF, incorporating four modalities in a multimodal fashion, demonstrated accuracy at 87.09%, surpassing the performance of previously reported studies using the same dataset. Cai et al. (2020a,b) constructed a multimodal model that combined data from three different EEG sources. These sources captured EEG signals under various audio stimuli, including neutral, negative, and positive audio inputs, with the goal of distinguishing between individuals with depression and those without. The dataset of the study comprised 86 participants diagnosed with depression and 92 non-depressed individuals. The EEG data were recorded using three electrodes while subjects were exposed to different audio stimuli.

To prepare the dataset for training, several filtering techniques, such as the Notch and Kalman filters, were applied during the preprocessing stage. Subsequently, both linear and non-linear features were extracted from the EEG signals. After this feature extraction process, a phase for feature fusion was implemented. Following feature fusion, a combination of *t*-tests and genetic algorithms was employed

for feature selection and feature weighting, with the aim of enhancing the overall performance of the recognition framework. The researchers evaluated three different classifiers (i.e. DT, SVM, and KNN) using the prepared dataset. The experimental results revealed a notable improvement in classification accuracy when the data from positive and negative audio stimuli were fused. Particularly, the KNN achieved the highest accuracy, reaching 86.98%. Mallol-Ragolta et al. (2018) introduced a multimodal approach that leverages both skin conductance (SC) physiology and self-reported data obtained from questionnaires to predict the severity of symptoms in individuals with PTSD. To assess the effectiveness of their proposed model, the authors used the Engagement Arousal Self-Efficacy (EASE) dataset (Dhamija and Boult, 2017). This dataset consists of diverse forms of data, such as facial and audio recordings, physiological signals, and self-reported measures, gathered from individuals participating in online trauma-recovery therapy. The SC signal within the dataset underwent preprocessing, which included a series of steps such as applying a low-pass Butterworth filter and decomposing the signal into tonic and phasic components. Furthermore, two sets of features were extracted from the SC signals, while the questionnaire-based text features were rooted in a trauma-focused coping self-efficacy measure (CSE-T). The CSE-T questionnaire inquired about the subjects' perceived capability to cope with various situations. The overall CSE-T score was computed as the average of all questionnaire items, resulting in an absolute value ranging from 1 to 7.

For the system classification task, a support vector regressor was employed, and system performance was assessed by computing the mean squared error between the actual and predicted PTSD symptom severity scores for each subject. The conducted experiments demonstrated that changes in PTSD symptom severity were notably more accurately modeled using this novel multimodal approach compared to relying solely on self-reports or SC data. Sun et al. (2021) developed a depression-level estimation system using a transformer-based framework that relies on two unimodal sources: audio and video data. The primary aim of this framework was to address two critical challenges. First, the framework aimed to process extended sequences of data, a necessity in depression detection. Second, the framework sought to overcome the inherent challenges in DL by leveraging a multimodal learning approach to enhance its performance. To tackle the former, the researchers devised a transformer model capable of processing prolonged sequences. For the latter challenge, the researchers introduced an adaptive late fusion strategy, which involved assigning greater weights to effective modalities or features while reducing the influence of less effective ones. The effectiveness of the proposed framework was evaluated using the AVEC 2019 DDS dataset (Ringeval et al., 2019), which encompasses audio and video modalities, each with various features, such as MFCC from audio. The evaluation metric employed to assess depression detection was the concordance correlation coefficient (CCC), which quantifies the agreement between actual and predicted Patient Health Questionnaire Depression Scale scores. The experimental outcomes indicated that the proposed framework outperformed the current state-of-the-art methods, achieving a CCC score of 0.733.

Finding a subset of EEG channels to detect schizophrenia is the main challenge handled by Hassan et al. (2023). The proposed approach to overcome this challenge involved the development of a channel selection mechanism rooted in the performance analysis of a CNN when considering individual EEG channels from distinct brain regions. The chosen channels were subsequently amalgamated and fused through CNN, followed by the integration of ML classifiers such as SVM and LR. The evaluation of the model was carried out using a publicly available EEG dataset (Olejarczyk and Jernajczyk, 2017), comprising EEG signals from 28 subjects, with half of them diagnosed with paranoid schizophrenia. Notably, the data collection process entailed the use of 19 electrodes, corresponding to 19 channels. In preparation for feature extraction, a preliminary preprocessing stage involved the application of a Butterworth filter, and subsequent z-score normalization was executed. In addition, to circumvent the need for hand-crafted feature extraction, CNN was harnessed as a feature extractor. For the classification phase, an array of classifiers was explored. The experimental results revealed that the most effective combination involved the use of CNN for feature extraction in tandem with an LR classifier. This combination achieved an impressive accuracy rate of 98% for testing not dependent on the subject, using EEG signals from only three channels. Mellem et al. (2020) introduced a multimodal framework designed for the prediction of symptom severity in psychiatric disorders, specifically targeting anhedonia, dysregulated mood, and anxiety. To estimate symptom severity, the researchers employed three distinct regression algorithms, including two linear models (i.e. least absolute shrinkage and selection operator and elastic net) and a non-linear model (i.e. RF). The evaluation of these models and their associated features was conducted using the ds000030 dataset (Poldrack et al., 2016), a dataset characterized by its rich array of features, encompassing clinical scale assessments and MRI data.

To identify the most predictive features, a data-driven feature selection approach was developed. This approach led to the suggestion of seven combinations of feature types, along with the creation of six distinct symptom severity scores based on clinical scales. Subsequently, an extensive array of combinations involving feature sets, symptom severity scores, and regression models was explored and assessed (i.e. 126 different combinations). The results of this comprehensive approach demonstrated a notable enhancement in modeling effectiveness, explaining a substantial proportion of variance across the three symptom domains. In particular, the approach achieved a level of explanation ranging from 65% to 90% of variance, in contrast to the baseline of 22% when not using the feature selection approach.

Fairness applications

ML models rely on data, which means that bias can be included in the model decisions or even in the data itself. The source model's bias is related to the unbalanced distribution of data among different subgroups, as well as the fact that the model's predictions are influenced by some protected factors (Paul et al., 2020). Optimally, algorithms would possess

comprehensive access to patient's electronic health record (EHR) data to construct representative models for disease diagnosis, negative effect prediction, or continuing treatment recommendations (Chen et al., 2019). Enhancing fairness in prediction has been promoted by modifying models through regularization, restrictions, and representation learning.

These efforts can be roughly classified as fairness techniques that rely on models. Some studies have implemented data preprocessing techniques to mitigate discrimination. However, constraining the model's complexity or altering the training data to enhance fairness might negatively impact the model's predictive accuracy. It can be challenging to justify sacrificing forecast accuracy in favor of fairness, especially when predictions have a major impact on critical decisions. Specifically, employing posthoc corrective techniques that rely on randomizing predictions is ethically unacceptable in clinical activities due to its potential for decreased predictive accuracy. As shown in Table 3, the analysis of fairness in predictive models should consider model bias, model variation, and outcome noise prior to imposing fairness requirements (Chen et al., 2018). When dealing with mental illnesses, there is strong evidence that bias is not only related to ML diagnosis models or data processing but also related to practitioner's mental training and stigma (Peris et al., 2008).Peris et al. (2008) explored the relationship between the stigma of mental illness and clinical decision-making. The study examines both implicit and explicit biases toward mental illness among 1539 participants with varying levels of mental health training, including mental health professionals, healthcare/social service specialists, undergraduate students, and the public. The findings reveal that those with mental health training have more positive implicit and explicit evaluations of individuals with mental illness. Explicit biases, but not implicit, were found to predict more negative prognoses for patients, whereas implicit biases (and not explicit) predicted a higher likelihood of over-diagnosing. This finding suggests the importance of considering both implicit and explicit measures when studying the stigma of mental illness and its effects on clinical care. The study underlines the complex ways in which biases can influence the clinical process, highlighting the need for clinicians to be aware of their own biases, which can affect their decisions regarding diagnoses and prognoses.

Adarsh et al. (2023) investigated the importance of early detection in diagnosing depression based on patient's social media posts. The study raised the issue of biased data used in related works due to unequal data distribution. Their work addressed the imbalance in participation across different age groups and demographics using the one-shot decision approach. The study introduced an ensemble model that combines SVM and KNN, ensuring data classification without bias. This model achieved a classification accuracy of 98.05% and applied LIME explainability approach. This application was on the label-corrected data to find the keywords that contributed to classifying the posts into two categories, i.e. with and without suicidal thoughts. The study also discusses the challenges in accurately diagnosing depression and the potential risks of suicidal thoughts or actions if they are not addressed. Park et al. (2022) investigate the potential gender biases in mobile phone-based mental health prediction algorithms. With the backdrop that approximately one in five American adults experience mental illness annually, the study emphasizes the growing significance of mobile phone apps leveraging AI for mental health assessments. However, concerns arise as various AI technologies, including facial recognition, have shown biases related to age, gender, and race. The study's objective was to understand the gender bias susceptibility in ML models used for mobile mental health assessments and to explore methods to reduce this bias without compromising accuracy. Using a dataset of 55 participants, the research revealed that while the highest accuracy achieved was 78.57%, there was a significant gender-based performance disparity in the algorithm. This gender disparity was significantly reduced after implementing the disparate-impact remover approach. The findings underscore the importance of algorithmic auditing in mental health assessment algorithms, emphasizing the need for fairness and accuracy in such tools. Tanqueray et al. (2022) explored the intersection of gender norms and social robotics, particularly in the context of peripartum depression (PPD) screening. This study emphasizes the importance of understanding social structures and potential divisions before simplifying them into algorithms. The study involves semi-structured interviews with experts, exploring gender norms in current medical practices surrounding PPD screening. It also examines the role of power and stakeholders in the development of new technologies, stressing the need for inclusivity and representation. The research highlights the necessity of a relational approach in designing robots for PPD screening, considering the unique life experiences and backgrounds of pregnant women. The study emphasizes the complexities of integrating technology in healthcare, advocating for the need to understand the social dynamics and power structures

Table 3: Fairness-based machine learning studies for mental health disorders.

Ref.	ML/DL model	Fairness technique	Data	Dataset	Task
Adarsh et al. (2023)	Ensemble ML	Preprocessing	Text	Reddit communities	Depression/suicidal thoughts
Park et al. (2022)	MPN, SVM, LR, KNN, and RF	Preprocessing	Questionnaires	Phone metadata	Depression, stress, flourishing, and loneliness
Chen et al. (2019)	LR	Preprocessing	Clinical notes	MIMIC-III	Psychiatric readmissions
Paul et al. (2020)	DL network	Preprocessing and in-processing	Text	OSMI Mental Health Survey	Likelihood mental health treatment

Abbreviations: DL, deep learning; KNN, K-nearest neighbor; ML, machine learning; OSMI, open sourcing mental health; RF, random forest; SVM, support vector machine.

to ensure gender fairness in social robotics Tanqueray et al. (2022). Chen et al. (2019) explore the potential of AI in addressing disparities in healthcare. The study examines two case studies using ML algorithms on clinical and psychiatric notes to predict intensive care unit mortality and 30-day psychiatric readmission, focusing on race, gender, and insurance payer type as proxies for socioeconomic status. It reveals that clinical note topics are heterogeneous with respect to these demographics, reflecting known clinical findings. The research highlights differences in prediction accuracy and machine bias, particularly with respect to gender, insurance type, and insurance policy for psychiatric readmission. The study also emphasizes the importance of understanding and addressing algorithmic biases in mental healthcare AI applications, advocating for a cooperative relationship between clinicians and AI to improve patient care and reduce disparities. Training and representation alteration (TARA) is a novel method proposed by Paul et al. (2020) to enforce AI fairness with respect to sensitive variables. This method employs dual preprocessing and in-processing approaches. The first approach involves representation learning alteration *via* adversarial independence to suppress bias-inducing dependence of data representation on sensitive factors. The second approach involves training set alteration *via* intelligent augmentation using generative models for fine control of sensitive factors related to underrepresented populations *via* domain adaptation and latent space manipulation. The study shows that TARA significantly debiases baseline models with considerable gains in overall accuracy, presents novel conjunctive debiasing metrics, and emphasizes the ability of these metrics to assess the effectiveness of the suggested methods.

XAI applications

XAI is one of the crucial components of the TAI-based system. In this section, we review a few studies that explored XAI in an ML-based approach for MHDs. Table 4 shows the summary of the reviewed studies. Ahmed et al. (2022b) presented a novel approach for enhancing Internet-delivered psychological treatment through NLP and DL models. The study focused on the challenge of emotion segmentation in psychological texts, where emotional biases can lead to incorrect analysis. The solution offered by the authors to the challenge is an assistance tool for psychologists, which leverages an NLP-based method to create word embeddings using an emotional lexicon. This creation is followed by attention-based deep clustering to visualize the emotional aspects of patient-authored texts. The authors' approach involves expanding patient-authored text using synonymous semantic expansion and clustering the semantic representation with an Explainable Attention Network-based Deep adaptive Clustering model (EANDC). They employ similarity metrics for text selection and curriculum-based optimization for better learning explainability. Experimental results showed that the EANDC

Table 4: XAI-based machine learning studies for mental health disorders.

Ref.	ML/DL model	XAI technique	Data	Dataset	Task
Wang et al. (2021)	LSTM	SHAP	Time series	Private	Mental illness risk
Nguyen and Byeon (2022)	DNN	LIME	Questionnaires	36,258 patients	Depression
Nemesure et al. (2021)	XGBoost	SHAP	Biomedical and demographic	4184 patients	MDD/GAD
Ahmed et al. (2022b)	DNN	Explainable attention network visualizations	Social media text	15,044 social posts	Psychological disorders
Wang et al. (2021)	DNN	SHAP	Sensory data	2069 days of patient records	Anxiety, depression, and schizophrenia
Ghosh et al. (2023)	LSTM+CNN	Feature-level explainability	Social media text posts	13,678 samples	Depression detection
Alam and Kapadia (2020)	N-gram language model	Linguistic inquiry and word count explainability	Social media text posts	2423 patients	Posttraumatic stress disorder
Ellis et al. (2022)	CNN-LSTM	Spatial and spectral explainability	EGG signals	101 patients	Schizophrenia
Zanwar et al. (2023)	MentalRoBERTa	Feature-level explainability	Social media text posts	8675 users	Attention-deficit hyperactivity disorder, anxiety, bipolar disorder, depression, and psychological stress
Han et al. (2022)	BERT attention network	Visualization of attention weights	Social media text posts	4208 users	Depression detection
Toleubay et al. (2023)	Logical neural network	Rule-based explainability	Medical notes	48 patient sessions	Anxiety, depression, suicidal thoughts, and schizophrenia
Yang et al. (2019)	Graph attention network	Activation maps	sMRI	106 subjects	Bipolar disorder
Nguyen and Byeon (2023)	Stacking ensemble of LR, LGBM, KNN, RF, and ET	LIME	Numerical features	526 patients	Parkinson's disease depression

Abbreviations: BERT, bidirectional encoder representations from transformers; CNN, convolutional neural network; DL, deep learning; DNN, deep neural network; ET, extra trees; GAD, generalized anxiety disorder; KNN, K-nearest neighbor; LGBM, light gradient boosting machine; LSTM, long short-term memory; MDD, major depressive disorder; ML, machine learning; RF, random forest; sMRI, structural MRI; XAI, explainable AI.

model, particularly the attention method with a bidirectional LSTM architecture, achieved a significant 0.81 ROC in blind tests. The model also helped in symptom recognition of mental disorders, proving that the synonym expansion based on the emotion lexicon increases accuracy. The EANDC model can assist mental health professionals in understanding and treating mental health conditions more effectively.

Wang et al. (2021) employed deep learning neural networks (DLNNs) to assess the illness risk of mental disorders in Nanjing, potentially influenced by various air pollutants and meteorological conditions. The study leverages the SHAP method to interpret the predictions of the proposed DLNNs, emphasizing the non-linear association between outpatient visits for mental disorders and environmental stressors. Through enhanced model interpretability, the study identifies and elucidates low-frequency, high-impact non-linear risk factors and their potential interactions, contributing to the broader understanding of air pollution epidemiology and its impact on mental health. Nguyen and Byeon (2022) focused on the implications of the COVID-19 pandemic on the mental health of the elderly. The study developed a DNN model to predict depression in the elderly based on 22 social characteristics using data from the 2020 Community Health Survey of the Republic of Korea, which comprised 97,230 adults over the age of 60 years. To offer explainability on the model's predictions, the model was further integrated with a LIME-based explainable model. The model achieved a prediction accuracy of 89.92%, with an F1-score of 92%, precision of 93.55%, and recall of 97.32%, according to the study. The findings indicate the COVID-19 pandemic's large impact on the likelihood of depression in the elderly, as well as the promise of the explainable DNN model in the early detection and treatment of depression in patients. Nemesure et al. (2021) predicted major depressive disorder (MDD) and generalized anxiety disorder (GAD) using basic medical examination and an ML approach. The study used 59 biomedical and demographic features from 4184 undergraduate students who underwent a general health screening and psychiatric assessment for MDD and GAD. The results showed that the model could predict MDD and GAD with an AUC of 0.67 and 0.73, respectively. Important predictors for MDD included satisfaction with living conditions and having public health insurance, while for GAD, the top predictors were up-to-date vaccinations and marijuana use.

Wang et al. (2021) investigated into how environmental factors like air pollution and meteorological conditions can be linked to the risk of mental disorders using DLNNs. The authors used a permutation-based SHAP method to interpret the DLNN's predictions. The study revealed that air pollutants like NO_2, SO_2, and CO are significant predictors of outpatient visits for mental disorders, indicating their substantial impact on mental health risks. The study combined data on daily outpatient visits from two major hospitals in Nanjing with environmental data from air quality monitoring stations, capturing the non-linear relationship between environmental stressors and mental health outcomes. The model's performance was robust, and SHAP analysis provided insights into the variable importance and interaction effects. Notably, the findings suggest that high levels of NO_2 increase the risk of MHDs, while SO_2 and CO showed mixed effects. The study holds potential implications for public health policies, emphasizing the importance of air pollution control in mitigating mental health risks. The authors also acknowledge limitations of the study such as the inability to measure individual exposure levels to pollution and the broader categorization of mental disorders, suggesting avenues for more focused future research. The paper concludes that accurate modeling of illness risk, combined with interpretability, is crucial for effective air pollution management strategies to address mental health concerns.

Mental disorder datasets

In this section, we surveyed the existing mainly open-access datasets for mental disorders. Each dataset is described in terms of the abbreviation, the size, the mental disorder, whether the dataset is open access or not, whether the dataset is time series or not, the modalities, the task including detection, prediction, and diagnosis, and the access link if the dataset is open access. Table 5 includes the description of 43 datasets. All these datasets are open access except Andrzejak et al. (2001), Tasnim et al. (2022), Yoon et al. (2022), Cummins et al. (2023), and Nigg et al. (2023), where we could not find a direct download link. Most datasets are single modality, where the dataset contains one type of data. For example, Ramírez-Dolores et al. (2022) presents a questionnaire, Olejarczyk and Jernajczyk (2017), Andrzejak et al. (2001), Mane (2023), Nasrabadi et al. (2020), Obeid and Picone (2016), Shoeb (2009), Temko et al. (2015), Gupta et al. (2018), Brinkmann et al. (2016), Detti et al. (2020), and Stevenson et al. (2019) present EEGs; Yates et al. (2017), Cohan et al. (2018), and Turcan and McKeown (2019) are about social media posts; Singh et al. (2022) and Villa-Pérez and Trejo (2023) present tweets; Yoon et al. (2022) contains videos; Mauriello et al. (2021) presents short messages; Tasnim et al. (2022) and Cummins et al. (2023) present speeches; Wang et al. (2016), Tanaka et al. (2021), and ADHD-200 (2023) contain MRIs; El Haouij et al. (2018), Schmidt et al. (2018), Jakobsen et al. (2020), Hosseini et al. (2022), and Kang et al. (2023) present sensory data; and Edgar et al. (2002) and Wu et al. (2017) are about genetic data. Using single modality data to build an ML model results in a model which cannot provide personalized or customized medicine. The main reason for this limitation is because in real environments, medical experts use multimodal data about the patient to provide a deeper representation and understanding of the patient's conditions. We collected a set of multimodal datasets with different formats which support the building of more robust and TAI models. For example, Albrecht et al. (2019) collected EEG, medication, and behavioral modalities for schizophrenia, and Valstar et al. (2013, 2014) are speech and video modalities for depression. In addition, Shen et al. (2022) collected the two modalities of speech and text for depression, Siless et al. (2020) collected MRI, clinical, and behavioral modality for anxiety and depression, and Cai et al. (2020a,b) contains the two modalities of EEG and speech for depression. Likewise, Hicks et al. (2021) had the demographics, neuropsychological tests, and activity and heart rates modalities

Table 5: Mental health disorders' open-access datasets.

Ref.	Year	Abbreviation	Size	Disease	Open access	Time series	Modalities	Access link	Task
Albrecht et al. (2019)	2019	MPRC	80	Schizophrenia	Yes	Yes	EEG, medication, and behavioral	link	Diagnosis
Valstar et al. (2013)	2013	AVEC2013	292	Depression	Yes	No	Audio and video	link	Detection
Valstar et al. (2014)	2014	AVEC2014		Depression	Yes	No	Audio and video	-	Detection
Ramírez-Dolores et al. (2022)	2022	-	316	Thermal comfort, stress, and anxiety	Yes	No	Questionnaire	link	Detection
Andrzejak et al. (2001)	2001	-	11,500	Epileptic seizure	No	No	EEG	-	Diagnosis
Cohan et al. (2018)		SMHD	Millions	Multiple disorders	Yes	No	Social media posts	link	Diagnosis
Valstar et al. (2016)	2014	DAIC-WOZ	189	Anxiety, depression, and stress	Yes	No	Audio and video	link	Diagnosis
Wang et al. (2020)	2013	AViD-Corpus	340	Depression	Yes	Yes	Audio and video	link	Diagnosis
Shen et al. (2022)	2022	EATD-Corpus	162	Depression	Yes	No	Audio and text	link	Diagnosis
Yoon et al. (2022)	2022	D-Vlog	916	Depression	No	No	Video	-	Diagnosis
Tasnim et al. (2022)	2022	DEPAC	2674	Depression and anxiety	No	Yes	Audio	-	Detection
Cummins et al. (2023)	2023	RADAR-MDD	585	Depression	No	No	Audio	-	Assessment
Siless et al. (2020)	2015	BANDA	225	Anxiety and depression	Yes	Yes	MRI, clinical, and behavioral data	link	Diagnosis
Cai et al. (2020a,b)	2015	MODMA	160	Depression	Yes	Yes	EEG and audio	link	Diagnosis
Mane (2023)	2023	-	112	Stress	Yes	No	EEG	link	Detection
ADHD-200 (2023)	2011	ADHD200	200	ADHD	Yes	Yes	fMRI	link	Diagnosis
Hicks et al. (2021)	2021	HYPERAKTIV	103	ADHD	Yes	Yes	Demographics, neuropsychological tests, and activity and heart rates	link	Diagnosis
Nigg et al. (2023)	2022	Oregon ADHD-1000	1000	ADHD	No	Yes	DNA, EEG, MRI, psychophysiological, psychosocial, clinical, and functional	link	Diagnosis
Nasrabadi et al. (2020)	2020	-	121	ADHD	Yes	No	EEG	link	Diagnosis
Villa-Pérez and Trejo (2023)	2023	-	3000	Nine disorders	Yes	Yes	Text (Twitter)	link	Diagnosis
Yates et al. (2017)	2017	RSDD	9000	Depression	Yes	No	Text (posts)	link	Detection
Turcan and McKeown (2019)	2019	Dreaddit	190,000	Stress	Yes	No	Text (posts)	link	Identification
Schmidt et al. (2018)	2018	WESAD	15	Stress	Yes	Yes	Sensor data	link	Detection
Jaiswal et al. (2020)	2020	Ulm-TSST	105	Stress	Yes	Yes	ECG, EDA, and BPM	link	Detection
Mauriello et al. (2021)	2021	SAD	6850	Stress	Yes	No	Text (SMS sentences)	link	Classification
Obeid and Picone (2016)	2016	TUH	642	Seizure	Yes	Yes	EEG	link	Detection
El Haouij et al. (2018)	2018	AffectiveROAD	10	Stress	Yes	Yes	Wearables	link	Detection
Singh et al. (2022)	2022	Twitter-STMHD	33,860	Eight disorders	Yes	No	Text (Twitter)	link	Diagnosis
Edgar et al. (2002)	2011	GSE26415	42	Autism spectrum	Yes	No	Genetic	link	Prediction
MaND (2023)	2011	MaND	225	Depression	Yes	No	MRI and CT	link	Diagnosis
Shoeb (2009)	2009	CHB-MIT	182	Intractable seizures	Yes	Yes	EEG	link	Detection and treatment
Temko et al. (2015)	2015	KUMCS	-	Seizures	Yes	Yes	EEG	link	Detection
Gupta et al. (2018)	2016	-	10	Seizures	Yes	Yes	EEG	link	Detection
Brinkmann et al. (2016)	2016	AESSPC	8003	Seizures	Yes	Yes	EEG	link	Prediction
Detti et al. (2020)	2020	SSED	14	Epilepsy	Yes	Yes	EEG	link	Prediction
Wang et al. (2016)	2016	SchizConnect	1392	Schizophrenia	Yes	No	MRI	link	Detection
Jakobsen et al. (2020)	2020	PSYKOSE	54	Schizophrenia	Yes	Yes	Sensory	link	Detection

Table 5: Continued.

Ref.	Year	Abbreviation	Size	Disease	Open access	Time series	Modalities	Access link	Task
Larivière et al. (2021)	2015	ENIGMA	6023	Epilepsy	Yes	No	MRI and DTI	link	Detection
Tanaka et al. (2021)	2016	SRPBS	2414	Autism, depression, schizophrenia, and obsessive-compulsive disorder	No	Yes	MRI	link	Detection
Hosseini et al. (2022)	2021	-	15	Stress	Yes	Yes	Sensory	link	Detection
Kang et al. (2023)	2023	K-EmoPhone	77	Stress	Yes	Yes	Sensory	link	Detection
Wu et al. (2017)	2017	SZDB	Multiple	Schizophrenia			Genetic	link	Diagnosis
Stevenson et al. (2019)	2018	-	87	Seizures	Yes	Yes	EEG	link	Detection
Olejarczyk and Jernajczyk (2017)	2017	EEG in schizophrenia	28	Schizophrenia	Yes	Yes	EEG	link	Classification
Poldrack et al. (2016)	2016	ds000030	272	Schizophrenia, bipolar, and ADHD	Yes	Yes	MRI, sMRI, and clinical scale assessments	link	Diagnosis
Ciftci et al. (2018)	2018	BDC	46	Bipolar	Upon request	Yes	Audio/video	link	Classification
Di Martino et al. (2014)	2014	ABIDE	1112	Autism spectrum	Yes	Yes	fMRI/ROI	link	Diagnosis
Mayer et al. (2013)	2012	COBRE	147	Schizophrenia	Yes	No	sMRI/rs-fMRI/phenotypic	link	Diagnosis
Dhamija and Boult (2017)	2017	EASE	110	Posttraumatic stress	Yes	Yes	Audio, video, text, and physiological signals	link	Diagnosis

for ADHD, Nigg et al. (2023) had the DNA, EEG, MRI, psychophysiological, psychosocial, clinical, and functional data for ADHD. Jaiswal et al. (2020) had the ECG, EDA, and respiration, and heart rate modalities for stress, Kempton et al. (2023) had the two imaging modalities of MRI and computed tomography (CT) for the depression, and Larivière et al. (2021) had the two imaging modalities of MRI and diffusion tensor imaging for epilepsy. Another important dimension for analysis is the longitudinal nature of the dataset. Most datasets are not time series data which means that each patient has been represented by a single record or single observation. Some datasets tracked patients over time (Schmidt et al., 2018; Albrecht et al., 2019; Tasnim et al., 2022; Kang et al., 2023). We noticed that depression disease has many datasets, but we also noticed that most datasets have small samples of data. For comprehensive datasets concentrated on specific type of data, Li et al. (2019) provided a comprehensive survey of the speech dataset for mental disorders. In addition, Wong et al. (2023) provided a survey of the EEG datasets for seizure detection and prediction, and Garg (2023) surveyed the mental health analysis process based on social media posts datasets.

CHALLENGES AND FUTURE RESEARCH DIRECTIONS

The literature of TAI in general has been reviewed in this study. In addition, we reviewed and evaluated the literature of TAI in the medical disorder domain considering the TAI requirements. We found a great shortage in MHD literature regarding TAI requirements. These results could be the reason for the little applications of AI-based systems in the real environment of MHDs. This review provides the researchers in the AI and mental disorders domains with an accurate analysis and evaluation of the current literature. In this section, we highlight the current limitations and possible directions for enhancements of AI-based systems for MHDs. The following is a list of possible research directions considering the previously discussed literature:

- There is no unified definition of the TAI term and its requirements. There is no unified framework for applying TAI principles. There are no best practices and machine learning operations guidelines to implement an ecosystem for supporting RAI use in the medical context (Sivarajah et al., 2023). There is an urgent need to develop deep research in RAI in the medical domain to develop medically trustworthy applications of AI.
- Available datasets in the MHD domain are mostly not sufficient and have a small number of patients. MHDs need time series and multimodal data. Most available datasets are not prepared for building personalized ML models that are suitable for building personalized CDSSs. Existing datasets have short sequences which are not suitable for advanced DL models such as transformers. New data augmentation algorithms like stable diffusion and GAN can be used to generate samples. Data preprocessing, division into train/validation/test, data labeling, data balancing, etc. could result in biased datasets which over optimize

the results and result in overfitted models that are difficult to generalize.

- The available datasets must be examined for potential data bias and uncertainty. Most datasets are biased due to the inclusion and exclusion criteria employed to choose respondents. The combination of patient's EHR data leads to a robust ML model that gives individualized and patient-tailored decisions. In the current literature, the study of multimodal time series data is insufficient. Medical data are usually multimodal in nature including neuroimages such as MRI and PET, structured data such as lab tests, graph data such as medicine and diseases, text data such as medical reports, EEG data, sensory data, etc. Fusion of these data results in comprehensive datasets that could be used to build personalized models and for discovering new biomarkers.
- All ML/DL models either neglected TAI at all or concentrated on specific dimension of TAI. For example, some studies concentrated on model robustness, especially related to security issues. At the same time, the study neglected the model transparency which is very much related to security. Some studies concentrated on model fairness and neglected model robustness, but increasing model fairness could result in lower performance models. Some studies focused on providing comprehensive XAI features and made the proposed model fully transparent. However, this transparency affects the model's security and user's data privacy. No study in the literature implemented a comprehensive TAI framework that can balance all these, possibly conflicting, requirements.
- ML/DL models can be used in real environment only if they are stable. Model reproducibility could be used to measure the generalizability of models. Reproducibility can be achieved by (i) developing standard datasets, (ii) applying standardized data management, (iii) using standard cross-validation techniques to validate the models, (iv) testing models on different datasets, and (v) publicly releasing the code (El-Sappagh et al., 2023).
- Building a federated ML model requires the integration of data from different sources. In medical domain, these data could have different standards, use different terminologies, and have different data structures. Data interoperability solutions can solve this challenging issue arising from the above differences. This solution facilitates the modeling of distributed CDSSs for use in distributed medical systems. Furthermore, the model can be trained on a dataset from one source and tested on data from another. This allows us to develop a model with data from one study and confirm it with data from another. This has the potential to improve the model's generalizability. Notably, finding standard mapping rules between different datasets is crucial. Standard ontologies and terminologies, such as Systematized Nomenclature of Medicine Clinical Term, International Classification of Disease, Logical Observation Identifiers Names and Codes, and Unified Medical Language System, and unified data formats such as Fast Healthcare Interoperability Resources and OpenEHR can help in handling this challenge (El-Sappagh et al., 2019).
- XAI features of ML/DL models are crucial necessities for medical applications. There are many limitations of the current XAI literature. In the MHD domain, there are many modalities to consider, and building consistent multimodal XAI features is important. This provides the domain experts with XAI features with different formats like feature importance, case-based reasoning, fuzzy and natural language explanations, visual explanations, etc. However, explanations should be context based. In other words, the provided explanations should depend on the experience of the user. Testing the causality and the uncertainty of the provided XAI is also crucial. Domain experts must participate in the design process of the XAI features of the AI-based system. In addition, using natural language models like GPT-4 to provide interactive XAI features needs to be explored (Salahuddin et al., 2022; Ali et al., 2023). Graph neural networks can combine multimodal data and knowledge bases for providing interactive explainability. There is an urgent need to provide new techniques for evaluating the XAI methods which combine human-centered and quantitative evaluations.
- Fusion of different data formats such as MRI, EEG, text, video, audio, and clinical data is crucial to improve the accuracy of the ML/DL models. The fusion of multimodal time series data needs complex models to understand these data. Ensemble and hybrid model optimization is a critical field to export to build stable and robust ML/DL models. AutoML techniques can be used to optimize the ML pipeline and select the best model architectures. The optimization can include the selection of the best XAI features that have the best fidelity with the main ML/DL models.
- Knowledge-guided ML/DL architectures are state-of-the-art to improve the performance and the interpretability of ML/DL models. Interpretable knowledge-based systems are based on domain experts, standard clinical practice guideline, and literature knowledge. This knowledge can be represented as Bayesian networks, knowledge graphs, IF–THEN rules, or semantic ontologies. Integrating this knowledge with the data-driven AI models improves the learning and performance of the resulting ML/DL model and improves its interpretability.
- The TAI techniques and tools are immature. For example, fairness tools, techniques, and metrics face many dilemmas, including the lack of a unified definition of fairness (Mehrabi et al., 2019). Most fairness definitions concentrated on the equality where each individual or group is given the same resources, but definitions neglected the equity where individuals and groups given the needed resources to succeed. Model robustness is based on providing countermeasures of adversarial attacks, but there is no universal method to measure the level of robustness of ML model against unknown attacks. Model explainability faces many challenges such as multimodal XAI, measuring XAI uncertainty, context-based XAI, human-in-the-loop in XAI design, interactive XAI, etc. These challenges need further investigation.
- The role of non-functional TAI parts such as regulation, standardization, certification, education, awareness, and accountability needs further study (Sivarajah et al., 2023).

CONCLUSION

In this study, we explored the literature of ML and DL models in the MHD domain. We evaluate the literature of TAI in MHD detection and prediction. We investigated and explored the use of multimodal data to build customized models. We evaluated the models' robustness, fairness, and transparency compared to the TAI standards and guidelines. Most existing datasets for MHDs have been collected. The main results of this study are as follows: (i) existing ML/DL models are not robust because these studies did not implement suitable security and privacy measures, have not performed any external validation, and have used small datasets; (ii) existing literature is not fair because no evaluation has been done for data and algorithmic bias, and no suitable measures have been found to remove these biases; (iii) existing methods have used limited XAI features because no multimodal, multiformat, context-sensitive, robust, and interactive XAI has been implemented; and (iv) existing studies were mainly depended on single modality such as images or EEG signals. Based on the analysis of the literature, we highlighted the current limitations of the literature of AI-based models in the MHD domain, and we suggested future research directions that could improve this domain. The investigation presented in this work is crucial to build medically relevant and acceptable AI-based models that could play an effective role in enhancing the quality of mental health management.

FUNDING

This work received funding from the King Salman Center for Disability Research through Research Group no. KSRG-2022-101 (funder ID: http://dx.doi.org/10.13039/501100019345).

AUTHOR CONTRIBUTIONS

SE, WN, MA, and TA designed the review. All authors provided feedback on survey items. SE and WN wrote the first draft of the manuscript. SE, WN, MA, and TA read and approved the final manuscript.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGMENTS

The authors extend their appreciation to the King Salman Center for Disability Research for funding this work through Research Group no. KSRG-2022-101 (funder ID: http://dx.doi.org/10.13039/501100019345).

REFERENCES

Abbas S.Q., Chi L. and Chen Y.P.P. (2023). DeepMNF: deep multimodal neuroimaging framework for diagnosing autism spectrum disorder. *Artif. Intell. Med.*, 136, 102475. 10.1016/J.ARTMED.2022.102475.

Abd Rahman R., Omar K., Noah S.A.M., Danuri M.S.N.M. and Al-Garadi M.A. (2020). Application of machine learning methods in mental health detection: a systematic review. *IEEE Access*, 8, 183952-183964.

Abuhmed T., El-sappagh S. and Alonso J.M. (2021). Robust hybrid deep learning models for Alzheimer's progression detection. *Knowl. Based Syst.*, 213, 106688. 10.1016/j.knosys.2020.106688.

Adarsh V., Arun Kumar P., Lavanya V. and Gangadharan G.R. (2023). Fair and explainable depression detection in social media. *Inf. Process. Manag.*, 60, 103168. 10.1016/j.ipm.2022.103168.

ADHD-200. (2023). *ADHD-200-Webpage*. http://fcon_1000.projects.nitrc.org/indi/adhd200/. Accessed September 9, 2023.

Adhikari A., Wenink E., van der Waa J., Bouter C., Tolios I. and Raaijmakers S. (2022). Towards FAIR explainable AI: a standardized ontology for mapping XAI solutions to use cases, explanations, and AI systems. In: *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA' 22)*, Corfu, Greece, 11 July 2022; pp. 562-568.

Ahmed A., Aziz S., Toro C.T., Alzubaidi M., Irshaidat S., Serhan H.A., et al. (2022a). Machine learning models to detect anxiety and depression through social media: a scoping review. *Comput. Methods Programs Biomed. Update*, 2, 100066.

Ahmed U., Srivastava G., Yun U. and Lin J.C.W. (2022b). EANDC: an explainable attention network based deep adaptive clustering model for mental health treatment. *Future Gener. Comput. Syst.*, 130, 106-113. 10.1016/j.future.2021.12.008.

Alam M.A.U. and Kapadia D. (2020). LAXARY: A Trustworthy Explainable Twitter Analysis Model for Post-traumatic Stress Disorder Assessment. http://arxiv.org/abs/2003.07433.

Albahri A. S., Duhaim A.M., Fadhel M.A., Alnoor A., Baqer N.S., Alzubaidi L., et al. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. *Inf. Fusion*, 96, 156-191.

Albrecht M.A., Waltz J.A., Cavanagh J.F., Frank M.J. and Gold J.M. (2019). Increased conflict-induced slowing, but no differences in conflict-induced positive or negative prediction error learning in patients with schizophrenia. *Neuropsychologia*, 123, 131-140. 10.1016/j.neuropsychologia.2018.04.031.

Alghadeer S.M., Alhossan A.M., Al-Arifi M.N., Alrabiah Z.S., Ali S.W., Babelghaith S.D., et al. (2018). Prevalence of mental disorders among patients attending primary health care centers in the capital of Saudi Arabia. *Neurosci. J.*, 23, 238-243.

Ali S., Abuhmed T., El-Sappagh S., Muhammad K., Alonso-Moral J.M., Confalonieri R., et al. (2023). Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion*, 99, 101805. 10.1016/j.inffus.2023.101805.

Altwaijri Y., Kazdin A.E., Al-Subaie A., Al-Habeeb A., Hyder S., Bilal L., et al. (2023). Lifetime prevalence and treatment of mental disorders in Saudi youth and adolescents. *Sci. Rep.*, 13, 1-13. 10.1038/s41598-023-33005-5.

Andrzejak R.G., Lehnertz K., Mormann F., Rieke C., David P., and Elger C.E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. *Phys. Rev. E.*, 64, 61907.

Apostolidis K.D. and Papakostas G.A. (2021). A survey on adversarial deep learning robustness in medical image analysis. *Electronics (Switzerland)*, 10, 2132. 10.3390/electronics10172132.

Arji G., Erfannia L., Alirezaei S. and Hemmat M. (2023). A systematic literature review and analysis of deep learning algorithms in mental disorders. *Inform. Med. Unlocked.*, 40, 101284.

Arrieta A.B., Díaz-Rodríguez N., Ser J.D., Bennetot A., Tabik S., Barbado A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58, 82-115.

Ashokan A. and Haas C. (2021). Fairness metrics and bias mitigation strategies for rating predictions. *Inf. Process. Manag.*, 58, 102646. 10.1016/j.ipm.2021.102646.

Baniecki H., Sobieski B., Bombiński P., Szatkowski P. and Biecek P. (2023). Hospital length of stay prediction based on multi-modal data towards trustworthy human-AI collaboration in radiomics. In: *Proceedings of the Artificial Intelligence in Medicine: 21st International Conference on Artificial Intelligence in Medicine (AIME'23)*, Portorož, Slovenia, 12-15 June 2023; pp. 65-74, 2023, 10.1007/978-3-031-34344-5_9.

Bellamy R.K.E., Kuntal Dey K., Hind M., Hoffman S.C., Houde S., Kannan K., et al. (2019). AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM J. Res. Dev.*, 63, 1-1.

Biggio B. and Roli F. (2018). Wild patterns: ten years after the rise of adversarial machine learning. *Pattern Recognit.*, 84, 317-331.

Brendel W., Rauber J. and Bethge M. (2017). Decision-based adversarial attacks: reliable attacks against black-box machine learning models. *arXiv preprint, arXiv:1712.04248.*

Brinkmann B.H., Wagenaar J., Abbot D., Adkins P., Bosshard S.C., Chen M., et al. (2016). Crowdsourcing reproducible seizure forecasting in human and canine epilepsy. *Brain*, 139, 1713-1722.

Buruk B., Ekmekci P.E. and Arda B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Med. Health Care Philos.*, 23, 387-399. 10.1007/s11019-020-09948-1.

Cai H., Gao Y., Sun S., Li N, Tian F., Xiao H., et al. (2020a). MODMA dataset: a multi-modal open dataset for mental-disorder analysis. *arXiv preprint arXiv:2002.09283*.

Cai H., Qu Z., Li Z., Zhang Y., Hu X. and Hu B. (2020b). Feature-level fusion approaches based on multimodal EEG data for depression recognition. *Inf. Fusion.*, 59, 127-138. 10.1016/J.INFFUS.2020.01.008.

Carey A.N. and Wu X. (2022). The causal fairness field guide: perspectives from social and formal sciences. *Front. Big Data*, 5, 1-40. 10.3389/fdata.2022.892837.

Caton S. and Haas C. (2020). Fairness in machine learning: a survey. *ArXiv*, 1-33.

Chang H., Nguyen T.D., Murakonda S.K., Kazemi E. and Shokri R. (2020). On adversarial bias and the robustness of fair machine learning. *arXiv preprint, arXiv:2006.08669.*

Chen I., Johansson F.D. and Sontag D. (2018). Why is my classifier discriminatory? *Adv. Neural. Inf. Process. Syst.*, 31, 1-12.

Chen I.Y., Szolovits P. and Ghassemi M. (2019). Can AI help reduce disparities in general medical and mental health care? *AMA J. Ethics.*, 21, 167-179.

Chiu C.H., Chung H.W., Chen Y.J., Shi Y. and Ho T.Y. (2023). Toward fairness through fair multi-exit framework for dermatological disease diagnosis. In: *Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Part III*, Vancouver, BC, Canada, 8-12 October 2023; pp. 97-107.

Cho G., Yim J., Choi Y., Ko J. and Lee S.H. (2019). Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investig.*, 16, 262-269. 10.30773/pi.2018.12.21.2.

Chou Y.L., Moreira C., Bruza P., Ouyang C. and Jorge J. (2022). Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. *Inf. Fusion*, 81, 59-83.

Christensen M.K., Lim C.C.W., Saha S., Plana-Ripoll O., Cannon D. and Presley F., et al. (2020). The cost of mental disorders: a systematic review. *Epidemiol. Psychiatr. Sci.*, 29, e161.

Chung J. and Teo J. (2022). Mental health prediction using machine learning: taxonomy, applications, and challenges. *Appl. Comput. Intell. Soft Comput.* 2022, 1-19.

Ciftci E., Kaya H., Gulec H. and Salah A.A. (2018). The Turkish audio-visual bipolar disorder corpus. In: *Proceedings of the 2018 1st Asian Conference on Affective Computing and Intelligent Interaction, ACII Asia 2018*, Beijing, China, 20-22 May 2018. 10.1109/ACIIASIA.2018.8470362.

Cohan A., Desmet B., Yates A., Soldaini L., MacAvaney S. and Goharian N. (2018). SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, 20-26 August 2018; pp. 1485-1497.

Crockett K.A., Gerber L., Latham A. and Colyer E. (2021). Building trustworthy AI solutions: a case for practical solutions for small businesses. *IEEE Trans. Artif. Intell.*, 4(1), 778-791. 10.1109/tai.2021.3137091.

Cuijpers P., Javed A. and Bhui K. (2023). The WHO world mental health report: a call for action. *Br. J. Psychiatry*, 222, 227-229.

Cummins N., Dineley J., Conde P., Matcham F., Siddi S., Lamers F., et al. (2023). Multilingual markers of depression in remotely collected speech samples: a preliminary analysis. *J. Affect. Disord.*, 341, 128-136. 10.1016/j.jad.2023.08.097.

de Bardeci M., Ip C.T. and Olbrich S. (2021). Deep learning applied to electroencephalogram data in mental disorders: a systematic review. *Biol. Psychol.*, 162, 108117.

de Laat P.B. (2021). Companies committed to responsible AI: from principles towards implementation and regulation? *Philos. Technol.*, 34, 1135-1193.

Detti P., Vatti G. and Zabalo Manrique de Lara G. (2020). EEG synchronization analysis for seizure prediction: a study on data of noninvasive recordings. *Processes*, 8, 846.

Dhamija S. and Boult T.E. (2017). Exploring contextual engagement for trauma recovery. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, HI, USA, 21-26 July 2017; pp. 2267-2277. 10.1109/CVPRW.2017.281.

Di Martino A., Yan C.G., Li Q., Denio E., Castellanos F.X., Alaerts K. et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry.*, 19, 659-667. 10.1038/MP.2013.78.

Díaz-Rodríguez N., Lamas A., Sanchez J., Franchi G., Donadello I., Tabik S., et al. (2022). EXplainable Neural-Symbolic Learning (X-NeSyL) methodology to fuse deep learning representations with expert knowledge graphs: the MonuMAI cultural heritage use case. *Inf. Fusion*, 79, 58-83. 10.1016/j.inffus.2021.09.022.

Dignum V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics Inf. Technol.*, 20, 1-3. 10.1007/s10676-018-9450-z.

Dignum V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in A Responsible Way*, Springer Nature.

Ding W., Abdel-Basset M., Hawash H. and Ali A.M. (2022). Explainability of artificial intelligence methods, applications and challenges: a comprehensive survey. *Inf. Sci.*, 615, 238-292.

Drukker K., Chen W., Gichoya J., Gruszauskas N., Kalpathy-Cramer J., Koyejo S., et al. (2023). Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *J. Med. Imaging*, 10, 061104. 10.1117/1.jmi.10.6.061104.

Edgar R., Domrachev M. and Lash A.E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic. Acids Res.*, 30, 207-210.

El Haouij N., Poggi J.M., Sevestre-Ghalila S., Ghozi R. and Jaïdane M. (2018). AffectiveROAD system and database to assess driver's attention. In: *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, Pau, France, 09 April 2018; pp. 800-803.

El-Ghany S.A., Azad M. and Elmogy M. (2023). Robustness fine-tuning deep learning model for cancers diagnosis based on histopathology image analysis. *Diagnostics*, 13, 1-18. 10.3390/diagnostics13040699.

Ellis C.A., Sattiraju A., Miller R. and Calhoun V. (2022). Examining effects of schizophrenia on EEG with explainable deep learning models. In: *Proceedings of the IEEE 22nd International Conference on Bioinformatics and Bioengineering, BIBE 2022*, Institute of Electrical and Electronics Engineers Inc., Taichung, Taiwan, 07-09 November 2022; pp. 301-304. 10.1109/BIBE55377.2022.00068.

El-Sappagh S., Ali F., El-Masri S., Kim K., Ali A. and Kwak K.S. (2019). Mobile health technologies for diabetes mellitus: current state and future challenges. In: *IEEE Access*, 7. 10.1109/ACCESS.2018.2881001.

El-Sappagh S., Alonso-Moral J.M., Abuhmed T., Ali F. and Bugarín-Diz A. (2023). Trustworthy artificial intelligence in Alzheimer's disease: state of the art, opportunities, and challenges. *Artif. Intell. Rev.*, 56, 1-148. 10.1007/s10462-023-10415-5.

Exforsys. (2011). *What is Monkey Testing*. https://www.exforsys.com/tutorials/testing-types/monkey-testing.html. Accessed August 23, 2023.

Garcia-Ceja E., Riegler M., Nordgreen T., Jakobsen P., Oedegaard K.J. and Tørresen J. (2018). Mental health monitoring with multimodal sensing and machine learning: a survey. *Pervasive Mob. Comput.*, 51, 1-26.

Garg M. (2023). Mental health analysis in social media posts: a survey. *Arch. Comput. Methods Eng.*, 30, 1819-1842.

GBD 2019 Mental Disorder Collaborators. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Psychiatry*, 9, 137-150.

Ghosh T., Al Banna M.H., Al Nahian M.J., Uddin M.N., Kaiser M.S. and Mahmud M. (2023). An attention-based hybrid architecture with explainability for depressive social media text detection in Bangla. *Expert. Syst. Appl.*, 213, 119007. 10.1016/j.eswa.2022.119007.

González-Gonzalo C., Thee E.F., Klaver C.C.W., Lee A.Y., Schlingemann R.O., Tufail A., et al. (2021). Trustworthy AI: closing the gap between development and integration of AI systems in ophthalmic practice. *Prog. Retin. Eye Res.*, 90, 101034. 10.1016/j.preteyeres.2021.101034.

Graham S., Depp C., Lee E.E., Nebeker C., Tu X., Kim H.C., et al. (2019). Artificial intelligence for mental health and mental illnesses: an overview. *Curr. Psychiatry Rep.*, 21, 1-18.

Greco C.M., Simeri A., Tagarelli A. and Zumpano E. (2023). Transformer-based language models for mental health issues: a survey. *Pattern Recognit. Lett.*, 167, 204-211.

Grgić-Hlača N., Zafar M.B., Gummadi K.P. and Weller A. (2018). Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning. *Proc. AAAI Conf. Artif. Intell.*, 32(1), 1-10. 10.1609/aaai.v32i1.11296.

Gundersen O.E., Shamsaliei S. and Isdahl R.J. (2022). Do machine learning platforms provide out-of-the-box reproducibility? *Future Gener. Comput. Syst.*, 126, 34-47. 10.1016/j.future.2021.06.014.

Gupta A., Singh P. and Karlekar M. (2018). A novel signal modeling approach for classification of seizure and seizure-free EEG signals. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 26, 925-935.

Han S.H. and Choi H.J. (2022). Checklist for validating trustworthy AI. In: *Proceedings of the 2022 IEEE International Conference on Big Data and Smart Computing, BigComp 2022*, Daegu, Korea, Republic of Korea, 17-20 January 2022; pp. 391-394. 10.1109/BigComp54360.2022.00088.

Han S., Mao R. and Cambria E. (2022). Hierarchical Attention Network for Explainable Depression Detection on Twitter Aided by Metaphor Concept Mappings. In: *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, 12-17 October 2022; pp. 94-104.

Hassan F., Hussain S.F. and Qaisar S.M. (2023). Fusion of multivariate EEG signals for schizophrenia detection using CNN and machine learning techniques. *Inf. Fusion.*, 92, 466-478. 10.1016/J.INFFUS.2022.12.019.

Hicks S.A., Stautland A., Fasmer O.F., Førland W., Hammer H.L., Halvorsen P., et al. (2021). HYPERAKTIV: an activity dataset from patients with attention-deficit/hyperactivity disorder (ADHD). In: *Proceedings of the 12th ACM Multimedia Systems Conference*, Istanbul Turkey, 22 September 2021; pp. 314-319.

Holzinger A., Dehmer M., Emmert-Streib F., Cucchiara R., Augenstein I., Del Ser J., et al. (2022). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Inf. Fusion*, 79, 263-278. 10.1016/J.INFFUS.2021.10.007.

Hosseini S., Gottumukkala R., Katragadda S., Bhupatiraju R.T., Ashkar Z., Borst C.W., et al. (2022). A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Sci. Data.*, 9, 255.

Hryniewska W., Bombiński P., Szatkowski P., Tomaszewska P., Przelaskowski A. and Biecek P. (2021). Checklist for responsible deep learning modeling of medical images based on COVID-19 detection studies. *Pattern Recognit.*, 118, 108035. 10.1016/j.patcog.2021.108035.

Institute for Health Metrics and Evaluation. (2023). *Global Health Data Exchange (GHDx)*. https://vizhub.healthdata.org/gbd-results/. Accessed September 12, 2023.

Iyortsuun N.K., Kim S.H., Jhon M., Yang H.J. and Pant S. (2023). A review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare*, 11, 285.

Jaiswal M., Bara C.P., Luo Y., Burzo M., Mihalcea R. and Provost E.M. (2020). MuSE: a multimodal dataset of stressed emotion. In: *Proceedings of the LREC 2020—12th International Conference on Language Resources and Evaluation, Conference*, Marseille, France, May 2020; pp. 1499-1510.

Jakobsen P., Garcia-Ceja E., Stabell L.A., Oedegaard K.J., Berle J.O., Thambawita V., et al. (2020). Psykose: a motor activity database of patients with schizophrenia. In: *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, Rochester, MN, USA, 28-30 July 2020; pp. 303-308.

Jiang J., Kahai S. and Yang M. (2022). Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *Int. J. Hum. Comput. Stud.*, 165, 102839.

Jobin A., Ienca M. and Vayena E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intell.*, 1, 389-399. 10.1038/s42256-019-0088-2.

Johnson B. and Brun Y. (2022). Fairkit-learn: a fairness evaluation and comparison toolkit. In: *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings (ICSE' 22)*, Ottawa, ON, Canada, 19 October 2022; pp. 70-74.

Joshi G., Walambe R. and Kotecha K. (2021). A review on explainability in multimodal deep neural nets. *IEEE Access*, 9, 59800-59821. 10.1109/ACCESS.2021.3070212.

Kang S., Choi W., Park C.Y., Cha N. Kim A., Khandoker A.H., et al. (2023). K-EmoPhone: a mobile and wearable dataset with in-situ emotion, stress, and attention labels. *Sci. Data.*, 10, 351.

Kaur D., Uslu S., Rittichier K.J. and Durresi A. (2023). Trustworthy artificial intelligence: a review. *ACM Comput. Surv.*, 55, 2. 10.1145/3491209.

Kempton M.J., Salvador Z., Munafò M.R., Geddes J.R., Simmons A., Frangou S., et al. (2023). Major Depressive Disorder Neuroimaging Database (MaND). [Online]. Retrieved September 21, 2024, from https://sites.google.com/site/depressiondatabase/.

Khare S.K., March S., Barua P.D., Gadre V.M. and Acharya U.R. (2023). Application of data fusion for automated detection of children with developmental and mental disorders: a systematic review of the last decade. *Inf. Fusion*, 99, 101898.

Kim D, Song Y., Kim S., Lee S., Wu Y., Shin J., et al. (2023). How should the results of artificial intelligence be explained to users? – Research on consumer preferences in user-centered explainable artificial intelligence. *Technol. Forecast. Soc. Change*, 188, 122343.

Kovalchuk S.V., Kopanitsa G.D., Derevitskii I.V., Matveev G.A. and Savitskaya D.A. (2022). Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. *J. Biomed. Inform.*, 127, 104013.

Kozodoi N., Jacob J. and Lessmann S. (2021). Fairness in credit scoring: assessment, implementation and profit implications. *Eur. J. Oper. Res.*, 293, 1083-1094. 10.1016/j.ejor.2021.06.023.

Kumar P., Dwivedi Y.K. and Anand A. (2021). Responsible artificial intelligence (AI) for value formation and market performance in healthcare: the mediating role of patient's cognitive engagement. *Inf. Syst. Front.*, 25, 2197-2220. 10.1007/s10796-021-10136-6.

Larivière S., Paquola C., Park B.Y., Royer J., Wang Y., Benkarim O., et al. (2021). The ENIGMA toolbox: multiscale neural contextualization of multisite neuroimaging datasets. *Nat. Methods*, 18, 698-700.

Leichtmann B., Humer C., Hinterreiter A., Streit M. and Mara M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Comput. Human Behav.*, 139, 107539. 10.1016/j.chb.2022.107539.

Leike J., Martic M., Krakovna V., Ortega P.A., Everitt T., Lefrancq A., et al. (2017). AI safety gridworlds. *arXiv preprint, arXiv:1711.09883*.

Li Y., Lin Y., Ding H. and Li C. (2019). Speech databases for mental disorders: a systematic review. *Gen. Psychiatr.*, 32, e100022.

Li B., Qi P., Liu B., Di S., Liu J., Pei J., et al. (2023). Trustworthy AI: from principles to practices. *ACM Comput. Surv.*, 55, 1-46.

Liang W., Tadesse G.A., Ho D., Fei-Fei L., Zaharia M., Zhang C., et al. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. *Nat. Mach. Intell.*, 4, 669-677. 10.1038/s42256-022-00516-1.

Liefgreen A., Weinstein N., Wachter S. and Mittelstadt B. (2023). Beyond ideals: why the (medical) AI industry needs to motivate behavioural change in line with fairness and transparency values, and how it can do it. *AI Soc.* 10.1007/s00146-023-01684-3.

Liu Q., Li P., Zhao W., Cai W., Yu S. and Leung V.C.M. (2018). A survey on security threats and defensive techniques of machine learning: a data driven view. *IEEE Access*, 6, 12103-12117.

Liu R., Gupta S. and Patel P. (2021). The application of the principles of responsible AI on social media marketing for digital health. *Inf. Syst. Front.*, 25, 2275-2299. 10.1007/s10796-021-10191-z.

Liu H., Wang Y., Fan W., Liu X., Li Y., Jain S., et al. (2022). Trustworthy AI: a computational perspective. *ACM Trans. Intell. Syst. Technol.*, 14, 1-55. 10.1145/3546872.

Liu S. and Vicente L.N. (2022). Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. *Comput. Manag. Sci.*, 19, 513-537.

Lofstrom H., Lofstrom T., Johansson U. and Sonstrod C. (2023). Calibrated explanations: with uncertainty information and counterfactuals. *arXiv preprint, arXiv:2305.02305.*

Lucieri A., Bajwa M.N., Braun S.A., Malik M.I., Dengel A. and Ahmed S. (2022). ExAID: a multimodal explanation framework for computer-aided diagnosis of skin lesions. *Comput. Methods Programs Biomed.*, 215, 106620.

Maalouf F.T., Alamiri B., Atweh S., Becker A.E., Cheour M., Darwish H., et al. (2019). Mental health research in the Arab region: challenges and call for action. *The Lancet Psychiatry*, 6(11), 961-966.

Ma J., Schneider L., Lapuschkin S., Achtibat R., Duchrau M., Krois J., et al. (2022). Towards trustworthy AI in dentistry. *J. Dent. Res.*, 101, 1263-1268. 10.1177/00220345221106086.

Malhotra A. and Jindal R. (2022). Deep learning techniques for suicide and depression detection from online social media: a scoping review. *Appl. Soft. Comput.*, 130, 109713.

Mallol-Ragolta A., Dhamija S. and Boult T.E. (2018). A multimodal approach for predicting changes in PTSD symptom severity. In: *ICMI 2018 – Proceedings of the 2018 International Conference on Multimodal Interaction*, CO, Boulder, USA, 02 October 2018; pp. 324-333. 10.1145/3242969.3242981.

MaND. (2023). *Major Depressive Disorder Neuroimaging Database (MaND)*. https://sites.google.com/site/depressiondatabase/. Accessed September 10, 2023.

Mane M. (2023). *An EEG recordings dataset for mental stress detection*. Mendeley Data, V1. 10.17632/wnshbvdxs2.1.

Mauriello M.L., Lincoln T., Hon G., Simon D., Jurafsky D. and Paredes P. (2021). SAD: a stress annotated dataset for recognizing everyday stressors in SMS-like conversational systems. In: *Proceedings of the Conference on Human Factors in Computing Systems*, New York, NY, USA, 08 May 2021; pp. 1-7. 10.1145/3411763.3451799.

Mayer A.R., Ruhl D., Merideth F., Ling J., Hanlon F.M., Bustillo J., et al. (2013). Functional imaging of the hemodynamic sensory gating response in schizophrenia. *Hum. Brain Mapp.*, 34, 2302-2312.

McDermott M.B., Wang S., Marinsek N., Ranganath R., Foschini L. and Ghassemi M. (2021). Reproducibility in machine learning for health research: Still a ways to go. Science Translational Medicine, 13(586), p.eabb1655.

Mehdiyev N., Majlatow M. and Fettke P. (2023). Communicating uncertainty in machine learning explanations: a visualization analytics approach for predictive process monitoring. *arXiv preprint, arXiv:2304.05736.*

Mehrabi N., Morstatter F., Saxena N., Lerman K. and Galstyan A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint, arXiv:1908.09635. ACM Comput. Surv.*, 54, 1-35.

Mehrabi N., Morstatter F., Saxena N., Lerman K. and Galstyan A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54, 1-35. 10.1145/3457607.

Mellem M.S., Liu Y., Gonzalez H., Kollada M., Martin W.J. and Ahammad P. (2020). Machine learning models identify multimodal measurements highly predictive of transdiagnostic symptom severity for mood, anhedonia, and anxiety. *Biol. Psychiatry. Cogn. Neurosci. Neuroimaging.*, 5, 56-67. 10.1016/J.BPSC.2019.07.007.

Mendelson T. and Eaton W.W. (2018). Recent advances in the prevention of mental disorders. *Soc. Psychiatry Psychiatr. Epidemiol.*, 53, 325-339.

Miller T. (2019). Explainable artificial intelligence: what were you thinking? In: *Artificial Intelligence: For Better or Worse*, Futureleaders, Australia; pp. 19-38.

Moher D., Liberati A., Tetzlaff J., Altman D.G. and Group P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.*, 6, e1000097.

Moulouel K., Chibani A. and Amirat Y. (2023). Ontology-based hybrid commonsense reasoning framework for handling context abnormalities in uncertain and partially observable environments. *Inf. Sci.*, 631, 468-486. 10.1016/j.ins.2023.02.078.

Moura I., Teles A., Silva F., Viana D., Coutinho L., Barros F., et al. (2020). Mental health ubiquitous monitoring supported by social situation awareness: a systematic review. *J. Biomed. Inform.*, 107, 103454.

Moustapha M., Marelli S. and Sudret B. (2022). Active learning for structural reliability: survey, general framework and benchmark. *Struct. Saf.*, 96, 102174.

Nasrabadi A.M., Allahverdy A., Samavati M. and Mohammadi M.R. (2020). EEG data for ADHD/control children. *IEEE Dataport.* 10.21227/rzfh-zn36.

National Alliance in Mental Illness. (2023). *Mental Health by the Numbers*. https://nami.org/mhstats. Accessed September 12, 2023.

National Center for Health Statistics. (2022). *Mental Health Treatment Among Adults Aged 18–44: United States, 2019–2021*. https://www.cdc.gov/nchs/products/databriefs/db444.htm. Accessed September 12, 2023.

National Institute of Mental Health. (2023). *Mental Illness*. https://www.nimh.nih.gov/health/statistics/mental-illness. Accessed September 12, 2023.

Nemesure M.D., Heinz M.V., Huang R. and Jacobson N.C. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Sci. Rep.*, 11, 1980.

Nguyen H.V. and Byeon H. (2022). Explainable deep-learning-based depression modeling of elderly community after COVID-19 Pandemic. *Mathematics*, 10, 1-10. 10.3390/math10234408.

Nguyen H.V. and Byeon H. (2023). Prediction of Parkinson's disease depression using LIME-based stacking ensemble model. *Mathematics*, 11, 708. 10.3390/math11030708.

Nicolae M.I., Sinn M., Tran M.N., Rawat A., Wistuba M., Zantedeschi V., et al. (2018). Adversarial robustness toolbox v0.4.0. *ArXiv*, 1-34.

Nigg J.T., Karalunas S.L., Mooney M.A., Wilmot B., Nikolas M.A., Martel M.M., et al. (2023). The Oregon ADHD-1000: a new longitudinal data resource enriched for clinical cases and multiple levels of analysis. *Dev. Cogn. Neurosci.*, 60, 101222.

Nyrup R. and Robinson D. (2022). Explanatory pragmatism: a context-sensitive framework for explainable medical AI. *Ethics Inf. Technol.*, 24, 13.

Obeid I. and Picone J. (2016). The temple university hospital EEG data corpus. *Front. Neurosci.*, 10, 196.

Olejarczyk E. and Jernajczyk W. (2017). EEG in schizophrenia. *RepOD*. 10.18150/repod.0107441.

Panigutti C., Beretta A., Fadda D., Giannotti F., Pedreschi D., Perotti A., et al. (2023). Co-design of human-centered, explainable AI for clinical decision support. *ACM Trans. Interact. Intell. Syst*., 13, 1-35.

Panigutti C., Perotti A. and Pedreschi D. (2020). Doctor XAI an ontology-based approach to black-box sequential data classification explanations. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*' 20)*, Barcelona, Spain, 27-30 January 2020; pp. 629-639. 10.1145/3351095.3372855.

Park J., Arunachalam R., Silenzio V. and Singh V.K. (2022). Fairness in mobile phone-based mental health assessment algorithms: exploratory study. *JMIR Form. Res.*, 6, e34366. 10.2196/34366.

Paul W., Hadić A., Joshi N.J., Alajaji F. and Burlina P. (2020). TARA: training and representation alteration for AI fairness and domain generalization. *Neural. Comput.*, 34, 716-753.

Peris T.S., Teachman B.A. and Nosek B.A. (2008). Implicit and explicit stigma of mental illness: links to clinical care. *J. Nerv. Ment. Dis.*, 196, 752-760. 10.1097/NMD.0b013e3181879dfd.

Pessach D. and Shmueli E. (2021). Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings. *Expert. Syst. Appl.*, 185, 115667. 10.1016/j.eswa.2021.115667.

Pfohl S.R., Foryciarz A. and Shah N.H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. *J. Biomed. Inform.*, 113, 103621. 10.1016/j.jbi.2020.103621.

Poldrack R.A., Congdon E., Triplett W., Gorgolewski K.J., Karlsgodt K.H., Mumford J.A., et al. (2016). A phenome-wide examination of neural and cognitive function. *Sci. Data.*, 3, 160110. 10.1038/SDATA.2016.110.

Qayyum A., Qadir J., Bilal M. and Al-Fuqaha A. (2021). Secure and robust machine learning for healthcare: a survey. *IEEE Rev. Biomed. Eng.*, 14, 156-180. 10.1109/RBME.2020.3013489.

Qureshi S.A., Saha S., Hasanuzzaman M., Dias G. and Cambria E. (2019). Multitask representation learning for multimodal estimation of depression level. *IEEE Intell. Syst.*, 34, 45-52. 10.1109/MIS.2019.2925204.

Raghunathan A., Xie S.M., Yang F., Duchi J.C. and Liang P. (2019). Adversarial training can hurt generalization. *arXiv preprint, arXiv:1906.06032.*

Ramírez-Dolores C., Lugo-Ramírez L.A., Hernández-Cortaza B.A., Alcalá G., Lara-Valdés J. and Andaverde J. (2022). Dataset on

thermal comfort, perceived stress, and anxiety in university students under confinement due to COVID-19 in a hot and humid region of Mexico. *Data. Brief.*, 41, 107996.

Rasheed K., Qayyum A., Ghaly M., Al-Fuqaha A., Razi A. and Qadir J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: a survey. *Comput. Biol. Med.*, 149, 106043. 10.1016/j.compbiomed.2022.106043.

Ringeval F., Schuller B., Valstar M., Cummins N., Cowie R., Tavabi L., et al. (2019). AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: *AVEC 2019 – Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop, co-located with MM 2019*, New York, NY, USA, 15 October 2019; pp. 3-12. 10.1145/3347320.3357688.

Rivera M.J., Teruel M.A., Mate A. and Trujillo J. (2022). Diagnosis and prognosis of mental disorders by means of EEG and deep learning: a systematic mapping study. *Artif. Intell. Rev.*, 55, 1-43.

Rong Y., Leemann T., Nguyen T.T., Fiedler L., Qian P., Unhelkar V., et al. (2023). Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 4, pp. 2104-2122, April 2024. 10.1109/TPAMI.2023.3331846.

Rojat T., Puget R., Filliat D., Del Ser J., Gelin R. and Díaz-Rodríguez N. (2021). Explainable artificial intelligence (XAI) on TimeSeries data: a survey. *arXiv preprint, arXiv:2104.00950.*

Rožanec J.M. and Mladenić D. (2021). Semantic XAI for contextualized demand forecasting explanations. *arXiv preprint, arXiv:2104.00452.*

Rožanec J.M., Fortuna B. and Mladenić D. (2022). Knowledge graph-based rich and confidentiality preserving explainable artificial intelligence (XAI). *Inf. Fusion*, 81, 91-102. 10.1016/j.inffus.2021.11.015.

Saeed W. and Omlin C. (2023). Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. *Knowl. Based Syst.*, 263, 110273.

Safa R., Bayat P. and Moghtader L. (2022). Automatic detection of depression symptoms in twitter using multimodal analysis. *J. Supercomput.*, 78, 4709-4744. 10.1007/S11227-021-04040-8.

Salahuddin Z., Woodruff H.C., Chatterjee A. and Lambin P. (2022). Transparency of deep neural networks for medical image analysis: a review of interpretability methods. *Comput. Biol. Med.*, 140, 105111.

Saranya A. and Subhashini R. (2023). A systematic review of explainable artificial intelligence models and applications: recent developments and future trends. *J. Decis. Anal.*, 7, 100230.

Schmidt P., Reiss A., Duerichen R. and Van Laerhoven K. (2018). Introducing WeSAD, a multimodal dataset for wearable stress and affect detection. In: *ICMI 2018 – Proceedings of the 2018 International Conference on Multimodal Interaction*, Boulder, CO, USA, 02 October 2018; pp. 400-408. 10.1145/3242969.3242985.

Schoonderwoerd T.A.J., Jorritsma W., Neerincx M.A. and van den Bosch K. (2021). Human-centered XAI: developing design patterns for explanations of clinical decision support systems. *Int. J. Hum. Comput. Stud.*, 154, 102684. 10.1016/j.ijhcs.2021.102684.

Serban A., Van Der Blom K., Hoos H. and Visser J. (2021). Practices for engineering trustworthy machine learning applications. In: *Proceedings of the 2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI, WAIN*, Madrid, Spain, 30-31 May 2021; pp. 97-100. 10.1109/WAIN52551.2021.00021.

Shen Y., Yang H. and Lin L. (2022). Automatic depression detection: an emotional audio-textual corpus and a Gru/BiLSTM-based model. In: *Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Singapore, 23-27 May 2022; pp. 6247-6251.

Shneiderman B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Trans. Interact. Intell. Syst.*, 10, 1-31. 10.1145/3419764.

Shoeb A.H. (2009). *Application of machine learning to epileptic seizure onset detection and treatment*. Massachusetts Institute of Technology.

Siala H. and Wang Y. (2022). SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. *Soc. Sci. Med.*, 296, 114782. 10.1016/j.socscimed.2022.114782.

Siless V., Hubbard N.A., Jones R., Wang J., Lo N., Bauer C.C.C., et al. (2020). Image acquisition and quality assurance in the Boston adolescent neuroimaging of depression and anxiety study. *Neuroimage Clin.*, 26, 102242. 10.1016/j.nicl.2020.102242.

Silva S.H. and Najafirad P. (2020). Opportunities and challenges in deep learning adversarial robustness: a survey. *arXiv preprint, arXiv:2007.00753*, 1-20.

Singh M., Ghalachyan G., Varshney K.R. and Bryant R.E. (2021). An empirical study of accuracy, fairness, explainability, distributional robustness, and adversarial robustness. *arXiv preprint, arXiv:2109.14653.*

Singh A.K., Arora U., Shrivastava S., Singh A., Shah R.R. and Kumaraguru P. (2022). Twitter-STMHD: an extensive user-level database of multiple mental health disorders. *Proc Int AAAI Conf Weblogs Soc Media.*, 16, 1182-1191.

Sivarajah U., Wang Y., Olya H. and Mathew S. (2023). Responsible artificial intelligence (AI) for digital health and medical analytics. *Inf. Syst. Front.*, 25, 2117-2122. 10.1007/s10796-023-10412-7.

Smuha, N.A. (2019). The EU approach to ethics guidelines for trustworthy artificial intelligence. Computer Law Review International, 20(4), pp. 97-106.

Sovrano F., Vitali F. and Palmirani M. (2020). Modelling GDPR-compliant explanations for trustworthy AI. In Electronic Government and the Information Systems Perspective: 9th International Conference, EGOVIS 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 9 (pp. 219-233). Springer International Publishing.

Squires M., Tao X., Elangovan S., Gururajan R., Zhou X., Acharya U.R., et al. (2023). Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Inform.*, 10, 1-19.

Stanford P.K. (2002). The manifest connection: causation, meaning, and David Hume. *J. Hist. Philos.*, 40, 339-360.

Stevenson N.J., Tapani K., Lauronen L. and Vanhatalo S. (2019). A dataset of neonatal EEG recordings with seizure annotations. *Sci. Data.*, 6, 1-8.

Su C., Xu Z., Pathak J. and Wang F. (2020). Deep learning in mental health outcome research: a scoping review. *Transl. Psychiatry*, 10, 116.

Sui J., Jiang R., Bustillo J. and Calhoun V. (2020). Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biol. Psychiatry*, 88, 818-828.

Sun H., Liu J., Chai S., Qiu Z., Lin L., Huang X., et al. (2021). Multi-modal adaptive fusion transformer network for the estimation of depression level. *Sensors.*, 21, 4764. 10.3390/S21144764.

Tan W., Chen L., Zhang Y., Xi J., Hao Y., Jia F., et al. (2022). Regional years of life lost, years lived with disability, and disability-adjusted life-years for severe mental disorders in Guangdong Province, China: a real-world longitudinal study. *Glob. Health Res. Policy*, 7, 1-14.

Tanaka S.C., Yamashita A., Yahata N., Itahashi T., Lisi G., Yamada T., et al. (2021). A multi-site, multi-disorder resting-state magnetic resonance image database. *Sci. Data.*, 8, 227.

Tang M., Kumar P., Chen H. and Shrivastava A. (2020). Deep multimodal learning for the diagnosis of autism spectrum disorder. *J. Imaging*., 6, 47. 10.3390/JIMAGING6060047.

Tanqueray L., Paulsson T., Zhong M., Larsson S. and Castellano G. (2022). Gender fairness in social robotics: exploring a future care of peripartum depression. In: *HRI' 22: Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, IEEE Press, Sapporo, Hokkaido, Japan, 07 March 2022; pp. 598-607.

Tasnim M., Ehghaghi M., Diep B. and Novikova J. (2022). DEPAC: a corpus for depression and anxiety detection from speech. In: *Proceedings of the CLPsych 2022 – 8th Workshop on Computational Linguistics and Clinical Psychology*, Seattle, USA, July 2022; pp. 1-16. 10.18653/v1/2022.clpsych-1.1.

Temko A., Sarkar A. and Lightbody G. (2015). Detection of seizures in intracranial EEG: UPenn and Mayo clinic's seizure detection challenge. In: *Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, Milan, Italy, 25-29 August 2015; pp. 6582-6585.

The Lancet Global Health. (2020). Mental health matters. *Lancet Glob. Health*, 8, e1352. 10.1016/S2214-109X(20)30432-0.

Thieme A., Belgrave D. and Doherty G. (2020). Machine learning in mental health: a systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Trans. Comput. Hum. Interact*., 5, 1-53.

Toleubay Y., Agravante D.J., Kimura D., Lin B., Bouneffouf D. and Tatsubori M. (2023). Utterance classification with logical neural network: explainable AI for mental disorder diagnosis. *arXiv:2306.03902*

Torkzadehmahani R., Nasirigerdeh R., Blumenthal D.B., Kacprowski T., List M., Matschinske J., et al. (2022). Privacy-preserving artificial intelligence techniques in biomedicine. *Methods Inf. Med.*, 61, E12-E27. 10.1055/s-0041-1740630.

Turcan E. and McKeown K. (2019). Dreaddit: a reddit dataset for stress analysis in social media. In: *Proceedings of the LOUHI@EMNLP 2019 – 10th International Workshop on Health Text Mining and Information Analysis*, Hong Kong, November 2019; pp. 97-107. 10.18653/v1/d19-6213.

Valstar M., Schuller B., Smith K., Eyben F., Jiang B., Bilakhia S., et al. (2013). AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, Barcelona, Spain, 21 October 2013; pp. 3-10.

Valstar M., Schuller B., Smith K., Almaev T., Eyben F., Krajewski J., et al. (2014). Avec 2014: 3D dimensional affect and depression recognition challenge. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, Florida, Orlando, USA, 07 November 2014; pp. 3-10.

Valstar M., Gratch J., Schuller B., Ringeval F., Lalanne D., Torres M. et al. (2016). Avec 2016: depression, mood, and emotion recognition workshop and challenge. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, New York, NY, USA, 16 October 2016; pp. 3-10.

Vaz M., Summavielle T., Sebastião R. and Ribeiro R.P. (2023). Multimodal classification of anxiety based on physiological signals. *Appl. Sci.*, 13, 6368. 10.3390/app13116368.

Venugopal V.K., Gupta A., Takhar R., Yee C.L.J., Jones C. and Szarf G. (2023). Navigating fairness in radiology AI: concepts, consequences, and crucial considerations. *arXiv preprint, arXiv:2306.01333.*

Villa-Pérez M.E. and Trejo L.A. (2023). Twitter dataset for mental disorders detection. *IEEE Dataport*. 10.21227/6pxp-4t91.

Vilone G. and Longo L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion*, 76, 89-106. 10.1016/j.inffus.2021.05.009.

Wang L., Alpert K.I., Calhoun V.D., Cobia D.J., Keator D.B., King M.D., et al. (2016). Schizconnect: mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration. *NeuroImage*, 124 (Pt. B), 1155-1167.

Wang X., Li J., Kuang X., Tan Y. and Li J. (2019). The security of machine learning in an adversarial setting: a survey. *J. Parallel Distrib. Comput.*, 130, 12-23.

Wang Z., Chen L., Wang L. and Diao G. (2020). Recognition of audio depression based on convolutional neural network and generative antagonism network model. *IEEE Access*, 8, 101181-101191. 10.1109/ACCESS.2020.2998532.

Wang C., Feng L. and Qi Y. (2021). Explainable deep learning predictions for illness risk of mental disorders in Nanjing, China. *Environ. Res.*, 202, 111740. 10.1016/j.envres.2021.111740.

Wei Q., Xu X., Xu X. and Cheng Q. (2023). Early identification of autism spectrum disorder by multi-instrument fusion: a clinically applicable machine learning approach. *Psychiatry Res.*, 320, 115050. 10.1016/J.PSYCHRES.2023.115050.

Wickramasinghe C.S., Marino D.L., Grandio J. and Manic M. (2020). Trustworthy AI development guidelines for human system interaction. In: *Proceedings of the 2020 13th International Conference on Human System Interaction (HSI)*, Tokyo, Japan, 06-08 June 2020.

Wong S., Simmons A., Rivera-Villicana J., Barnett S., Sivathamboo S., Perucca P., et al. (2023). EEG datasets for seizure detection and prediction—a review. *Epilepsia. Open.*, 8, 252-267.

World Health Organization (WHO). (2021). *Comprehensive Mental Health Action Plan 2013–2030.*

World Health Organization (WHO). (2022). *Mental Health and COVID-19: Early evidence of the Pandemic's Impact: Scientific Brief, 2 March 2022*, World Health Organization.

Wu Y., Yao Y.G. and Luo X.J. (2017). SZDB: a database for schizophrenia genetic research. *Schizophr Bull.*, 43, 459-471.

Wu C., Li Y. and Bouvry P. (2023). Survey of trustworthy AI : a meta decision of AI. *arXiv preprint, arXiv:2306.00380.*

Xiong P., Buffett S., Iqbal S., Lamontagne P., Mamun M. and Molyneaux H. (2022). Towards a robust and trustworthy machine learning system development: an engineering perspective. *J. Inf. Secur. Appl.*, 65, 103121.

Xu H. and Mannor S. (2012). Robustness and generalization. *Mach. Learn.*, 86, 391-423.

Xue M., Yuan C., Wu H., Zhang Y. and Liu W. (2020). Machine learning security: threats, countermeasures, and evaluations. *IEEE Access*, 8, 74720-74742.

Yang H., Li X., Wu Y, Li S., Lu S., Duncan J.S., et al. (2019). Interpretable multimodality embedding of cerebral cortex using attention graph network for identifying bipolar disorder. In: *Medical Image Computing and Computer Assisted Intervention*, Springer, Cham. 10.1007/978-3-030-32248-9_89.

Yang J., Soltan A.A.S., Yang Y. and Clifton D.A. (2022). Algorithmic fairness and bias mitigation for clinical machine learning: insights from rapid COVID-19 diagnosis by adversarial learning. *medRxiv.*, 10.1101/2022.01.13.22268948.

Yates A., Cohan A. and Goharian N. (2017). Depression and self-harm risk assessment in online forums. In: *Proceedings of the EMNLP 2017 – Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 2017; pp. 2968-2978. 10.18653/v1/d17-1322.

Yazdavar A.H., Mahdavinejad M.S., Bajaj G., Romine W., Sheth A., Monadjemi A.H., et al. (2020). Multimodal mental health analysis in social media. *PLoS One.*, 15, e0226248. 10.1371/journal.pone.0226248.

Yoon J., Kang C., Kim S. and Han J. (2022). D-vlog: multimodal vlog dataset for depression detection. *Proc. 36th AAAI Conf. Artif. Intell.*, 36, 12226-12234. 10.1609/aaai.v36i11.21483.

Zanwar S., Wiechmann D., Li X., Qiao Y. and Kerz E. (2023). What to fuse and how to fuse: exploring emotion and personality fusion strategies for explainable mental disorder detection. In: *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, pp. 8926-8940.

Zhang J. and Zhang Z.M. (2023). Ethics and governance of trustworthy medical artificial intelligence. *BMC Med. Inform. Decis. Mak.*, 23, 1-15. 10.1186/s12911-023-02103-9.

Zhang Z., Lin W., Liu M., and Mahmoud M. (2020). Multimodal deep learning framework for mental disorder recognition. In: *Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020*, 16 November 2020; pp. 344-350. 10.1109/FG47880.2020.00033.

Zhang X., Chan F.T.S. and Mahadevan S. (2022a). Explainable machine learning in image classification models: an uncertainty quantification perspective. *Knowl. Based Syst.*, 243, 108418.

Zhang J.M., Harman M., Ma L. and Liu Y. (2022b). Machine learning testing: survey, landscapes and horizons. *IEEE Trans. Softw. Eng.*, 48, 1-36. 10.1109/tse.2019.2962027.

Zhou Y., Kantarcioglu M. and Clifton C. (2021). Improving fairness of AI systems with lossless de-biasing. *arXiv preprint, arXiv:2105.04534.*

Zhou X., Liu C., Zhai L., Jia Z., Guan C. and Liu Y. (2023). Interpretable and robust AI in EEG systems: a survey. *arXiv preprint, arXiv:2304.10755.*

Zou K., Chen Z., Yuan X., Shen X., Wang M. and Fu H. (2023). A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, 1, 1-12.