Artificial Intelligence Review (2023) 56:11149–11296
https://doi.org/10.1007/s10462-023-10415-5

Trustworthy artificial intelligence in Alzheimer's disease: state of the art, opportunities, and challenges

Shaker El-Sappagh[1,2,3] · Jose M. Alonso-Moral[4] · Tamer Abuhmed[3] · Farman Ali[5] · Alberto Bugarín-Diz[4]

Accepted: 31 January 2023 / Published online: 12 March 2023

Abstract
Medical applications of Artificial Intelligence (AI) have consistently shown remarkable performance in providing medical professionals and patients with support for complex tasks. Nevertheless, the use of these applications in sensitive clinical domains where high-stakes decisions are involved could be much more extensive if patients, medical professionals, and regulators were provided with mechanisms for trusting the results provided by AI systems. A key issue for achieving this is endowing AI systems with key dimensions of Trustworthy AI (TAI), such as fairness, transparency, robustness, or accountability, which are not usually considered within this context in a generalized and systematic manner. This paper reviews the recent advances in the TAI domain, including TAI standards and guidelines. We propose several requirements to be addressed in the design, development, and deployment of TAI systems and present a novel machine learning pipeline that contains TAI requirements as embedded components. Moreover, as an example of how current AI systems in medicine consider the TAI perspective, the study extensively reviews the recent literature (2017–2021) on AI systems in a prevalent and high social-impact disease: diagnosis and progression detection of Alzheimer's Disease (AD). The most relevant AI systems in the AD domain are compared and discussed (such as machine learning, deep learning, ensembles, time series, and multimodal multitask) from the perspective of how they address TAI in their design. Several open challenges are highlighted, which could be claimed as one of the main reasons to justify the rare application of AI systems in real clinical environments. The study provides a roadmap to measure the TAI status of an AI systems and highlights its limitations. In addition, it provides the main guidelines to overcome these limitations and build medically trusted AI-based applications in the medical domain.

Keywords Trustworthy AI · AI for Alzheimer's disease diagnosis and progression detection · Machine learning in medicine · Responsible AI · Fairness, accountability, and transparency in AI

✉ Tamer Abuhmed
tamer@skku.edu

Extended author information available on the last page of the article

1 Introduction

1.1 Artificial intelligence in medicine

Artificial Intelligence (AI) rapidly impacts everyday life (Ji et al. 2021). The healthcare industry is not an exception, where AI applications are becoming pervasive due to the complexity of health problems such as chronic disease management, treatment protocol creation, medication research, customized medicine, patient monitoring, and care (Vesnic-Alujevic et al. 2020; Buruk et al. 2020). Since it is challenging for a physician to comprehend such complicated health issues, AI can be utilized as a support tool that provides medical professionals with valuable insights on available data and information which help them to assess better illness diagnosis, detection, progression, or medication recommendations, among others (Liu et al. 2020). In this regard, Clinical Decision Support Systems (CDSSs) can assist medical professionals in reducing errors in decisions, improving the effectiveness and quality of care, and optimizing the delivery of personalized medicine at the right time, which is of crucial importance (Sappagh et al. 2021; Abuhmed et al. 2021). Qayyum et al. (2021) reviewed prognosis, diagnosis, treatment, and clinical workflow applications of Machine Learning (ML) in the medical domain, and El-Sappagh et al. implemented an ontology-based CDSS for diabetes diagnosis (2018). However, dealing with AI in the medical domain is not straightforward, it could carry potential pitfalls and inherent biases which raise the possibility for harmful errors with high social, legal, and economic consequences (Thiebes et al. 2021; Cutillo et al. 2020). Even if CDSSs were introduced in 1959 (Ledley and Lusted 1959), almost no actual practical application has been reported (Ji et al. 2021; Wang, et al. 2101). CDSSs are not ready for clinical use in real-time, indicating that existing approaches are inadequate (Kovalchuk et al. 2020). There are many reasons and barriers for this failure, including the lack of rigorous development, accurate evaluation of the system's effectiveness and usability, and the lack of reliability and accountability evaluations of predictive AI systems (Jacobs et al. 2021). Besides, ML and Deep Learning (DL) pipelines are stochastic and probabilistic in most of their steps. At the same time, uncertainty turns up in every step of CDSSs, such as (1) patients fail to describe their conditions, (2) physicians cannot interpret what they observe, (3) lab results contain some degree of errors, and (4) medical data are not consistent. As a result, CDSSs did not gain the trust of medical institutions to be used in real domains (Thiebes et al. 2021). Indeed, Hatherley (2020) suggested that AI should not be used in the medical field because it would erode the patient-physician trust relationship. Sánchez-Martínez et al. (2019) summarized the challenges influencing the integration and implementation of CDSSs, and the most critical parameter of success for CDSS systems is the patients' and physicians' trust: Building a CDSS based on trustworthy AI is expected to improve model acceptability in real-world environments (Middleton et al. 2016). Toreini et al. (2020) reviewed technologies that support building reliable ML systems. The trustworthiness of the CDSSs is guaranteed in a system that is based on robust, fair, explainable, scalable, auditable, secure, reliable, interoperable, safe, and responsible AI models (Wang et al. 2101; Markus et al. 2020; Dignum 2019). These properties must be considered across all stages of the model life cycle. ML deployment in healthcare should involve interdisciplinary teams of physicians, knowledge experts, and decision-makers (Wiens et al. 2019; Gillespie et al. 2020). Moreover, CDSS's patient-centric decisions should be based on the complete patient's data, including demographics, images, lab results, genetics, and others collected from distributed Electronic Health Records (EHRs) (He et al. 2019). For example, Jagannatha and Yu (Richards et al.

2018) trained a bidirectional Recurrent Neural Network (RNN) for cancer disease prediction and medication based on EHR data, and Choi et al. (2016) introduced Retain, a high-accuracy model for predicting heart failure using EHR data. The seamless integration in clinical workflow and the use of explainable and interpretable ML models greatly influence the user acceptance for CDSSs. However, the methods used to develop AI technology in healthcare are currently siloed. Models are built-in labs on a extremely narrow scale, isolated from real-world settings, and without the context of larger EHR ecosystems. This could result in systems that potentially do more harm than having a positive impact (Ji et al. 2021). Recently, Kovalchuk et al. (2020) proposed a three-stage trustworthy CDSS pipeline which engages medical experts in system design, merges knowledge-based and data-driven CDSSs, and supports continuous monitoring of performance and explainability.

1.2 Alzheimer's disease medical problem

Recently, the United Nations General Assembly declared 2021–2030 the Decade of Healthy Aging, which implies enabling the elders to remain active citizens from both physical and cognitive viewpoints, ultimately "improving the lives of older people and their families and communities." Accordingly, one area of healthcare in which researchers expect a substantial role of AI-based CDSSs is AD diagnosis and progression detection. AD is a progressive neurodegenerative disorder that affects 10% of the population over 65 years old (Tanveer et al. 2020). It is a multifactorial disease resulting from the complex interplay of multiple pathological, environmental, and genetic factors which cause functional and structural changes in a patient's brain (Chetelat 2018). According to the World Health Organization (WHO), 47 million people are living with dementia worldwide, which is expected to reach 82 million in 2030 and 150 million by 2050. This means a person develops AD every 3 s (Ahmed et al. 2019a). The number of AD patients is estimated to double in the next 20 years (Alberdi et al. 2016). The death rate caused by AD has increased by 68% between 2000 and 2010 (Association 2019), and in 2017, AD was ranked among the top 5 causes of death worldwide, with 2.44 million (4.5%) deaths (Khatami et al. 2020). In 2017, the Alzheimer's Association report asserted that the growing costs for managing AD are estimated at about $259 billion (Forouzannezhad, et al. 2018). Although it is considered one of the most common illnesses of later life, the treatments of AD are symptomatic, i.e., administered after advanced and irreversible stages of the disease process. In other words, AD cannot be diagnosed until the disease has been clinically confirmed, and cognitive and functional declines impair the patient's ability to meet life's demands. (Weiner et al. 2017). This has enormous social and economic consequences. Only a few symptomatic treatments are available, which are only effective for a short time and for a small number of patients (Nordberg et al. 2010). AD is chronically and gradually developed in 3 phases (Khatami et al. 2020). The first phase is the preclinical (pre-symptomatic) AD stage. In this stage, changes in brain structure, blood, and Cerebrospinal Fluid (CSF) may start happening, but the patient does not experience any symptoms. Nowadays, detecting this stage is challenging because it can start 20 years before any symptom is evidenced (Rathore et al. 2017). Some studies linked the youth's linguistic ability and late-life progression to AD (Riley et al. 2005). The prodromal phase, i.e., Mild Cognitive Impairment (MCI), is the second phase, where symptoms related to thinking ability may start to be noticeable, but they do not affect a patient's daily life. For unknown reasons, only 10–15% of MCI patients progress to AD (Alberdi et al. 2016). This group is called progressive MCI (pMCI), while non-progressive MCI is called stable MCI (sMCI). Some studies divide this phase into

early MCI and late MCI (Weiner et al. 2017) according to the severity levels of symptoms. Other studies distinguish MCI patients based on the memory impairment into amnestic MCI (aMCI) and non-amnestic MCI (naMCI), where the first group is more likely to develop AD (Varghese et al. 2013). It is worth noting that MCI has multiple pathologies, and AD is just one of them. The third phase is AD, where memory, thinking, and behavioral symptoms are evident and affect the patient's daily life.

On the one hand, AD symptoms are always confused with the normal aging process, and physicians are not consulted until too late, resulting in a late diagnosis. On the other hand, AD clinical diagnosis and progression detection processes are complex tasks and depend on heterogeneous types of pathological, physiological, behavioral, cognitive, and psychological assessments. Difficult as it is to comprehend the etiology of AD, significant efforts have been made to identify markers and biomarkers for AD pathological change, including neuroimaging, CSF samples, and numerous amyloid data (Blennow and Zetterberg 2018). As AD is a multifactorial disease, the combination of different markers and biomarkers is critical to building CDSSs, e.g., cerebrovascular amyloid protein and tau protein, nerve cell degeneration; cognitive evaluation to determine cognitive and functional declines; drugs such as tacrine to measure symptoms, neuroimaging biomarkers (e.g., Magnetic Resonance Imaging [MRI] and Positron Emission Tomography [PET]), genetics measurement to determine Down syndrome, comorbidities such as diabetes and hypertension, neuropsychological measures, symptoms, text data from interviews, and blood lab tests. As a result, AD diagnosis and progression detection are based on multimodal data. These data can collectively quantify brain morphology, function, connectivity, and pathology changes, including amyloid plaques, hypometabolism, and neurological disorders (atrophy). Most of these methods used for AD management (e.g., early diagnosis and progression detection) involve many factors; hence, they are complex, costly, time-consuming, and require expert interventions. For example, MRI is increasingly used because it allows the physician to analyze a patient's brain by eye. Still, when changes become visible to an eye, it usually is too late because the brain has evident signs of atrophy. Given the multifactorial nature of the disease, it is not a good idea to rely on a single category of markers and biomarkers (e.g., MRI) to make a clinical diagnosis. Only a panel of markers and biomarkers fusion may offer the appropriate sensitivity, specificity, and positive and negative predictive values.

Reliable, early, automatic, continuous, and trustworthy AD detection and prediction are highly important, especially in preclinical/predementia stages (Ahmed et al. 2019a; Varghese et al. 2013). This is the most effective method for controlling AD's progression, which helps preventive and disease-modifying therapies be more effective in delaying symptoms. In addition, the accurate progression detection of the disease assists physicians in predicting the future status of patients, which improves the AD survival rate (Khedher et al. 2015). Disease management improves the patient's quality of life and avoids high healthcare costs (Alberdi et al. 2016). Because AD is a chronic disease, multimodalities are collected over time which results in time series data (Sappagh et al. 2021; El-Sappagh 2021; El-Sappagh et al. 2020). Time series data can be handled using RNN (Cui and Liu 2019), Convolutional Neural Network (CNN) (Khvostikov et al. 1809), stacked RNN-CNN (Abuhmed et al. 2021; Dua et al. 2020), or a hybrid model (Abuhmed et al. 2021). CDSSs have the potential to automatically, quickly, and accurately assist clinical experts in classifying and predicting AD using sophisticated, objective, and automatic classification and regression techniques that can analyze and learn complex patterns from high-dimensional data (Tanveer et al. 2020). Previous computer-aided methods targeted a single aspect of the disease, but these methods achieved limited results because of the complex nature

of AD. Integrative disease modeling based on heterogeneous modalities (i.e., built on a wide range of markers and biomarkers) supports comprehensive analysis (Khatami et al. 2020). ML models can extract and highlight tiny and not noticeable changes in many features that facilitate the accurate and early detection of the disease. Many classification techniques have been used, including Support Vector Machine (SVM), Artificial Neural Network (ANN), and DL (Ahmed et al. 2019a; Falahati et al. 2014). Regular ML models include separate steps for feature engineering, but DL incorporates the representation learning phase in the learning process itself (Ebrahimighahnavieh et al. 2020). As a result, DL is the best choice for big data analysis such as neuroimages (Zhang et al. 2017). Some studies used ensemble techniques to improve the classification accuracy of AD (Ramírez et al. 2018; Yao et al. 2018; Lebedev et al. 2014). In binary classification, the accuracy is higher regarding Cognitively Normal (CN) vs. AD. It is harder to classify CN vs. MCI and even harder for MCI vs. AD (Beheshti et al. 2017). Additionally, the classification of pMCI vs. sMCI and aMCI vs. naMCI are also challenging tasks (Rémi Cuingnet et al. 2011). The classification problem has also been formulated as a 3-class classification (e.g., CN vs. MCI vs. AD) and a 4-class classification task (e.g., CN vs. sMCI vs. pMCI vs. AD). Accurate prediction or diagnosis of AD requires monitor of multiple markers and biomarkers at the same time. This is called multitask modeling (Tabarestani et al. 2019) and has been used in several medical domains (Ali et al. 2022; El-Rashidy et al. 2022; Juraev et al. 2022). Multitask modeling has been explored in AD diagnosis and progression detection using single modality or multiple modalities, e.g., single-modal single-task classification (Lu et al. 2017a) and regression (Ito et al. 2011) learning, single-modal multitask regression learning (Zhou et al. 2013), multimodal single-task classification (Liu et al. 2014) and regression (Duchesne et al. 2009) learning, and multimodal multitask learning (Ding et al. 2018). AD progression has also been modeled using a multimodal multitask process based on time series data (Abuhmed et al. 2021; El-Sappagh et al. 2020).

1.3 Trustworthy AI for AD

Despite the seminal studies on the technical aspects of applying ML and DL techniques to build efficient and accurate CDSSs for AD, these models are rarely integrated into clinical practice. Current research focuses mainly on creating and optimizing highly accurate medical-assisted CDSSs based on ML algorithms and less representative datasets (Fang et al. 2020). However, improving model accuracy alone is not always correlated with overall performance when deployed in the real world. Due to the lack of large-scale deployment, a considerable gap exists between the research and clinical practice of AI-based CDSSs. For deploying AI systems in clinical environments, there are many issues beyond accuracy, including social, technical, ethical, and organizational challenges that must be carefully considered to design trustworthy CDSSs, e.g., low context awareness, patient privacy, and poor workflow integration (Jacobs et al. 2021). In addition, medical professionals may resist accepting CDSSs because they are afraid of being replaced by artificial agents (Pesapane et al. 2020). In addition, in other domains, such as criminal recidivism and job applications, the employed data-driven systems have shown signs of social biases and racial and ethnic discrimination (Varona et al. 2021; Wiśniewski and Biecek 2021a). The medical domain is highly sensitive because it involves human lives. As a result, AI's ethical and social implications rise to the forefront of public, political, and research agendas. Even if domain experts promised to improve diagnostic efficiency and accuracy, it is still challenging for CDSSs to be deployed in real hospitals. The main reason for this problem

is that CDSSs are not trustworthy for the following challenges (Neri et al. 2020; Wang, et al. 2101; Gillespie et al. 2020; Pesapane et al. 2020; Expertengruppe and für Künstliche Intelligenz (HEG-KI), High-Level Expert Group on Artificial Intelligence 2020; ; ; ; ; ; ; ;): (1) Extensive computing resources are required for the model training; (2) vast and high-quality datasets are required to be collected from EHRs for the training process, but medical institutions operate independently and must maintain data privacy; (3) heterogeneous EHR data sources are not standardized and they are poorly integrated with CDSSs; (4) used data are biased which creates a fairness challenge; (5) data and model provenance should be assured; (6) models are not sufficiently validated and tested; (7) models are not sufficiently accurate, stable, and robust; (7) most models act as complex black boxes and their decisions are neither transparent nor interpretable nor explainable; (8) patients and physicians do not collaborate in model's building, training, validation, and testing (human-in-the-loop challenge); (9) models are not seamlessly integrated in the clinical workflow; (10) lack of personalized decisions based on whole patient's medical history; (11) security issues are not handled; and (12) responsibility and accountability of AI decisions need new regulations.

Trustworthy AI (TAI) is a collection of sociotechnical requirements which must be handled altogether throughout the life cycle of predictive model construction: data collection, data preprocessing, feature engineering, model training and optimization, model testing, model deployment, and model use and monitoring (Wang, et al. 2101; Gillespie et al. 2020; Suryanarayanan, et al. 2020). This is known as the "chain of trust" (Toreini et al. 2020). To be acceptable in a real environment, trustworthy AI-based CDSSs must be endowed with accountability, responsibility, fairness, robustness, transparency, reproducibility, safety, and privacy. Some studies address only part of the problem. For example, considering only explainability (Rashed-Al-Mahfuz et al. 2021), security and privacy (Arachchige et al. 2020), explainability and security (Huang et al. 2020), robustness (Hossain et al. 2020), or just reliability (Durán and Jongsma 2021) in the search for getting the user's trust. To effectively communicate how AI-based systems are perceived and utilized, as well as what constraints and challenges exist for their implementation, considerable work remains despite current research in several domains to tackle these issues (Arnold et al. 2018). To get an in-depth and empirical grasp of these aspects, substantial research is required to comprehend how AI-based CDSSs should be used in clinical practice and their operational characteristics. Our study helps to analyze the current state of AI-based CDSSs in the AD domain.

1.4 Study contributions

It is believed that TAI is the bedrock for acceptable, human-centered, and responsible AI systems. Based on the ALTAI guidelines, TAI has seven requirements (human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; accountability) (Expertengruppe and für Künstliche Intelligenz (HEG-KI), High-Level Expert Group on Artificial Intelligence 2020). When people recognize that an AI-based CDSS adheres to these requirements, it is likely to be trusted by users (Gillespie et al. 2020). Trustworthiness will be checked in the model's design, development, procurement, and deployment phases by concentrating on the data, algorithms, and processes. To the best of our knowledge, there are no studies in the literature that comprehensively reviewed and discussed the TAI concepts, requirements, techniques, guidelines, and applications in the AD domain. In addition, no studies proposed a framework to map the high-level TAI

guidelines into practical steps in the ML pipeline. Most previous surveys either concentrate on Explainable AI (XAI) or discuss TAI in general without focusing on a specific domain (Buruk et al. 2020; Markus et al. 2020; Huang et al. 2020). This study will review the current literature on ML and DL models in the AD domain by focusing not only on the performance of the proposed models, but also regarding the utilized datasets and their properties, the utilized validation methods, data preprocessing steps, and so on. Moreover, the study evaluates the level of trustworthiness achieved by each reviewed study based on the ALTAI guidelines. To do so, the study discusses the AD research considering TAI principles. Namely, this work concentrates on robustness, fairness, and transparency. The main contributions are as follows:

1. The study provides an overview of the latest TAI guidelines, standards, and requirements.
2. The TAI requirements are mapped to a quantitative questionnaire that AI practitioners and users could use to evaluate the trustworthiness of an AI system.
3. A comprehensive ML pipeline is proposed, which includes embedded components at various stages to manage, check, or assess different TAI requirements.
4. Because AD is a critical medical problem and because the AD actual domain has not benefited well from AI, the AD domain is considered as our case study to apply our formulations of TAI requirements. A comprehensive survey of the latest AD research studies from many search databases is provided. First, the study concentrates on the studies that do not explicitly consider TAI in their design. Second, it concentrates on the studies that explicitly consider TAI requirements in their implementation. The focus is on the most critical requirements of robustness, fairness, and transparency.
5. A set of limitations that are believed to be the main reasons for the current limitation of AI literature in the AD domain are explored and highlighted.

This review is organized as follows. Section 2 discusses the methods used to prepare the study, and Sect. 3 discusses existing guidelines for trustworthy AI. Then, Sect. 4 goes deeper and discusses the technical methods for implementing trustworthy requirements in AI-based CDSSs. Section 5 discusses the application of TAI in the AD domain as a case study. Section 6 discuss the effort to apply trustworthy AI in industry and other domains. Finally, Sect. 7 concludes the study and proposes future research directions.

2 Methods

2.1 Search strategy

To identify relevant studies, candidate search terms are first set based on a preliminary search. Based on the resulting terms, the following search text words and MeSH terms containing generalized keywords were used to avoid potential bias in searching for the state-of-the-art studies: [Alzheimer's, AD, dementia, mild cognitive impairment, MCI], [time series, multimodal, multitask, optimization, longitudinal], [conversion, detection, prediction, diagnosis, progression], [Trustworthy, trustworthiness, user-centric, human-centric, ethical, robustness, privacy, fairness, responsible, interpretable, explainable, trust, reproducibility, reliable, XAI, or accountability], [neuroimaging, neuropsychological, cognitive

score, symptom, genetic, MRI, PET, EEG, fMRI], [transfer learning, CDSS, deep learning, machine learning, ensemble, hybrid, decision support], [classification, prediction, detection, regression]. The controlled terms were combined using "AND" and "OR" for a better searching strategy. Synonyms of the candidate terms are included using the "OR" operator to maximize efficiency. First, a rigorous literature search is conducted through five leading electronic databases, including PubMed, ScienceDirect, SpringerLink, Nature, and IEEE Xplore. Also, Web of Science and Scopus are queried to cross-check the findings. These specific digital libraries were chosen since they offer the most important and recent peer-reviewed full-text journals covering the field of ML in the healthcare field. In addition, to collect the most recent papers, all manuscripts uploaded to arXiv and medRxiv are considered in the defined timeframe. Because of duplication, any published articles are removed from the collected list from arXiv and medRxiv. The study concentrates only on highly ranked journal papers, and chapter and conference papers are excluded. The terms are searched in the title, abstract, and keywords of the papers. Second, each article's titles, keywords, and abstract were independently and manually scanned and interpreted. Third, articles meeting our inclusion criteria are considered for the full-text review. For each candidate article, the full text is carefully read with the aim of keeping in the results only the most relevant articles. Moreover, reference lists of the selected articles are manually scanned to extract additional articles to get a complete overview of the field. The final list of articles is included in the review process. These articles are reviewed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moheret al. 2009). PRISMA is the most popular methodology for conducting accurate and comprehensive literature studies (Page et al. 2021).

2.2 Eligibility criteria

As a use case, the study focuses on reviewing state-of-the-art AD studies. The study concentrates on reviewing recent ML and DL studies in the AD domain and investigating the implemented trustworthy guidelines in these studies. The inclusion and exclusion criteria were set up through rigorous discussions and brainstorming among the authors. Inclusion criteria are: (1) The focus of the collected studies develop, test, and discuss regular ML, DL, trustworthiness, neuroimaging, multimodality, time series, multitasking, and any other hybrid algorithms in the AD domain. (2) Only English language studies are considered. (3) The studies are published from January 2017 to December 25, 2021. (4) The focus is on peer-reviewed original journal papers only. (5) Only AD diagnosis and progression detection studies are included. Articles which do not meet these criteria are excluded from further processing. For example, preliminary studies, editorials, opinion papers, book chapters, conference papers, or reviews are out of the scope of this study. In addition, other brain diseases such as Parkinson's, Down syndrome, Schizophrenia, and Autism are not considered in this study. Furthermore, articles that could not be accessed in full text are excluded. In the end, the full texts of 110 papers were included in the full-text review.

2.3 Study selection

The initial query search in the electronic databases resulted in $n = 3088$ records that satisfied the defined search criteria. Searching in reference lists results in $n = 30$ additional publications. We have $n = 3038$ candidate papers in terms of their title, keywords, and abstract after removing 80 papers that are duplications and published in a language other than

English. After scanning titles, keywords, and abstracts, $n=2238$ papers were excluded, resulting in $n=800$ eligible papers for full-text screening. The full-text screening retained $n=310$ papers, of which, after quality review, $n=110$ papers were finally included for discussion and analysis in this study (see Fig. 1). As will be seen in detail in Sect. 5, 86 out of 110 papers (78.18%) do not explicitly take care of trustworthy AI (see Sect. 5.1). More precisely, 10 out of 86 papers are survey papers (see Table 2), 18 papers focus on classical ML models (see Table 3), 17 papers are related to DL models (see Table 4), 13 papers are related to ensemble models (see Table 5), 15 papers pay attention to time-series data (see Table 6), and 23 papers deal with multimodal multitask problems (see Table 7). Accordingly, only 21.82% of papers under consideration deal with trustworthy issues (see Table 8 in Sect. 5.2).

3 Trustworthy AI guidelines

There is no single definition of trustworthiness in AI. The study defines TAI as the property for which an AI system adheres to the ethical principles of fairness, explicability, respect for human autonomy, and harm prevention. Numerous high-level guidelines and standards (i.e., more than 60) have been proposed to evaluate and monitor trustworthiness and ethics in AI, such as AI4People and Asilomar AI principles (Thiebes et al. 2021; Toreini, et al. 2020; Gillespie et al. 2020). Nevertheless, TAI is a dynamic and multidisciplinary field of research. As noted in the European Union (EU) guidelines, some TAI principles may contradict each other and be redundant. The main reason for TAI is to develop AI systems in agreement with human values and needs. In 2018, the European Commission created a High-Level Expert Group (HLEG) on AI. This group defined TAI in terms of three desirable properties of AI systems (Hleg 2019): lawful, ethical, and robust. *First*, the HLEG defined four ethical principles to achieve trustworthiness: (1) respect for human autonomy; (2) prevention of harm; 3 fairness; and (4) explicability. *Second*, the HLEG outlined seven critical requirements for the four principles: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination, and fairness; (6) societal and environmental wellbeing; and (7) accountability. It is worth noting that some of these requirements, like "Human agency and oversight" and "Societal and environmental wellbeing" are the least relevant in the scope of our study. Other requirements such as robustness, fairness, security, and privacy have accurate measures to evaluate. *Third*, regarding assessment and validation of AI systems, the HLEG translated the seven high-level requirements into a trustworthy assessment checklist (ALTAI) (Expertengruppe and für Künstliche Intelligenz (HEG-KI), High-Level Expert Group on Artificial Intelligence 2020). ALTAI is the most applicable guideline for measuring trustworthiness of AI-based systems (Jacovi et al. 2021). This paper follows ALTAI comprehensive guidelines for evaluating the status of TAI in AD literature and finding out the main reasons for the current minor impact of CDSSs in real clinical environments. This is the first study that discusses TAI in AD domain. To the best of our knowledge, there is no deployed system in a real hospital for diagnosing and predicting the progression of AD.

Each HLEG requirement is considered a contract that collectively creates the "broad trust." However, trust and trustworthiness are different terms. Trust has a lack of conceptual clarity on its meaning and dynamics. If party A believes that party B will act in A's best interest, and accept the vulnerabilities of B's actions, then A trusts B. Jacovi et al.

(2021) formalized the "contractual trust," where the trust between stakeholders and CDSSs ensures that some implicit or explicit contract will hold. Such a contract implies an obligation by AI developers to achieve the agreement. It could be said that the AI model is trustworthy to a contract if it can maintain this contract.

On the other hand, trustworthiness is a property of a syst em with some characteristics that make it deserve users' trust. However, users can trust the untrustworthy system, and a trustworthy system does not necessarily get trust (Jacovi et al. 2021). Formally, if user H believes that AI model M is trustworthy to contract C, and accept M's vulnerabilities, then H trusts M contractually to C. As a result, AI users always perceive model risks. This way, the level of trustworthiness for CDSSs could be quantified.

3.1 Human agency and oversight

This requirement includes three components that describe the principle of respect for human autonomy:

1) *Fundamental rights:* AI systems can positively or negatively affect human rights. A fundamental rights impact assessment should be used before system development to evaluate the level of the risk. We believe that this high-level requirement cannot be applied to AI-based CDSSs.
2) *Human agency:* AI-based CDSSs are aimed at assisting physicians in making accurate, personalized, and on-time decisions. AI-based CDSSs will never replace physicians (Durán and Jongsma 2021). In addition, it is early to decide whether patients can use AI-based CDSSs away from physicians. As a result, both patients and physicians should be able to make autonomous decisions with the support of AI systems. They must interact with the system, challenge it if needed, and potentially edit and override the suggested decisions. Overreliance on AI systems can influence and affect human behavior and knowledge, posing a threat to personal autonomy. As a result, physicians and patients should have the right not to be subject to an automated decision when this has legal effects on them.
3) *Human oversight:* this requirement ensures that AI-based CDSSs do not cause any adverse effects or restrict human autonomy. Users need governance mechanisms, including human-in-the-loop (HITL), Human-On-The-Loop (HOTL), and Human-In-Command (HIC) methods (On 2019; Budd et al. 2019). HITL is the ability of the user to interfere in the decision cycle of the system, HOTL gives users the capability to take part in the system design cycle, and with HIC, users can monitor the impact of CDSSs and decide when and how to use them.

3.2 Technical robustness and safety

Robustness is the most critical requirement for the evaluation of AI-based CDSSs. It measures the sensitivity of the system's outcome to a change in the input. Robustness also measures the model's ability to work accurately under uncertain conditions. This property is evaluated by exposing the model to adversarial inputs, and the system should achieve an error rate close to training error. Mechanisms are required to prevent possible risks, ensure reliable behavior as planned, and minimize unexpected damages. Robustness needs to be guaranteed in case of potential change in the operating

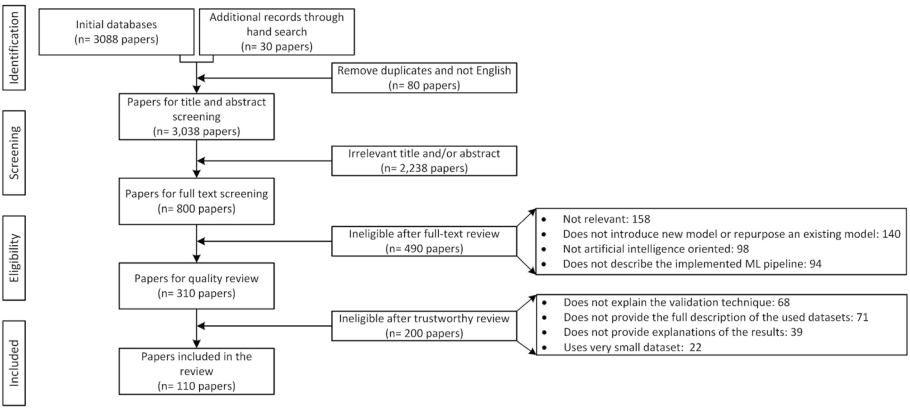

Fig. 1 PRISMA (Huang et al. 2020) flowchart for the study

environment or the existence of agents that interact in an adversarial manner with AI-based CDSSs. Nicolae et al. (2018) proposed a tool to create defense techniques for ensuring the robustness of AI systems:

1) *Resilience to attack and security:* an AI system should be secure against vulnerabilities and adversarial attacks. These attacks can affect data (data poisoning), the model (model leakage), or the system infrastructure (i.e., hardware and software) (Toreini et al. 2020). If attacked, the system could take wrong or at least different decisions, and data may become corrupted by malicious intention. Sufficient mechanisms should be in place to prevent adversarial attacks. Issues related to system availability, data storage and encryption, and access control are outside the scope of this study. The study concentrates on the security issues related to the ML domain, like malicious attacks based on poisoning training or online data, spoofing attacks, and inversion attacks (Toreini et al. 2020).
2) *Fallback plan and general safety:* This requirement determines the safeguards which enable a fallback plan when an attack or error occurs. Perfect performance in a controlled lab environment is not proof of safety (Qayyum et al. 2021). For example, the system can stop working or switch to another CDSS like a rule-based system. The AI-based CDSS solution should ensure that the system performs as planned and minimize the risks associated with errors. In each phase of the life cycle of AI-based CDSSs, a set of precautionary measures must be applied.
3) *Accuracy:* AI-based CDSSs must achieve high training and testing accuracies. Accuracy is the model's ability to predict outcomes by reducing false positives and false negatives. Training pipeline, robust validation mechanism, sufficient and high-quality data, ML model selection and optimization, optimized feature engineering, and handling of biased and imbalanced data, among others, are defined to support, mitigate, and correct unintended risks from inaccurate decisions. As errors cannot be prevented entirely, at least they should be quantified and reported.
4) *Reliability and reproducibility:* With identical input parameters and operating conditions, CDSSs should produce identical outcomes. A minor input modification should not significantly change the system's behavior. The system is reliable if it functions correctly in a variety of inputs and settings. Reproducibility is attained when identical results are obtained under identical settings.

3.3 Privacy and data governance

Privacy is related to the principle of prevention of harm, where adequate data governance for quality and integrity, data relevance, access protocols, and data processing should be adopted. This requirement has three components:

1) *Privacy and data protection:* Human-centric CDSS requires access to distributed EHR data to make personalized decisions, which increases the challenge of protecting patient's privacy. Other data that need privacy protection include the ML model parameters and hyperparameters. If the AI-based CDSS integrates expert-defined knowledge bases and rules for data preprocessing, data selection and data filtering should be protected. CDSSs should protect patients' training data and identity, internal system logic, and any intellectual property. Patient data to be protected include data used for building the model, data collected by the physician when interacting with the system, and patient decisions. If the AI system accesses the whole patient's medical record, this allows the AI system to infer not only the patient's conditions but also other private data, such as family conditions. Suitable measures should be implemented to ensure that the collected data will not be used unlawfully. There is always a tradeoff between model transparency and privacy (Schneeberger et al. 2020). He et al. (2020) discussed different types of attacks on the AI system and measures that should be taken to prevent them.
2) *Quality and integrity of data:* The quality of CDSSs depends on the quality of the training data. The medical domain is well known for its low-quality data. Data are always sparse and include missing values. Data regarding symptoms, medications, and comorbidities are always collected in the raw text without standardizations. All these issues must be addressed before using the data in model training or testing. Data integrity ensures that no malicious data are fed to the system.
3) *Access to data:* Data access controls determine who can access specific data and for what purpose. The data to be handled by AI systems include input data (e.g., images, numerical and categorical structured data, and text data), model data (e.g., features, model parameters and hyperparameters, algorithms), and output data (e.g., predictions).

3.4 Transparency

This is a crucial requirement for TAI, which is closely linked to the principle of explicability. It comprises three components:

1) *Traceability:* All data, processes, and algorithms should be documented. System decisions should also be documented to track decisions and determine mistakes. As a result, traceability supports auditability and explainability; and contributes to creating responsible AI.
2) *Explainability:* Users have the right to get an explanation for the AI decisions and technical processes as stated by the "Right to an Explanation" in the GDPR (Schneeberger et al. 2020). Even if XAI enables models to be more predictable and transparent (Alejandro Barredo Arrieta et al. 2020), there is no obligation or law for that right. In addition, the GDPR does not clarify the level of correctness or details of the explanation. Many open-source packages implement explainability features (Arya et al. 2020).

Tradeoffs always exist between explainability and security, privacy, and performance. The suitable explanation depends on factors in the system's context, such as the available time, the experience of the user, the available explanation mechanisms, or the security and privacy issues regarding the data and model parameters. Explanations should be dynamic based on the context and interactive using human–computer interaction techniques. The AI system should provide different XAI mechanisms (e.g., visualization, feature importance, rules, natural language explanations, etc.).

3) *Communication:* Patients should know when they are dealing with an AI system and be able to select between AI and medical expert decisions. The limitations of the AI system should be communicated to its users.

3.5 Diversity, non-discrimination, and fairness

All affected stakeholders should have equal access to the AI system. This means the absence of biases. Diversity must be enabled in the entire AI system's life cycle. This requirement implements the principle of fairness, and it has three components as follows.

1) *Avoidance of unfair bias:* Datasets and algorithmic biases result in unfair discrimination against certain people. The AI system can suffer from different kinds of biases (Varona et al. 2021). Identifiable data biases should be removed from the data collection phase. Furthermore, the system development process may suffer from bias. AI-based CDSSs should not be affected by patient characteristics that are not related to the medical problem, such as place of birth, ethnic or social origin, and abilities. Many open-source packages provide the understanding and mitigating of biases (Bellamy et al. 2019).
2) *Accessibility and universal design:* CDSSs should not have a one-size-fits-all approach and universal design methods that cover all possible users.
3) *Stakeholder participation:* CDSSs require the active involvement of all stakeholders who may be directly or indirectly affected in all life cycle phases. Specific mechanisms should be considered to collect and implement user feedback during system design and deployment.

3.6 Societal and environmental wellbeing

The principle of fairness requires the AI system to prevent harm to the environment and wellbeing. This requirement has three components: (1) sustainability and environmental friendliness, (2) social impact, and (3) democracy. However, this requirement has no role in designing an AI-based CDSS.

3.7 Accountability

"Who is responsible for the AI decision?" is a crucial question that affects model trustworthiness. This is called the responsibility gap (Durán and Jongsma 2021), where either physicians or AI-based CDSSs may be responsible for the medical decision. Thus, CDSSs must have the ability to justify automated decisions. This requirement is closely linked to the principle of fairness and pays attention to the mechanisms through the model life

cycle, ensuring the responsibility and accountability of AI decisions. Katell et al. (2020) discussed different types of accountabilities and their relationships to different user types. XAI can help the physician to interpret why (or why not) and how the AI system came to a specific decision. There are two primary components for this requirement:

1) *Auditability:* Internal and external auditors can assess an AI system's algorithms, data, and design processes. The designer should take responsibility for proper testing and auditing of the system. Providing mechanisms for auditing ensures that the model is accountable for the decisions it produces.
2) *Risk management:* This is the ability to identify, assess, report, and minimize the potential negative impact of an AI system.

4 Technical methods for implementing TAI requirements in AI-based CDSSs

This section formalizes the definition of TAI and its requirements according to HLEG guidelines. For each requirement, we collect the existing evaluation measures and possible dimensions to evaluate. In addition, we summarize the literature for medical and non-medical studies which are associated with each requirement. The ALTAI checklist proposed by HLEG is connected to the ML pipeline phases, including design, development, implementation, deployment, and use. TAI components (lawful, ethical, and robust) and requirements must be embedded into AI services and products to create human-centric systems. TAI requires "a holistic and systemic approach, encompassing the trustworthy of all actors and processes that are part of the system's socio-technical context through its entire life cycle" (On 2019). *Lawful AI* means that the AI system operates according to legal rules which establish what cannot be done, what should be done, and what may be done. Nevertheless, the HLEG guidelines do not concentrate on AI laws and regulations. *Ethical AI* means that the AI system is aligned with ethical norms. *Robust AI* means that the system should operate in a safe, secure, and reliable manner. Robustness has technical and social perspectives in each context (On 2019). Note that issues like "respect for human dignity and rights," "freedom of the individual," and "respect for democracy and justice" are out of the scope of this study. The concentration is mainly on the trustworthy requirements to deliver an AI-based CDSS for AD. A full discussion of the high-level ethical principles can be found in (Expertengruppe and für Künstliche Intelligenz (HEG-KI), High-Level Expert Group on Artificial Intelligence 2020; On 2019). To implement TAI in CDSSs, there are functional and non-function methods. On the one hand, Toreini et al. (2020) organized trustworthy functional requirements into two classes: (1) *data-centric trustworthy,* which concentrates on data collection, data preprocessing, and feature engineering and (2) *model-centric trustworthy,* which concentrates on model training and validation, model testing, and model deployment. On the other hand, non-functional methods (e.g., regulation, standardization, certification, education and awareness, accountability by government, or inclusion of stakeholders) are out of the scope of this study.

For each requirement, a list of questions has been suggested that will be used to evaluate AD literature regarding trustworthy AI. The following sections will discuss the details of the functional methods to achieve TAI. The study considers the proposed standardized documents to define each HLEG requirement's contract formally. These documents include

data statements (Bender and Friedman 2018), datasheets for datasets (Gebru et al. 2021), model cards (Mitchell et al. 2019), reproducibility checklists (School and of Computer Science 2020), fairness checklists (Madaio et al. 2020), and factsheets (Arnold et al. 2018).

4.1 Trustworthy AI architecture

As shown in Fig. 2, a pipeline for checking the TAI requirements at every stage of the model's life cycle is proposed. Note that the emphasis in this figure is on ML models, as this is now the most popular direction in AI. Applying TAI requirements at each phase of the pipeline creates what is known as a chain of trust, since trust at one point will manifest in subsequent phases. It is recommended to optimize the pipeline as a whole rather than each step individually (Cearns et al. 2019). Recent research studies successfully combined ML and metaheuristics optimization techniques such as swarm intelligence and proved to be able to obtain outstanding results in different areas (Bacanin Ruxandra et al. 2021; Malakar et al. 2020)). An AI system can be formulated as $y = f_\theta(X)$, where y is the target or dependent feature, X is the list of inputs or independent features, f_θ is the AI system parameterized by θ. Inspired by Toreini et al. (2020), we consider data-centric, model-centric, and user-centric trustworthy. The study measures the level of acceptance for the AI system by healthcare users, including patients, physicians, and caregivers. This acceptance may include the suitability of explanation, privacy, and accuracy. It may also regard to the ease of use for the model interface and the level of interoperability between the CDSS and EHR.

4.1.1 Data-centric trustworthy

Data is a crucial part of any AI-based system. The efficiency of any system depends mainly on the quality of data as much as on the algorithm's capabilities (Ashmore et al. 2019). Creating high-quality datasets is the first step in any AI pipeline, which consumes most of the time and effort of the project team (Ashmore et al. 2019). Data availability, privacy, malicious training data, and data biases are examples of the challenges to cope with (Thiebes et al. 2021). The following four steps are proposed to ensure data quality: (1) *Data standardization, coding, and unification:* Heterogeneous data are collected from the distributed EHR system and need to be standardized using standard ontologies like SNOMED CT, LOINC, NDFRT, and ICD 11 and unified using standard data models like HL7 FHIR and openEHR. The challenge here is to solve the interoperability problem between heterogeneous EHR systems. (2) *Problem definition:* The main reason CDSSs are not applicable in the medical domain is that engineers do not understand the medical problem. This makes the resulting system either very shallow or extremely sophisticated, which does not convince domain experts. To achieve this step, data scientists must work closely with medical domain experts to understand the problem, determine the data requirements, collect and interpret data via secure transmission media, meet fairness and robustness requirements, and define the generalization performance requirements and metrics. Data can be collected from multiple sources, including distributed EHR databases. The collected data must satisfy four main properties: relevance, completeness, balance, and accuracy. Ashmore et al. (2019) comprehensively surveyed these properties and how to measure and enhance them. Collected data could be stored on a local server or the cloud. (3) *Data preparation and data augmentation:* This step includes a range of data cleaning activities, missing values imputation, outlier detection, data normalization, data fusion, time-series data preparation, and data balancing. Different exploratory data analysis (EDA) techniques

facilitate understanding data distributions and correlations to uncover biases and select the best preprocessing methods. Data leakage makes the training dataset irrelevant because the data will include information that will not be available to the system after deployment. EDA techniques can identify sources of leakage. For example, a tight correlation between a feature and a label may indicate a leakage (Ashmore et al. 2019). The most critical issue is to apply data preparation steps on training datasets only to prevent data leakage. Imprecise labeling of the data leads to severe loss in the quality of the model. Data augmentation is required to appropriately label the data and add more samples to the collected data. To assign labels, domain experts, clustering, and reinforcement learning can be utilized. (4) *Feature engineering and optimization:* Feature selection (FS) and enrichment enhance model performance and its interpretability. The use of genetic algorithms, recursive feature elimination, or Bayesian optimizers has recently become popular. Many automated ML tools (He et al. 2021) and techniques have been published to do the data preparation and feature engineering steps easily, automatically, and correctly.

4.1.2 Model-centric trustworthy

Building and deploying a simple, robust, and fair AI system is challenging in sensitive domains like medicine. Model availability, privacy, uncertainty, bias, and opacity are examples of challenges of TAI (Thiebes et al. 2021). The following six steps are considered: (1) *ML algorithm selection and optimization:* This step involves validating multiple algorithms like decision trees (DTs), SVMs, and logistic regression. In addition, static and dynamic ensembles and DL neural networks with different architectures may achieve better results. The selected algorithm needs to be validated and optimized using a hyperparameter optimization technique like grid search, random search, Bayesian optimization, genetic algorithms, or particle swarm optimization (Yang and Shami 2020; Yu and Zhu 2020). Many AutoML tools exist to automate the ML pipeline and select the algorithm with the best hyperparameters θ that achieve the best validation performance with the best model complexity. (2) *Build explainability features:* Interpretation of glass-box models regarding feature importance, visualization, CBR, and natural language generation ((Alonso and Bugar 2019; Mariotti et al. 2021)), which provide post-hoc explanations of black-box models. (3) *ML model (offline) testing (i.e., model verification):* ML model f_θ should generalize well to previously unseen data and reasonably handle edge cases. Based on domain expert priorities, specific performance metrics need to be defined and measured. Testing data should be collected from the same distribution as training data in addition to adversarial samples for robustness testing. If the generalization performance requirements are not reached, data preparation and model training steps should be repeated. (4) *ML small scale deployment:* Because the ideal scenario testing should be done in a real-life setting, and because of the safety, privacy, and accuracy constraints, the resulting AI system should be first deployed and tested in a small-scale real environment, such as single department in a hospital or small clinic where domain experts check that the AI system adheres to medical regulations. Afterwards the system should be also tested with external data from other hospitals. Real domain experts or external committees need to test the system's explainability, reliability, fairness, privacy, robustness, and auditability. The integration of CDSSs in the EHR ecosystem needs to be verified too. (5) *AI system full deployment:* The accepted model by medical experts is fully deployed in the real hospital and integrated as a component in the EHR ecosystem. The interoperability between AI-based CDSSs should be implemented, and the online learning setting should be defined. Data access controls

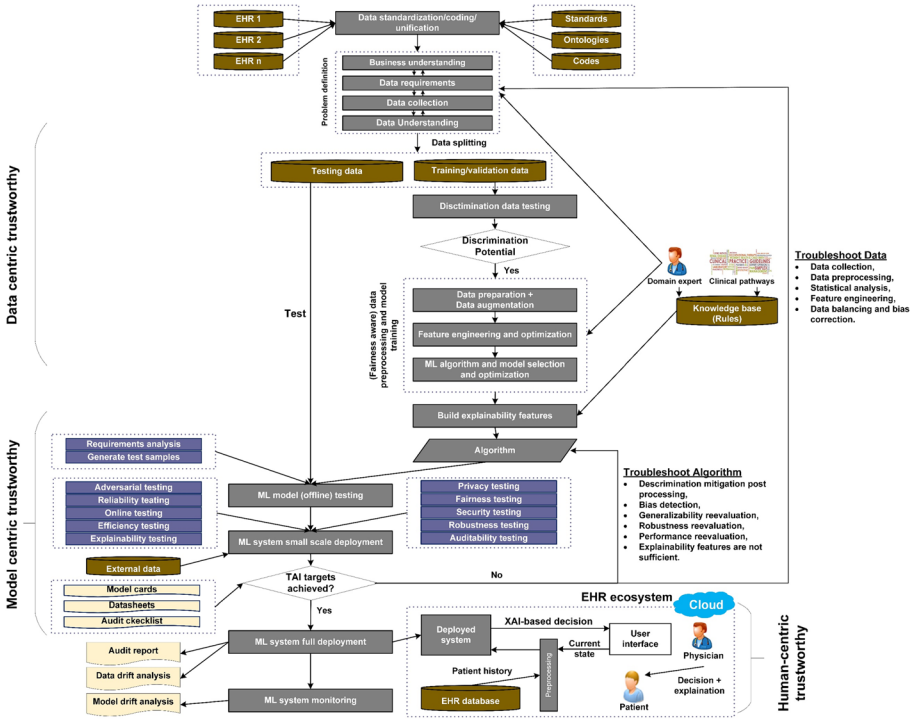

Fig. 2 TAI pipeline

must be defined to determine who (and when) can access to CDSSs. There is no standard for transferring an AI system from the training to the production phases. The system can be deployed as REST services, Flask-based system, mobile app, or cloud service. Online query data should be (a) passed through the same preprocessing steps, (b) integrated with other EHR data in a standard way. (6) *AI system monitoring:* Complex AI systems require continuous monitoring and maintenance. The dynamic training using online learning may cause the system to learn wrong patterns over time. This is called concept drift which is a change observed in the joint distribution $p(X, y)$, for X input and y output because of re-tuning of θ using newly arrived data. This phenomenon could have major adverse effect on model performance even happen in a microscopic scale (Celik and Vanschoren 2021). In addition, a model update is required to make sure it reflects the most recent trends in data and medicine. Based on the practical considerations, the model can be scheduled for regular retraining. Domain experts should define what metrics to monitor the data and system continuously and when and how to perform the updates. Notice that the software engineering dimension of the system deployment is out of scope of the study.

4.1.3 User-centric trustworthy

This is the last phase in the TAI pipeline. Users mainly include patients and physicians who interact with the AI system. The homogeneity level of integration of AI-based CDSSs in medical workflow affects the acceptance level of users. If the system works in isolation

from EHR, this requires physicians to manually interact with the system, which takes a long time for domain experts. Responsibility and accountability are crucial requirements for CDSSs, where domain experts need to evaluate the traceability and auditability features of the deployed system. The ease of use of CDSSs should also be evaluated (Budd et al. 2019). In addition, the context-aware personalized explainability features should be evaluated by users in a real environment and over time. The level of generated details, the number of explainers, the complexity of explanations that depend on the level of experience of receivers, and the availability of time for their investigation should be considered.

4.2 Robustness methods

The model accuracy has been the long-lasting standard for comparing models. Recently, model robustness became a more crucial measure for model acceptance and evaluation. Although AI has a prominent role in improving healthcare, there are still lingering doubts regarding the robustness of these techniques in healthcare settings (Qayyum et al. 2021), primarily because of security and privacy issues resulting from adversarial attacks. Achieving robustness for AI-based systems requires handling of three main challenges, including (1) security and privacy, (2) accuracy, and (3) reliability and reproducibility (Xiong et al. 2021). Most literature concentrated only on the accuracy of the model as a measure of robustness ((Abuhmed et al. 2021; Liu et al. 2021a)), but there is always a trade-off between performance and robustness (Su et al. 2018) which should not be disregarded. This section briefly discusses these requirements.

4.2.1 Security and privacy

AI-based systems are vulnerable to different adversarial attacks at every pipeline phase (Xiong et al. 2021). Undesirable effects include performance degradation, system misbehavior, and privacy breach. The robustness of an AI system is the ability to resist malicious attacks and protect its integrity, availability, and confidentiality. Adversaries can attack an AI-based CDSS in many ways (Arnold, et al. 2018) (Finlayson et al. 2019). For example, (1) small perturbation in MRI images could cause the CDSS to misclassify patients to any label that the attacker desire, (2) training data and models can be poisoned to reduce model performance, and (3) sensitive information about model and training data can be extracted by observing outputs for different inputs. Discovering that AI-based CDSSs are not secure and private significantly hinders their practical deployment (Qayyum et al. 2021). Yuan et al. (Xiaoyong Yuan et al. 2019) discussed different attack mechanisms on DL systems at the training and validation stages. Su et al. (Jiawei Su and Vargas 2019) discussed a DL vulnerability in an image classification task where a perturbation of just a single pixel could dramatically modify the classifier performance and facilitate attacks. Pitropakis et al. (Nikolaos Pitropakis et al. 2019) surveyed existing ML vulnerabilities and associated attack methods. Huang et al. (Huang et al. 2020) comprehensively surveyed the safety, XAI, verification, testing, and adversarial attacks of DL models. Huang et al. surveyed the mathematical formulation of robustness. Let f be an ML model, $C(f)$ is the correctness of f, and $\delta(f)$ be f with perturbations on any of its components such as data, model, or learning framework. Robustness of f is the measurement of the difference between $C(f)$ and $C(\delta(f))$. There are two main categories of robustness: *(1) local robustness* that measure the robustness for specific test input. Let x be a test input for ML model f. Let $\hat{x}$ be an adversarial input of x. Model f is δ-local robust at input x if for any $\hat{x}$, $\forall \hat{x} : \|x - \hat{x}\|_p \leq \delta \rightarrow f(x) = f(\hat{x})$,

for $\|.\|_p$ represent the p distance; *(2) global robustness* measures the robustness for all inputs. Model f is ϵ-global robust for any x and $\hat{x}$ if$\forall x, \hat{x} : \|x - \hat{x}\|_p \leq \delta \rightarrow f(x) - f(\hat{x}) \leq \epsilon$. Barreno et al. (Barreno et al. 2010) proposed a popular taxonomy for vulnerabilities and attacks on ML-based systems, which has three dimensions:

(1) *Influence*: either the attempt to control training data (causative or poisoning attack) or new data is compromised (explorative or evasion).
(2) *Security violation*: either the exploitation concentrates on the attack for increasing false negative rate (integrity attack), overload of false positives (availability attack), or the unveiling of sensitive information of training data or trained model (privacy violation attack).
(3) *Specificity*: either the attack is targeted to a particular instance (targeted attack) or a wider class (indiscriminate attack).

4.2.1.1 Threat modeling For managing effectively and systematically possible vulnerabilities to an AI system, a *threat model* should be adopted to identify potential system-specific threats, suitable security objectives, relevant attack, and vulnerability vectors, and design appropriate defense methods. Threat modeling in the ML domain focuses on the following aspects.

(1) *Attack surface*: this is the type of potential attack in every AI pipeline stage, including poisoning attack during model training, testing, and retraining; model stealing; evasion attack in prediction; gradient descent attack in learning; and polymorphic attack in feature extraction.
(2) *Attacker goals*: they may include (a) *security violation* where attackers could undermine the system functionality (integrity and availability), e.g., significantly degrading the model accuracy, or deducing sensitive information about the system (confidentiality and privacy), e.g., obtaining model parameters or training data, (b) *attack specificity* where attacker launches a targeted or indiscriminate attack against specific or any system, and (c) *error specificity* where attacker either fool the system to misclassify an input sample to a specific class or any class different from the legitimate class.
(3) *Attacker knowledge*: attackers could have different access levels to sensitive information about AI systems. This information includes training data, algorithms and architecture, hyperparameters, objective function, and trained model parameters. The type of attack (black box, gray box, and white box) depends on the level of access to sensitive information. If the attackers know everything about the model, they can run white-box attacks. Suppose attackers know some information like feature set and algorithm and its architecture, but not the training data and parameters. In that case, they can run gray-box attacks to collect surrogate datasets from similar sources and collect label outputs from the model for these data. If attackers do not know anything about the model, they can run black-box attacks. Based on the normal distance that evaluates the difference between original and perturbed input, adversarial attacks can be classified as L_0, L_1, L_2, and L_∞-attacks (Huang et al. 2020).
(4) *Attacker capability*: the attacker's ability to access and manipulate training data and input examples or monitor the corresponding output of a trained model.
(5) *Attacking effect (causative or exploratory)*: causative effect occurs when an attacker can manipulate training and input samples during training and online learning. This attack

aims to corrupt the model under training to cause either integrity violation (which makes the model produce adversary desired output) or availability violation because of the corrupted model. The exploratory effect occurs when attackers can manipulate input samples only during the prediction phase to cause the model to produce incorrect outputs or violate privacy for deducing sensitive information about the model or data.

4.2.1.2 Security attacks Adversarial attacks exist in every stage of the AI pipeline (Xiong et al. 2021). These attacks are based on manipulating and/or extracting input data, training data, or models (Qayyum et al. 2021; McGraw et al. 2019). Adversarial sampling (i.e., adversarial input perturbation) based attacks are the most widespread attacks in the ML domain where the attacker intentionally perturbs a small portion of train/test/input data to adjust the availability, integrity, and confidentiality of the system. Because the ML model is based on relatively small training and test sets compared to the whole population, the large space of the data distribution remains unexplored by the learned model (Ben Braiek and Khomh 2020). This phenomenon becomes more critical for sophisticated models like DL neural networks, which have complex decision boundaries. This complexity creates vulnerabilities that adversaries can exploit.

Moreover, the decision boundaries of some linear models are extrapolated to many regions of high-dimensional spaces that are untrained. The generalizability of the resulting models is negatively affected (McGraw et al. 2019), and the model does not account for adversarial samples (Xiong et al. 2021). Adversarial ML techniques find the regions where the model exhibits an erroneous behavior. Adversarial sampling is achieved in three main ways: perturbation of valid samples, transferring adversarial samples across different models, and generative adversarial networks (GANs) (Dasgupta and Collins 2019). Huang et al. (Huang et al. 2020) formally defined adversarial examples as follows. Given a trained DL model with its association function $f : \mathbb{R}^{s_1} \rightarrow \mathbb{R}^{s_k}$, a human decision oracle $\mathcal{H} : \mathbb{R}^{s_1} \rightarrow \mathbb{R}^{s_k}$, and input $x \in \mathbb{R}^{s_1}$ with $\underset{j}{\text{argmax}} f_j(x) = \underset{j}{\text{argmax}} \mathcal{H}_j(x)$ that mean DL model works correctly on the original example, and adversarial example is defined as:

$$\exists \hat{x} : \underset{j}{\text{argmax}} \mathcal{H}_j(\hat{x}) = \underset{j}{\text{argmax}} \mathcal{H}_j(x) \wedge \|x - \hat{x}\|_p \leq d \wedge \underset{j}{\text{argmax}} f_j(\hat{x}) \neq \underset{j}{\text{argmax}} f_j(x)$$

where $p \in N, p \geq 1, d \in \mathbb{R}$, $\mathbb{R}^{s_1}$ is the input vector, $\|.\|_p$ is L_p-norm distance, $\hat{x}$ is the adversarial input, $\mathbb{R}^{s_k}$ is the output vector, f is the model with scalar or vector output, and $d > 0$. For example, adding a small amount of adversarial perturbation caused the DL model to incorrectly classify fMRI images from "malignant tumor" to "benign tumor" (Finlayson et al. 2018). Papangelou et al. (Papangelou et al. 2018) introduced the concept of adversarial patients, which are unintentional adversarial examples that could lead to several ethical issues (e.g., patients with identical predictive features but different individual treatment effects). Most adversarial attacks concentrate on computer vision models (Huang et al. 2020). As asserted before, vulnerabilities exist in every phase of the AI pipeline. For a comprehensive survey of these issues, the interested readers are kindly referred to Qayyum et al. (2021). In brief, the sources of these vulnerabilities can be one of the following:

(1) *Data collection vulnerabilities*: the collected data from EHRs, neuroimages, and medical reports introduce vulnerabilities such as instrumental and environmental noises (e.g., MRI image is highly sensitive to motion), unqualified personnel (e.g., build-

ing and maintaining AI systems by a non-technical physician or depending on the physician-data engineer relationship).

(2) *Data annotation vulnerabilities*: labeling samples for supervised ML models is called data annotation. Domain experts or automated unsupervised learning techniques are always used to label data, introducing problems, including class imbalance, label leakage, and coarse-grained labels. The major causes of the labeling challenge are that the medical ground truth is ambiguous and the perturbance of data by malicious users. AI systems can be used to detect rare diseases and to detect hidden patterns in the medical environment. However, data are usually limited, imbalanced, biased, and sparse. These issues further complicate data labeling.

(3) *Model training vulnerabilities*: incomplete or improper training, privacy breaches, data poisoning, model poisoning, and stealing are popular issues in model training. Improper training involves using improper parameters like learning rate, batch size, or epochs in the learning process. *A poisoning attack* happens during the model training and retraining phases, when the attacker injects some "poisoned" samples into train/test sets to modify the statistical properties of the data. This attack affects the integrity and availability of the model, where an attacker can launch error-generic or error-specific poisoning attacks. Model retraining takes place in applications like intrusion detection, spam detection, and facial recognition to reconsider new updates in the domain. At that point, poisoned data can be fed to the operational system, and an attacker can launch white or black-box attacks during the (re)training process based on the level of access (Liu et al. 2018a). Label-flipping, gradient descent, backdoor, and Trojans attacks are the main types of poisoning attacks that affect supervised learning algorithms (Liu et al. 2018a; Biggio and Roli 2018; Gardiner and Nagaraja 2016).

(4) *Testing vulnerabilities:* These issues are related to the interpretation of the results of models, including misinterpretation, false positive, and false negative outcomes. An *evasion attack* is an adversarial attack against the system in the prediction phase. This attack evades the trained model by adversarial input examples (Biggio and Roli 2018). These examples can be generated using many algorithms. Adversarial sampling techniques include the fast gradient sign method (Tramèr et al. 2017), universal attack approach (Moosavi-Dezfooli et al. 2017), DeepFool (Moosavi-Dezfooli et al. 2016), among other algorithms (Xiong et al. 2021). Gradient-based attacks use a gradient descent function to search for adversarial examples that can mislead the system. This attack is widely used in attacking differentiable learning algorithms such as DL and SVM with the differentiable kernel. For non-differentiable learning as a DT, an attacker can use differentiable surrogate models.

(5) *Deployment vulnerabilities*: the most critical issues in the deployment phase are distribution shifts and incomplete data. *Distribution shifts* are expected because the distribution of data used in model building is very different from the data distribution in a real environment. This causes significant degradation in model performance. In addition, this difference can be exploited to generate adversarial examples. For *incomplete data*, in real settings, the collected data for patient care usually contain missing observations and variables. Handing this incompleteness is a challenge because data may not be missing at random.

4.2.1.3 Defense mechanisms Defense techniques are dual to attack techniques by either improving the robustness of the model to immune the adversarial cases or differentiating

the adversarial cases from the correct ones. There are many countermeasures for different attacks. Wang et al. (2019a), Qayyum et al. (2021), Xiong et al. (2021), Huang et al. (2020), and Liu et al. (2018a) provided comprehensive surveys for these mechanisms, which are categorized as follows:

(1) Modifying model: they concentrate on training a robust model against adversarial attacks by using different methods like adversarial training, data compression, foveation-based method, gradient masking, network verification, classifier robustifying, explainability modeling, and defensive distillation methods (Qayyum et al. 2021; Ben Braiek and Khomh 2020; Wang et al. 2019a). Another method adds an external defense layer in front of the trained model. This layering process inputs data before they are used by the trained models in the prediction phase. These techniques include input monitoring against anomaly inputs that detect and filter adversarial inputs and input transformation maps input samples far from the training data in feature space (McGraw et al. 2019; Biggio and Roli 2018). Another technique is the masking model technique, where the adversarial attack is formulated as a learning plus masking problem (Nguyen et al. 2018), where the masking step introduces noise in the logit output to defend against attacks.

(2) Modifying data: these techniques do not touch the model but modify the data or features to prevent adversarial attacks. Related techniques include *adversarial (re)training* (i.e., training the model again using augmented training data with adversarial examples. This method fails to catch iterative adversarial perturbation generation methods as basic iterative method (BIM) (Kurakin et al. 2018)), *input reconstruction* (i.e., cleaning adversarial noises from input examples using techniques like denoising autoencoder which removes harmful effects on the model (Gu and Rigazio 2015)), *feature squeezing* (i.e., squeezing features that an adversary can use to construct adversarial examples (Xu et al. 2017)), *feature masking* (i.e., masking the most sensitive features by adding a masking layer before the classification layer to set the corresponding weights to zero (Gao et al. 2017)), and *data sanitization and robust learning* (i.e., attacker alters the statistical distribution of training data or insert adversarial examples (Biggio and Roli 2018)).

(3) Adding auxiliary model(s): These methods integrate auxiliary models along with the main model for detecting adversarial attacks. Standard techniques in this category include adversarial detection (i.e., a binary classifier is added to distinguish between adversarial and original examples (Lu et al. 2017b)), ensembles defenses (i.e., sequentially or in parallel integration of multiple defensive strategies such as PixelDefend (Song et al. 2017)), and using GAN models (i.e., using GAN to defend against adversarial attacks such as Defense-GAN (Samangouei et al. 2018)). Wang et al. (Wang et al. 2018a) proposed model mutation testing to detect adversarial examples at runtime. They used mutation operators like Gaussian Fuzzing and Neuron Switch to randomly mutate the DL model and compute the label change ratio of genuine and adversarial data. Authors discovered a higher label change ratio for adversarial examples under model mutation than for original examples. As a result, they proposed using this change to decide whether an input is adversarial or genuine at runtime.

Braiek and Khomh (Ben Braiek and Khomh 2020) provided a survey of the testing techniques for ML models as software programs. They surveyed the possible errors in models and the techniques to detect or mitigate these errors. The approaches for detecting potential errors in models are categorized into: (1) *White-box testing approaches* that consider the internal implementation logic of the model. These techniques include pseudo-test oracle as differential testing and metamorphic testing; adequacy evaluation testing as neuron coverage, modified conditions/decision coverage, and surprise adequacy; and automated input test generation as gradient-based optimization, GAN-based generation, constrained

programming solver, and differentiable fuzzing. (2) *Black-box testing approaches* that do not need access to the internal implementation details. These techniques include adversarial and mutation testing. Zhang et al. (Zhang et al. 2019a) asserted that there must be offline testing to detect bugs while modeling development and online testing for bug detection after the model deployment. The tested components include data, learning model, and used framework (e.g., TensorFlow, scikit-learn, and Weka). For each component, Zhang et al. identified possible bugs. For example, data bugs include data incompleteness, data not representativeness, data imbalance, biased labels, data poisoning, and data skewness.

4.2.1.4 Privacy preserving Preserving the privacy of patients is paramount. A privacy attack occurs when an adversary uses the model's output to infer values of sensitive features used as input to the model. It is worth noting that privacy in data storage is not considered, either in local servers or in the cloud. Data anonymization can partially solve privacy, but sensitive information can be inferred from anonymized data (Narayanan and Shmatikov 2008). This problem cannot be entirely eliminated because of the statistical nature of ML models. Because a valid model is generalized on inputs that were not used for model training, the model can breach privacy for those cases not considered in the training stage. Private information in the AI pipeline can be attacked using four main techniques (Al-Rubaie and Chang 2019): *reconstruction attack* (i.e., reconstructing raw training data using knowledge about feature vectors), *model inversion attack* (i.e., used responses for inputs sent to the system in an attempt to create feature vectors similar to those used in model training), *membership inference attack* (i.e., infer if an example was a member of the training set), and *model extraction* (i.e., the adversary duplicates the functionality of the system by building an equivalent model based on some obtained predictions from the input feature vectors). These attacks have two main threats: unveiling confidential information and malicious use of data (Qayyum et al. 2021). Equivalently, mechanisms have been proposed to prevent privacy breaches. Privacy can be protected by two main methodologies (Qayyum et al. 2021; Liu et al. 2018a; Wang et al. 2019a; Al-Rubaie and Chang 2019): perturbation mechanisms (e.g., differential privacy and dimensionality reduction methods) and cryptographic mechanisms (e.g., fully homomorphic encryption, Garbled circuits, secret sharing, secure multiparty computation, functional encryption, and crypto-oriented model architectures). Each of these mechanisms applies different techniques. For example, differential privacy, which adds perturbation or noise to the training data for protecting private information, has been implemented using techniques like the private aggregation of teacher ensemble (PATE) (Papernot et al. 2018), differentially private stochastic gradient descent (DP-SGD) algorithm (Abadi, et al. 2016), moments accountant (Wang et al. 2019b) and others (Qayyum et al. 2021).

4.2.2 Correctness

ML models have already achieved human-level performance in some applications in clinical medicine (Qayyum et al. 2021). For example, computer-aided diagnosis systems are fully automated without any human intervention.[1] ML models also benefitted from other technologies like cloud/edge computing, mobile communication, and big data technology. Using these technologies, AI systems can produce highly accurate and patient-centric predictions. However, the tangible effect of AI systems in hospitals has not been seen yet.

[1] https://tinyurl.com/FDA-AI-diabetic-eye

The way of measuring the performance of literature studies implied that medical experts were not trusting these results. In this section, we briefly consider the issues of performance validation techniques. These techniques are mainly based on the size of the dataset, the data division methodology (training/validation/testing), the validation technique used, the number of sources used to collect data, the used model, and performance evaluation metrics (Maleki et al. 2020). Testing error is used to estimate the model's generalization performance (*internal validation*), so the test set must be representative of the accurate distribution of the data. In addition, this set must be separated and untouched from the very beginning of the AI pipeline (before preprocessing step) to prevent data leakage and over-optimistic results. The causes of data leakage have been summarized in Wen et al. (2020). The simplest form of data leakage is making FS and/or missing values imputations on the entire dataset before partitioning it into train/validation/test. Moreover, model testing using external data from other sources provides a more accurate estimation of the generalization performance (*external or multisite validation*).

In the case of supervised ML, given a training dataset $D_{train} = \{X, Y\}$, an input vector x of dimension d as $x \in X \subset \mathbb{R}^d$ and y is a scalar or label target as $y \in Y \subset \mathbb{R}$ or $y \in Y \subset \{c_1, c_2, \ldots, c_l\}$. Let the exact relationship between X and Y be represented by the unknown model f, i.e., $Y = f(X) + \epsilon$. Let $\widehat{f}$ estimates f that predicts $\widehat{Y} = \widehat{f}(X)$, where $\widehat{Y}$ is the prediction of Y. The accuracy of $\widehat{Y}$ depends on the reducible and irreducible errors. In other words, $\mathrm{E}(Y - \widehat{Y})^2 = \mathrm{E}\left[f(X) + \epsilon - \widehat{f}(X)\right]^2 = \left[f(X) - \widehat{f}(X)\right]^2 + var(\epsilon)$ = reducible error + irreducible error, where $var(\epsilon)$ is the variance of ϵ. Independent from the selected algorithm and optimized model, ϵ is called irreducible error resulting from the inherent noise in the data. On the other hand, reducible errors of $\widehat{f}$ can be reduced by optimizing the ML pipeline. Overfitting and underfitting are defined based on the errors defined earlier, and optimizing an ML model requires a trade-off between bias and variance by controlling model complexity (Maleki et al. 2020). This is called the level of model relevance for the data (Zhang et al. 2019a). The reducible error can be divided into bias error and variance error, and the objective function of an ML model reflects these errors.

4.2.2.1 Model selection Robust validation of ML models supports their reproducibility, reliability, and generalizability (Maleki et al. 2020). There are many validation techniques in the literature (Raschka 2018). These techniques are usually combined with a model tuning technique such as grid search, random search, Bayesian optimization, genetic optimization, and practical swarm optimization (Yang and Shami 2020). All components of the ML life cycle are closely bonded, and error propagation is a serious problem (Zhang et al. 2019a), so optimizing the entire ML pipeline, including data preprocessing, feature engineering, and hyperparameters tuning is the best choice. AutoML tools like TPOT[2] are implemented to automate this process (He et al. 2021).

Holdout validation: It is the most common method for evaluating ML models, where the input dataset is randomly (and may be in a stratified manner) divided into three independent partitions with the same distribution, i.e., D_{train}, $D_{validation}$, and $D_{testing}$. Training data D_{train} are used for model building, validation data $D_{validation}$ are used for hyperparameter tuning and testing data $D_{testing}$ are used to measure the generalization performance. The proportion of each set depends on the model, the size of the dataset, and data variability. For example, a large portion (i.e., 70%) is assigned for training, 15% for validation, and 15% for testing (Maleki et al. 2020). To obtain a more robust estimate of performance that

[2] http://automl.info/tpot/

is less variant to how the data are split, the holdout method should be repeated k times with different random seeds, and average performance (along with standard deviation) should be reported, e.g., for accuracy metric, $ACC_{avg} = \frac{1}{k}\sum_{j=1}^{k} ACC_j$, $ACC_j = 1 - \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i)$, where m is the size of the dataset, and $\mathcal{L}$ is the loss function. The repeated holdout is called Monte Carlo cross validation (CV).

Cross-validation: Compared with the holdout method, CV provides a more accurate estimation of generalization error when dealing with small datasets. CV is useful for detecting overfitting, but it is not always clear how much overfitting is acceptable. CV cannot detect overfitting if test data is not representative of possible unseen data (Zhang et al. 2019a). Many versions of CV are available, including k-fold CV, repeated k-fold CV, stratified k-fold CV, repeated stratified k-fold CV, leave-one-out CV (LOOCV), leave-p-out CV, leave-one-group-out CV (LOGOCV), and nested CV (NCV). The basic idea of these validation schemes $\mathcal{V}$ s to estimate the out-of-sample error $\mathcal{L}$ is first to divide the whole dataset $D = D_{CV}$ and D_{test} where $D_{CV} \cap D_{test} = \varnothing$. Secondly, D_{CV} and k splits are used by $\mathcal{V}(D_{CV}, k)$ to divide data into k non-overlapping data splits, i.e., $\mathcal{V}(D_{CV}, k) = \left\{\left(I_i^t, I_i^v\right), I_i^t \cap I_i^v = \varnothing\right\}_{i=1}^{k-1}$, $k-1$ folds are used to train the model, and one fold is used for validation. The estimated validation error is the average out-of-sample loss over all k splits.

$$L\left(D_{CV}, \mathcal{V}\right) = \frac{1}{k}\sum_{i-0}^{k-1} \frac{1}{|I_i^v|} \sum_{(x,y)\in I_i^v} \mathcal{L}\hat{f}(x), y)$$

Thirdly, the D_{test} is used to measure the generalization performance of the validated model. The state-of-the-art method for model validation is the nested CV (the reader can refer to Raschka (2018) for more details). In all CV techniques except NCV, when data are divided into training and validation sets, experimenting with several models, and searching for the best hyperparameters usually makes the resulting validation error overoptimistic if used to estimate generalization error. The solution for this issue is that a test set should be locked and not be used for model training and hyperparameter tuning. Taking a single test set from a small input dataset estimates a generalization error with high variance and sensitivity to the selected test set. This solution is implemented in NCV, which has an outer CV loop and an inner CV loop. The outer loop uses different train, validation, and test splits. The inner loop takes training and validation sets chosen by the outer loop (Maleki et al. 2020). The inner loop train and validate the model, and the out loop tests its generalization error. In online learning, the model can gradually be overfitted by learning from new data because D_{test} cannot guarantee to represent the new online data.

4.2.2.2 Statistical analysis Statistical analysis is crucial to understand the dataset before starting the learning process to study the feature's normality, skewness, kurtosis, correlation, and statistical significance for model prediction. Moreover, the statistical analysis of the performance of the learned model is critical to measure the stability and certainty of this model. Measuring model stability can be achieved by repeating the holdout or CV process with different random seeds. Results should be reported with a central tendency (e.g., mean) and variation (e.g., variance). Another good practice is to compute the confidence interval around a performance estimate to judge the model's certainty or uncertainty (Raschka 2018). Under the normal approximation, the confidence interval can be calculated by $\text{ACC} \pm z\sqrt{\frac{1}{n}\text{ACC}(1-\text{ACC})}$, where α is the error quantile (e.g., $\alpha = 0.05$ for 95% confidence) and $z = 1 - \frac{\alpha}{2}$. Bootstrap method could be used to calculate the confidence interval

if we do not assume a normal distribution of the results. For a dataset of size n, and b bootstrap rounds, $\mathbf{ACC_{boot}} = \frac{1}{b}\sum_{j=1}^{b}\frac{1}{n}\sum_{i=1}^{n}(1-\mathcal{L}(\hat{y}_i, y_i))$, $\mathrm{SE_{boot}} = \sqrt{\frac{1}{b-1}\sum_{i=1}^{b}(\mathrm{ACC}_i - \mathrm{ACC_{boot}})^2}$, and confidence interval of $\mathrm{ACC_{boot}} \pm t \times \mathrm{SE_{boot}}$, for t comes from t-table, $\mathrm{ACC_{boot}}$ is average accuracy of b bootstraps, and SE_{boot} is standard error of all accuracies ACC_i. Confidence intervals must be reported for every model.

4.2.2.3 Model and algorithm comparison Most studies compare models by using no statistical test. These studies build models and measure their performance on a test set. Nevertheless, this method provides insufficient evidence about whether there exist true differences in models' performance. Gardner and Brooks (Gardner and Brooks 2017) discussed in detail what are the valid and invalid methods for making model comparisons. To determine the statistically significant differences of multiple models, a suitable statistical hypothesis testing approach should be selected. This selection depends on the size of the training dataset. The statistical test could be based on target predictions for independent test sets or fitting and evaluating CV models. Suppose that there is a separate test set to compare the performance of two classifiers. In that case, the difference of two estimated generalization accuracies can be compared, where we select a confidence interval (e.g., 95%) and assume a normal approximation. The simplest case is to use the z-score test. For example, if the 95% confidence intervals of the accuracies of the two models do not overlap, then reject the null hypothesis that the performance of the two models is the same. For using the same test set to compare two models, it is better to use a paired test (e.g., paired Student t-test) or, more accurately, use the McNemar test. For a full discussion about these techniques, readers are guided to Raschka (2018). For comparing more than two classifiers, we use multiple hypotheses testing approaches. For doing that, we first do an omnibus test (e.g., ANOVA, Cochran's Q test, or F-test) under the null hypothesis that there is no difference between classifiers. Then, if rejected, pairwise post-hoc tests are used to determine the different classifiers. However, post-hoc tests are regarded as fishing expeditions because there is no clear hypothesis of which models should be compared. As a result, all possible pairs must be compared, which leads to the multiple hypnosis problem, so a correction term (e.g., Bonferroni's correction) should be used to reduce the false positive rate in multiple comparison tests by adjusting the significance threshold α.

Model comparison tests do not consider the variance of the training sets into account. Algorithm comparison techniques are used for comparing sets of models where each set has been fit to different training sets. Many techniques are available, including 5×2-Fold CV, k-hold-out paired t-test, k-fold CV paired t-test, Dietterich's 5×2cv paired t test, and Alpaydin's combined 5×2cv F-test. These techniques are applied to results previously collected from many runs of the model based on independent test sets with CV techniques. The robustness of these techniques depends on the number of collected results and the normality and equivariance of such results. Thus, two main statistical hypothesis testing techniques can be used, including parametric tests (e.g., Student's t test and ANOVA) and non-parametric tests (Mann–Whitney U Test, Wilcoxon Signed-Rank Kruskal–Wallis H Test, and Friedman Test). The mathematics behind these techniques can be found in Raschka (2018), and a full description of model comparison techniques is given in Dietterich (1998). All these techniques have been implemented in the Python package of MLxtend.[3]

[3] http://rasbt.github.io/mlxtend/

4.2.3 Reliability and reproducibility

Building a reliable and reproducible model depends on the attributes of the input dataset, including size and variability, data balancing, and the availability of data and code.

4.2.3.1 Dataset size Because of the rarity of most phonotypes under study, limited resources, patient privacy, limited data preparation and annotation expertise, and other legal issues, collecting large datasets is not straightforward in the medical domain. As a result, most researchers in the medical domain work with small datasets. Dividing small datasets into training, validation, and testing reduces the model training and validation dataset. Furthermore, it leads to an unreliable estimation of the generalization error and accordingly hinders reproducibility. Patients in a small dataset tend to be more homogenous, not truly representing the intended population. There are many solutions to this issue, including: (1) using the LOOCV or LOGOCV approaches for model validation, but LOOCV suffers from high degrees of correlation between samples; (2) integrating public datasets with the local datasets; (3) increasing number of samples using patch-based approaches to select several patches from a single image; (4) using data augmentation for model training and validation, for example, using GANs. The geographic, demographic, and phenotypic diversity increases as the sample size increases. As a result, validation performance decreases, and generalizability increases.

4.2.3.2 Data variability In the medical domain, data are always collected using different devices and under various circumstances. For example, MRI images may be collected in the AD domain using different scanners with different resolutions, such as 1.5 T or 3 T, and using a scanner from several vendors. Data collected from separate hospitals may vary because of varying scanner settings, disease prevalence, and various protocols or standards. In addition, collected data will have different noises. The distribution of training, validation, and testing sets should represent these variations. It is a good practice to use data from some hospitals to train and optimize the model and use data from other hospitals to test the generalization performance of the learned model. However, this implies interoperability among information systems in different hospitals, which is not always true. The testing performance estimates the generalization performance, and the amount of shared and transferred data is reduced, which improves data privacy. One challenge for this technique is the data preparation because data may be encoded and stored using different techniques and standards. This results in a model tested in a different context from the one it was trained in.

4.2.3.3 Reproducibility requirements Reproducibility means obtaining the same results specified in a paper using the same data, code, and identical conditions. The repeatability of the experimental process is a crucial requirement to verify the reliability and consistency of research findings, especially in the medical domain (McDermott et al. 2021). More than 70% of researchers failed to reproduce other researchers' results, and over 50% of researchers failed to reproduce their results (Pineau, et al. 2019). This is a major challenge because (1) it is difficult to access the same training data or data with similar distribution; (2) trained model, model parameter and hyperparameters, model architecture, and code necessary to run the experiment are usually not available; (3) performance metrics and results are not always clearly specified; (4) statistical analysis of the results is sometimes done wrongly; (5) selective reporting and over-claiming of results; (6) cohort description and data preprocessing specification are not available; (7) random seed used to train model is absent; (8) analysis of model complexity

(time, space, sample size) is not clear enough; and (9) statistics, splitting approaches, and sample excluding policies of utilized datasets are not defined. The issue is more difficult when working with DL models due to several factors: (1) random weights initialization, (2) dataset shuffling, and (3) diversity of DL frameworks (e.g., changing between Theano and TensorFlow backends in Keras). As a result, in 2019, the international conference on Neural Information Processing Systems (NeurIPS), which is the premier international conference for ML research, highlighted the need for *ML reproducibility checklist* as a part of paper submission process.

Tatman et al. (Tatman et al. 2018) categorized reproducibility as low reproducibility (sharing of model characteristics), medium reproducibility (sharing code and data), and high reproducibility (sharing code, data, and complete computational environment). McDermott et al. (2021) categorized reproducibility into technical reproducibility (i.e., the ability to replicate the results in the paper using shared code and data), statistical reproducibility (i.e., the results are not significantly different under random resample conditions, the variance around results is reported, or internal validation), and conceptual reproducibility (i.e., how well the needed results can be reproduced in conditions that match the abstract description of the purposed effect, testing the system with data from multiple institutions, or external validation). Full reproducibility needs all these three categories. Reproducibility is an integrated requirement for TAI because it increases the transparency and quality of research process by sharing the code, data, results, and scientific communications. However, reproducibility requirements in the medical domain contradict other TAI requirements, such as security and privacy. Datasets may not be shared because of privacy and sensitivity reasons; code may not be shared because it contains intellectual property; running the code may need enormous computational resources (time and number of powerful machines). The publicly available medical datasets such as MIMIC-III and ADNI are frequently used, leading to possible dataset-specific overfitting. Even worse, medical studies did not share code or include a full description for reproducibility (McDermott et al. 2021). Nowadays, several services are used to host data and code, including for-profit services (e.g., Kaggle kernels, Google Collaboratory, Amazon SageMaker, IBM Watson Studio, and Microsoft Azure Notebooks) and not-for-profit services (e.g., MyBinder and Codalab). However, there is a lack of a centralized repository of trained models. Some well-known traditional file repositories include ModelZoo[4] on GitHub and ModelHubAI.[5]

4.2.4 Robustness quantification

Trained models should be evaluated by model performance metrics but also by their capability to resist adversarial attacks (McDaniel et al. 2016). Performance evaluation metrics for classification and regression are well known, and their selection depends on the task and the data balancing issue. Japkowicz (2006) provided a comparison between these metrics. However, there is a lack of standard assessment methodologies and metrics for measuring resilience to adversarial attacks. Performance evaluation metrics like accuracy, precision, recall, and F1-score can be used to assess security robustness by measuring the level of performance degradation after the model suffers from various adversarial attacks (Xiong et al. 2021). This is called the security evaluation curve (Biggio and Roli 2018). For example, Dunn et al. (2020) used accuracy, precision, and false positive and true positive rates to measure the negative impact of poisoning attacks on model integrity. Biggio et al. (2013) proposed a security evaluation framework that could be applied to various classifiers. Katzir et al. (2018) proposed a security robustness metric called model robustness score to measure the relative resilience of

[4] https://github.com/BVLC/caffe/wiki/Model-Zoo
[5] http://modelhub.ai/

models. This method is based on two concepts of total attack budget and feature manipulation cost to model an attacker's abilities. Bastani et al. (2016) proposed three robustness evaluation metrics: (1) pointwise robustness, which measures the minimum input change a classifier fails to be robust; (2) adversarial frequency to measure how often the change in input changes a classifier's performance; and (3) adversarial severity that measures the distance between an input and its nearest adversarial example. Some studies calculated the upper bound of the robustness of models (Carlini and Wagner 2017) or the global robustness upper bound and lower bound using test data (Ruan et al. 2019). Adversarial input generation is a widely used method to test the robustness of models (Zhang et al. 2019a). For reproducibility evaluation, there are three categories of metrics, including technical reproducibility metrics (i.e., code available and public dataset), statistical reproducibility (i.e., variance reported), and conceptual reproducibility (i.e., multiple datasets) (McDermott et al. 2021).

4.2.5 Robustness tools

There are tools for building robust models. TensorFlow Federated[6] supports distributed ML for training global models without sharing the data. CrypTen[7] is a PyTorch-based package to train ML models based on encrypted data. OpenMined[8] provides various libraries, including PySyft[9] PyGrid,[10] and SyferText[11], for building privacy-preserving ML. DeepTest[12] is a tool for automatic testing of DL models for autonomous cars. This tool uses the concept of neuron coverage to test CNN and RNN models. PySyft extends PyTorch, TensorFlow, and Keras to provide differential privacy, federated learning, and encryption. PyGrid extends PySyft to provide a peer-to-peer network for collectively building ML models. SyferText provides preserved NLP tasks. Foolbox[13] is a Python library for testing ML models like DL neural networks against adversarial attacks. It can work with models in PyTorch, TensorFlow, and JAX. Adversarial Robustness Toolbox[14] is a Python library for ML security. It is a set of tools to evaluate, defend, and verify ML models against adversarial threats. AdvBox[15] is a Python tool set for ML model security, including generating, detecting, and protecting adversarial examples. CLEVER[16] is a metric for measuring the robustness of DL models. It is attack-agnostic, which can be computed efficiently for large models like ResNet-50 and Inception-v3. Ditto[17] is a Python package to build robust models against data and model poisoning attacks. DeepCheck (Gopinath et al. 2019) is a lightweight symbolic-execution-based approach to performing symbolic analysis of DL models. It transforms the model into a program by translating the activations into IF-ELSE branch structures to create the program paths to be used to identify important pixels and crate

[6] https://www.tensorflow.org/federated
[7] https://github.com/facebookresearch/CrypTen
[8] https://www.openmined.org/
[9] https://github.com/OpenMined/PySyft
[10] https://github.com/OpenMined/PyGrid
[11] https://github.com/OpenMined/SyferText
[12] https://github.com/ARiSE-Lab/deepTest
[13] https://github.com/bethgelab/foolbox
[14] https://adversarial-robustness-toolbox.readthedocs.io/en/stable/
[15] https://github.com/advboxes/AdvBox
[16] https://github.com/IBM/CLEVER-Robustness-Score
[17] https://github.com/litian96/ditto

1-pixel and 2-pixel attacks. Adversarial Robustness Evaluation for Safety (ARES)[18] is a TensorFlow-based Python library for adversarial ML focusing on benchmarking adversarial robustness on image classification. It implements 15 attacks and 16 defenses. DeepXplore[19] automatically identifies erroneous behaviors in DL models, using differential testing, without requiring manual labeling. Bugbug[20] is a Python package that aims to detect bugs in ML programs, ML model's quality management, and defect prediction. TFX[21] is a TensorFlow-based Google-production-scale ML platform. It provides libraries to integrate common components needed to define, launch, and monitor ML systems. ActiveClean[22] and Alphaclean[23] provide an optimized and iterative way for data preprocessing and cleaning.

4.3 Fairness methods

Integrating ML with CDSSs and EHRs may introduce biases related to missing values, patients not identified by the algorithm, insufficient sample size and underestimation of the minority class, and misclassification (Gianfrancesco et al. 2018). A decision is unfair if it treats individuals with similar decision-related features differently based on personal characteristics that should not affect the final decision. There are many ways to introduce bias to AI systems. For example, transferring the model from one population to another with different distribution of features; intentionally or malicious introduction of bias to skew the performance; the structure of algorithms and equipment used to collect data may introduce bias too. There are many sources of bias, including historical, representation, evaluation, measurement, aggregation, population, sampling, linking, and deployment biases. Mehrabi et al. (Mehrabi et al. 2019) provided a comprehensive survey for 23 data biases and six discrimination challenges. However, AI systems must generate fair and non-discriminating outcomes. Many researchers, governments, and policies like general data protection regulation (GDPR) called for social understanding of ML (Caton and Haas 2020) because model unfairness negatively affects its robustness and interpretability (Ignatiev et al. 2020). Fairness is domain-specific because some features like gender might be considered a protected feature in trading but not in medicine. Fairness-aware AI-based CDSSs must cope with different biases (e.g., in data, algorithms, or output) based on domain expert knowledge. Most fairness techniques are based on the notion of protected or sensitive features (SFs) (these terms are used interchangeably) and on (un)privileged groups (groups defined by one or more SFs that are disproportionately (less) more likely to be positively classified). Generally, SFs include ethnicity, gender, age, color, nationality, religion, or socio-economic group. AI systems are not fair or unfair by nature, but they can become unfair for many reasons (Toreini et al. 2020), as there is a tradeoff between model accuracy, explainability, and fairness. For instance, the disclosure of SFs could negatively affect patients' privacy (Fung et al. 2010). Chang and Shokri (Chang and Shokri 2020) observed that the more biased the training data is, the higher the privacy cost to achieve fairness for unprivileged subgroups. Thus, fairness is manipulated as a set of constraints or performance metrics

[18] https://github.com/thu-ml/ares
[19] https://github.com/peikexin9/deepxplore
[20] https://github.com/mozilla/bugbug
[21] https://www.tensorflow.org/tfx/guide
[22] https://activeclean.github.io/
[23] https://github.com/sjyk/alphaclean

for the ML pipeline (Haas 2019). Cruz et al. (xxxx) jointly maximized accuracy and fairness using a multi-objective Bayesian optimization technique. This method handled fairness through the model's hyperparameter optimization. Moustakidis et al. (2020) optimized a DL model for knee osteoarthritis classification by combining performance metrics (i.e., accuracy, precision, and recall) with fairness metrics (i.e., demographic parity and balanced equalized odds). By adding fairness constraints to the objective function of DL models, Cherepanova et al. (2021) observed that models can overfit to fairness objectives. Technical approaches for ML fairness are applied before modeling (pre-processing), while modeling (in-processing), or after modeling (post-processing).

4.3.1 Fairness measurement and mitigation metrics

Measuring the fairness of data and algorithms depends mainly on the type of task at hand, i.e., supervised (e.g., classification or regression) and unsupervised (e.g., clustering) learning. The study concentrates on supervised learning tasks. To create a fair model, metrics must be used to: (1) assess the fairness, (2) remove or mitigate the unfairness, and (3) reduce the harm of bias if biases are still present (Bellamy et al. 2019; Mehrabi et al. 2019; Caton and Haas 2020). Fairness is either individual (e.g., everyone is treated equally) or group fairness (e.g., differentiate within or between groups). Narayanan (2018) provided more than 21 definitions of fairness in ML. Still, there is a lack of consistency between fairness definitions and metrics, and each metric measures different aspects of what could be considered "fair" (Caton and Haas 2020). Formally speaking (Kuznetsov 2001), fairness can be statistically defined, but there is no universal measure for fairness.. Because some fairness metrics are incompatible or conflicting, there is no model able to satisfy all of the following definitions of fairness simultaneously (Grgic-Hlaca et al. 2016). Let A be the set of SFs that define the groups for which it is required to measure fairness for a case$U$, X is the set of rest features, Y is the ground truth label (e.g., $Y \in \{0, 1\}$ for binary classification), and $\widehat{Y}$ is the predicted label by the ML model$f$.

4.3.1.1 Similarity-based fairness *(1) Fairness through unawareness:* A model is fair if its predictions are independent of the SFs, i.e., the model discards SFs (Mehrabi et al. 2019). This definition is satisfied if no SFs are explicitly used in the decision-making, so the outcome should be the same for individuals i and j who have the *same attributes*. This metric may fail because some features in X could be correlated with features in A.

$$X : X_i = X_j \rightarrow \widehat{Y}_i = \widehat{Y}_j$$

(2) Fairness through awareness: It assumes that the algorithm is fair if it gives similar predictions to similar individuals, where a distance metric defines similarity.

(3) Causal discrimination: A classifier produces the same classification for two subjects with the exact same attributes X except the SFs A, i.e., for $\{g_i, g_j\} \in A$: $(X_i = X_j) \bigwedge (g_i \neq g_j) \rightarrow \widehat{Y}_i \neq \widehat{Y}_j$.

4.3.1.2 Individual fairness This fairness is satisfied if similar cases have similar outcomes. For a metric d and a valueδ, if $d(U_i, U_j) < \delta$, the predictions forU_i, U_j should be similar.

$$\widehat{Y}(X_i, A_i) \approx \widehat{Y}(X_j, A_j)$$

Note that the level of fairness depends mainly on the definition of d, which defines how similar two given individuals are in the context of a decision-making task.

4.3.1.3 Group fairness This set of metrics measures the model's performance on the population group level, where SFs define groups. Most of these metrics are based on the confusion matrix. Let g_i and g_j denote two protected groups.

(1) Statistical/demographic parity metric (Corbett-Davies et al. 2017) considers the predicted positive rates, i.e., $P(\widehat{Y} = 1)$, across different groups and it is achieved if the outcomes are equal across groups and independently from the protected attribute A:

$$P\left(\widehat{Y} = 1|A = g_i\right) = P\left(\widehat{Y} = 1\middle|A = g_j\right)$$

This definition satisfies the *independence* mathematical property between protected attributes and algorithm (i.e., $\widehat{Y} \perp A$), but the metric concentrates on the outcomes only and ignores error rates and disparities. Positive decisions should have the same rate for all groups; this metric cannot handle the case where an individual is a member of multiple protected groups. By forcing group fairness for one group, the fairness for other groups is violated. This could still be unfair to individuals because the algorithm may eliminate qualified individuals to attain attribute independence. This is known as the problem of inverse discrimination.

(2) Disparate impact metric (Feldman et al. 2015) considers the ratio between unprivileged and privileged groups $\frac{P(\widehat{Y}=1|A=g_1)}{P(\widehat{Y}=1|A=g_2)}$.

(3) Equalized opportunity metric (Hardt et al. 2016) is achieved if equal outcomes happen across subgroups of true positives. It asserts that the True Positive Rate (TPR) should be equal for all groups. It satisfies the mathematical criteria of *separation* (i.e., $\widehat{Y} \perp A|Y$). It ensures that the same fraction of patients in each group will receive correct results but ignores any disparity in the Negative Error Rate (NER).

$$P\left(\widehat{Y} = 1|A = g_i, Y = 1\right) = P\left(\widehat{Y} = 1\middle|A = g_j, Y = 1\right)$$

(4) Equalized odds metric (Berk et al. 2018) is achieved if equality of outcomes happens across both groups and true labels. It asserts that False Positive Rate (FPR) and TPR should be equal across groups. It satisfies the mathematical criteria of *sufficiency* (i.e., $\mathrm{Y} \perp A|\widehat{Y}$). It balances both positive and negative error rates across both groups (i.e., ensures that both groups have equal sensitivity and specificity), but generally results in lower overall classification accuracy.

$$P\left(\widehat{Y} = 1|A = g_i, Y = \gamma\right) = P\left(\widehat{Y} = 1\middle|A = g_j, Y = \gamma\right), \gamma = \{0, 1\}$$

(5) Overall accuracy equality metric (Berk et al. 2018) measures the relative accuracy rates across different groups. If two groups have the same accuracy, they are considered equal.

$$P\left(\widehat{Y} = 0|A = g_i, Y = 0\right) + P\left(\widehat{Y} = 1\middle|A = g_i, Y = 1\right) = P\left(\widehat{Y} = 0|A = g_j, Y = 0\right) + P\left(\widehat{Y} = 1\middle|A = g_j, Y = 1\right)$$

(6) Conditional use accuracy equality metric looks at the positive and negative predictive values for each subgroup.

$$P\left(Y=1|A=g_i,\widehat{Y}=1\right)=P(Y=1|A=g_i,\widehat{Y}=1)\&P\left(Y=0|A=g_i,\widehat{Y}=0\right)$$
$$=P(Y=0|A=g_i,\widehat{Y}=0)$$

(7) Treatment equality metric considers the ratio of false-negative predictions to false-positive predictions.

$$\frac{P\left(\widehat{Y}=1|A=g_i,Y=0\right)}{P\left(\widehat{Y}=0|A=g_i,Y=1\right)}=\frac{P\left(\widehat{Y}=1|A=g_j,Y=0\right)}{P\left(\widehat{Y}=0|A=g_j,Y=1\right)}$$

(8) Equalizing disincentives metric compares the difference of TPR and FPR across groups:

$$P\left(\widehat{Y}=1|A=g_i,Y=1\right)-P\left(\widehat{Y}=1\middle|A=g_i,Y=0\right)=P\left(\widehat{Y}=1|A=g_j,Y=1\right)$$
$$-P\left(\widehat{Y}=1\middle|A=g_j,Y=0\right)$$

(9) Test fairness/calibration/matching conditional frequency metric (Chouldechova 2017): when two individuals from two different groups get the same predicted scores, they should have the same probability of belonging to $Y=1$.

For $A_1\in A, P\left(Y=1|A_1=a,g_i\right)=P\left(Y=1|A_1=a,g_j\right)$

(10) Well calibration (Kleinberg et al. 2016) is an extension of the previous metric where the probability of being in a positive class has to equal a specific value *s*.

For $A_1\in A, P\left(Y=1|A_1=a,g_i\right)=P\left(Y=1|A_1=a,g_j\right)=s$

Note that the study does not discuss other well-known statistical measures such as positive predictive value, false discovery rate, false omission rate, etc. Readers are guided to Verma and Rubin (2018) for a full explanation of these metrics.

4.3.1.4 Causal fairness These metrics consider the outcome for every single individual (Chiappa 2019). Given a causal fairness model, this fairness compares the probabilities of possible outcomes for different interventions. For any $X=x$ context, SF value $A=a$, and possible value a' of A:

$$P\left(\widehat{Y}_{A\leftarrow a}=y|X=x,A=a\right)=P\left(\widehat{Y}_{A\leftarrow a'}=y|X=x,A=a\right)$$

It is always hard for any classifier to maintain the same relations between subgroups, so what is the acceptable amount of bias? A general rule must be defined to determine the margins around perfect fairness. One rule is called the four-fifths rule (P. 1607-UGOES Procedures 1978) "*a selection rate for any race, sex, or ethnic group which is less than four-fifths of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact*." Fairness implementation depends mainly on the level of accessibility to the data and model. If training data exist, fairness checking is done to these data (*preprocessing fairness*). If discrimination is noticed, fairness-aware data preprocessing steps are performed on the data. The model's fairness is considered if there is no training data access. Again, depending on the level of access to the model, fairness can be implemented while training the model (*in-processing fairness*) or on the model's output (*post-processing fairness*).

4.3.2 Fairness of data

This is called preprocessing or model-agnostic fairness which happens in the data preparation stage. Collected data can reflect the unfairness and social discrimination in the real world. Data are biased if the sampling distribution is different from the population's true distribution. For example, the model is expected to be biased if it is trained using data from one hospital, tested by data from another hospital, or trained using data from one geographic region and tested by data from other regions. Using these data to train models can produce unfair models, and using these data to train explainers can produce unfair explainers. Notice that the unfair dataset will not become fair by just removing the SFs because correlations between protected features and other features could still create unfairness which remains after removing SFs. Unfair data can be divided into class imbalanced (i.e., labeling bias and under-representation of minority class) (Krawczyk 2016) and feature imbalanced (i.e., the distribution of protected attributes with different groups such as race, gender, native language, income level, religion, and sexual orientation are biased) (Mehrabi et al. 2019). It is well known in the literature that data are the primary source of discrimination, e.g., sample bias, inappropriate use, data veracity and quality, data context shift, and subjectivity filters (Caton and Haas 2020). For example, sample bias is a data sampling method where not all SF groups are adequately represented. Sample bias has two directions, i.e., over or under-representation. This yields models to consider the minority class as an outlier and optimizes its cost function to predict the majority class only. To solve this problem, we may alter the sample distribution of SFs and minority classes by resampling or collecting data where all SF groups are equally represented. A possible approach is to use the stratified resampling technique. In addition, working with a domain expert to fully understand the possible discrimination ways could help but not be sufficient to prevent bias. On the other hand, involving domain experts in the data collection step and performing the statistical demographic analysis can improve the process. In addition, data incompleteness affects the model fairness because there are different sources of incompleteness, including missing values at random and not at random (Goel et al. 2020). The critical issue of missing data is that the missing values may convey information, and ignoring this dependence could result in biased models. Data preprocessing steps are crucial to mitigate data biases. However, data preprocessing techniques are not aware of the internal operation of models, so they are not tuned to optimize the model performance. As a result, there is a trade-off between fairness and accuracy (Lou et al. 2013). The most popular methods for mitigating the unfairness of data concentrate on binary classification problems (Caton and Haas 2020). There are seven methods in the literature for preparing data with fewer biases, including adversarial learning (Feng et al. 2019), causal methods (Salimi et al. 2019), relabeling and perturbation (Cowgill and Tucker 2017), (re)sampling (Adler et al. 2018), reweighting (Kamiran and Calders 2012), transformation (Calmon et al. 2017), and variable blinding (Chouldechova and G'Sell 2017). Readers are referred to Mehrabi et al. (2019) for more details about these methods.

4.3.3 Fairness of model

Models must be designed such that they take potential data bias into account. Inspired by Hajian et al. (2016), algorithmic fairness has two main solutions: (1) *in-processing solutions* that design models to be inherently fair, and (2) *post-processing solutions* that adapt the model's outcomes to be fair. The in-processing methods either modify the model to

remove unfairness or add an independent agent that detects and prevents unfairness. Some solutions added regularization terms to the objective function for the in-processing methods, which penalizes the correlation between SFs and the predicted labels, and the resulting model maximizes performance and fairness (Kamishima et al. 2012). Other solutions add constraints to the model (Caton and Haas 2020). For example, Goh et al. (2016) used a non-convex constraint to optimize a classifier under a fairness constraint based on disparate impact. Calders et al. (2010) optimized naïve Bayes models for every possible SF value and thus balanced their outcomes. In adversarial learning, two models are trained at the same time. The first one models the data distribution, and the second estimates the probability that a sample comes from the training data rather than the model (Goodfellow, et al. 2014). Post-processing approaches study the outputs of models and then transform them to remove unfairness (Caton and Haas 2020). These methods require access to the SFs while making decisions. Some studies combine in-processing and post-processing methods by training different classifiers for different SF groups (Dwork et al. 2018). However, modifying data and output may have legal and explainability implications. In addition, these methods do not have any access to the optimization functions of models. In contrast, in-processing approaches can add regularization terms to the optimization functions to increase fairness, but this needs the functions to be accessible, replaceable, and modifiable. The most popular in-processing techniques for mitigating algorithmic biases are adversarial learning (Beutel et al. 2019), bandits (Gillen et al. 2018), constraint optimization (Celis et al. 2019; Cotter et al. 2019), regularization (Stefano et al. 2020), multitask modeling (Benton et al. 2017), multitask adversarial learning (Beutel et al. 2017), and reweighting (Krasanakis et al. 2018). For the post-processing bias mitigation methods, calibration (Liu et al. 2017), constraint optimization (Kim et al. 2018), thresholding (Iosifidis et al. 2019), and transformation (Chiappa 2019) are the most popular methods. Readers are guided to Caton and Haas (2020) for more details about these methods.

4.3.4 AI fairness open-source libraries

Table 1 enumerates open-source tools for providing pre-, in-, and post-processing approaches for classification and regression tasks by considering fairness assessment in the model development pipeline. Most tools are implemented in Python. These tools are used to detect and prevent bias either in data or in algorithms. However, their adoption in real-world applications is uncommon mainly because they need to access SFs at inference time or have high development and deployment costs (Cruz et al. xxxx). Other tools that could support a better understanding of algorithmic decisions and possible biases are aimed for supporting model transparency, explainability, and interpretability.

4.4 Explainability methods

The explainability of algorithmic decisions is a pressing TAI issue concerning patients, physicians, engineers, and regulation agencies (Diprose et al. 2020). Indeed, model transparency is considered one of the main barriers to implementing AI in healthcare (Markus et al. 2020). For example, physicians need to build trust and confidence in AI to make sensible decisions that affect human lives. Data scientists and engineers want to know whether the algorithm is stable and captures the relevant features. XAI can detect issues like dataset shift, the model's wrong or incomplete objectives, and reliability limitations. As a result, XAI requirements are starting to appear in legal regulations and

Table 1 Fairness tools

Tool	Reference	Algorithms	Language	Supported fairness		
				Pre-processing	In-processing	Post-processing
IBM AIF360	Bellamy et al. (2019)	Binary classification	Python	Yes	Yes	Yes
Microsoft Fairlean	https://fairlearn.org	Binary and multiclass classifications and regression	Python	Yes	No	Yes
Aequitas	Saleiro et al. (2018)	–	Python	Yes	No	No
Fairness	https://github.com/kozodoi/fairness	Binary classification	R	No	No	Yes
FairTest	Tramer et al. (2017)	Binary classification and regression	Python	No	No	Yes
Dataset Nutrition Label	https://datanutrition.org/	–	JavaScript	Yes	No	No
Silva	Yan et al. (2020)	Binary classification	Python	Yes	Yes	Yes
Dalex	Baniecki et al. (2020)	Binary classification and regression	Python	No	No	Yes
Fairmodels	Wiśniewski and Biecek (2021b)	Binary classification	R	Yes	No	Yes
Google What-If	Wexler et al. (2019)	Binary and multiclass classification and regression	Python	Yes	Yes	Yes

ethical guidelines, e.g., article 22 of the EU general data protection regulation (GDPR (Regulation (EU) 2016, European Union, General data protection regulation 2016)). Who is accountable for wrong automated decisions? Can we explain why things went wrong? If working well, can we say why and how much confidence we have for future decisions? All these questions can be solved by using suitable and robust XAI techniques. Arrieta et al. (2020) stated that "*given an audience, an XAI is one that produces details or reasons to make its functioning clear or easy to understand*." An audience of XAI determines the level of (local or global) required explainability, the time limitation to understand an explanation, and the user experience, which determines the required level of detail. Decision rules/trees and linear models are considered interpretable white-box models because they can be easily interpreted if they are not very large (i.e., if the number of rules, the size of trees, or the number of explanatory features are small enough to be processed by humans). In addition, users can mathematically analyze their algorithmic mechanisms.

DL, ensemble, SVM, and other non-linear models are considered black box opaque models because it is not easy to understand why they yield specific outputs for certain inputs. These models hide their sophisticated internal logic; thus, users do not understand their underline rationales. *After adding XAI features, these black-box models can be explained to humans.* Some studies advised against explaining black models, such as Rudin (2019). Thus, explainability is different from interpretability, but the study refers to both by using the XAI abbreviation. As a result, XAI, trustworthiness, and tractability are at the heart of the successful deployment of AI-based applications (Toreini, et al. 2020).

Implementing a trustworthy XAI is a challenge (Diprose et al. 2020). As mentioned by Molnar (2018), the explanation of ML models should take into account: accuracy, reliability, fidelity, consistency, stability, scalability, causality, representativeness, certainty, novelty, degree of importance, and comprehensibility. Markus et al. (2020) provided definitions for these terminologies. However, many questions need to be answered to achieve these goals, including: What does it mean that a model is explainable? How can explainability be quantified? Employing what metrics? When is the explanation considered comprehensible? How to define the audience of explanation? How much is the audience willing to lose in prediction accuracy to gain how much explainability? What is the best XAI format to the studied domain and/or audience? What time constraints are required for users to understand explanations? How to match the complexity of explanations with user experience and background knowledge? These are complex questions to answer. For example, Kononenko et al. (Štrumbelj and Kononenko 2010) proposed a qualitative method to measure explainability, but there is no quantitative evaluation framework to assess the XAI features of a model yet (Miller 2019).

Moreover, different data formats have different degrees of explainability. Even tabular data are the most popular format. Images, text, and time series are somehow understandable formats to humans. Data preprocessing steps, especially FS, significantly affect the model explainability (Gonzalez Zelaya 2019). Crone et al. (2006) studied the effect of the preprocessing steps like scaling, sampling, and encoding on the explainability. This is called interpretable data for interpretable models (Guidotti et al. 2018). However, the explainability of data collection and preprocessing steps is not fully explored compared to the model explainability. XAI techniques have been implemented to provide explainability in different formats to different black-box models regarding global and local explainability (Arrieta et al. 2020). Global XAI provides explainability for the whole model on the whole dataset. Local XAI regards a specific single instance. XAI techniques that are independent of the model to explain are called model agnostic explainers, but XAI techniques designed for specific models are called model-specific explainers. Readers are referred to (Arrieta

et al. 2020; Guidotti et al. 2018) for comprehensive surveys of these techniques and their application in different fields. Figure 3 illustrates our taxonomy of XAI techniques, which are revisited in the following subsections.

4.4.1 Interpretable ML models

Interpretable ML models include linear models (e.g., linear regression, logistic regression, general additive model, or (semantic) fuzzy model), discretization models (e.g., rule-based model, DT, or Bayesian network), example-based models (e.g., k-nearest neighbor (KNN) or case-based reasoning (CBR)).

A DT is a hierarchical tree structure composed of internal nodes for feature evaluations and leaf nodes for a class label or regression value (Babapour Mofrad, et al. 2019). Extracted rules from DTs have the conjunctive form (i.e., $IF condition_1 AND condition_2 AND \ldots AND condition_n THEN outcome$). Trees have graphical representation and rules have textual representation. DT has inadequate generalization capabilities that can be enhanced using tree ensembles like Random Forest (RF) and XGBoost (Arrieta et al. 2020), but the DT combination loses transparency features. Some techniques are available to extract rules from RFs (Wang et al. 2020a). The general form of rule-based models can be formed by conjunctions, disjunctions, and negations (Kuo and Fuh 2011) and can have the form of m-of-n rules, which means that if m of the n conditions in the antecedent part of a rule are satisfied, then the consequence of the rule is true (Guidotti et al. 2018). Weights and/or orders can be added to the rule list to force specific priority in rules execution. Feature importance can be extracted from DT based on the order of using features in the tree. Interpretability of rules can be enhanced by extending rules with semantic knowledge (Skillen et al. 2014), dealing with uncertainty with fuzzy logic (Gacto et al. 2011), or both (El-Sappagh et al. 2018). Accordingly, explainable fuzzy systems are a powerful tool in the context of XAI (Alonso et al. 2021). Note that deep DTs and systems with a long list of decision rules are considered less interpretable models (Markus et al. 2020; Blanco-Justicia et al. 2020).

Linear models have easy-to-interpret mathematical models (i.e., $y = w_1x_1 + w_2x_2 + \ldots w_nx_n + w_0$), being w_i the degree of how much x_i contributes to the prediction of y. Logistic regression adds nonlinearity to return the probability of a class. The feature weights and signs can be interpreted as feature importance for the global model. It is possible to make decisions by plugging feature values and weights into the model's formula. Nevertheless, linear models with a large number of non-zero weights are considered as not interpretable models (Guidotti et al. 2018). Instance-based methods like CBR and KNN are popular methods in XAI, especially for the medical domain, where similar cases represent an experience for physicians (Nasiri et al. 2019). The capabilities of CBR can be enhanced by integrating it with fuzzy ontology and accurate semantic similarity measures (El-Sappagh et al. 2015). Like CBR, KNN produces transparent models which predict the class for a test sample by voting the classes of the k nearest neighbors. In regression tasks, KNN takes the average of the k neighbors. KNN is based on a selected similarity function. However, KNN becomes less transparent if the number of features is large, the number of neighbors is large, and the similarity function is complex for physicians to understand (Arrieta et al. 2020). The general additive model (GAM) is a linear model where the dependent feature value is given by aggregating several unknown smooth functions defined for the predictor variables, and GAM learns these smooth functions. It has the general form of $g(E[y]) = \beta_0 + \sum f_j(x_j)$, where g is the link function. GAM facilitates

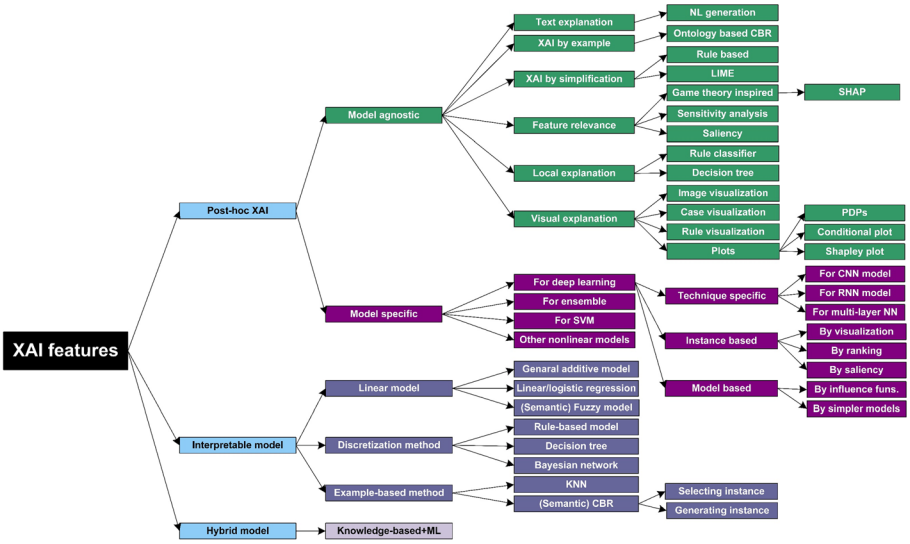

Fig. 3 XAI features

understanding of the relationships of the variables in the data (Caruana et al. 2015). Bayesian network has been used in many medical applications (McLachlan et al. 2020). It is a probabilistic directed acyclic graph model where links are conditional dependencies between a set of variables, e.g., the relationship between diseases and symptoms.

4.4.2 Post-hoc XAI techniques

Applying ML in sensitive domains such as healthcare needs accurate as well as explainable models (Markus et al. 2020). Black-box models (e.g., DL models, ensembles, and SVM) are usually more accurate than white-box interpretable models, but at the cost of a lack of interpretability. The decision function of a neural network is formulated as $y = h\Big(\sigma\Big(W_1\sigma\Big(W_2\sigma\Big(\ldots W_N\left[x_1,\ldots,x_n\right]^T + b_N \ldots\Big) + b_2\Big) + b_1\Big)\Big)$, with multiple implicit nonlinearity σ s, and so the role of x_i feature is not clear on model decisions. Black-box models can be accompanied by post-hoc XAI features as (1) *model agnostic methods* that provide local and global explanations for black-box models through other surrogated interpretable models, and (2) *model-specific methods* that provide tailored XAI for specific models (Arrieta et al. 2020). Even not knowing how the AI system works internally, post-hoc XAI techniques mimic and appwroximate the behavior of black-box models. Different post-hoc techniques have different scopes; some provide global explanations, and others provide local explanations. As shown in Fig. 3, these techniques include visual, textual, example-based, simplification-based, and feature-relevance explanations.

4.4.2.1 Model agnostic techniques This section provides examples of model agnostic techniques.

(A) *Feature relevance XAI*: It explains the function of an opaque model by ranking the importance of features that participate in predicting the outputs. These techniques

include game theory-based, saliency-based, and sensitivity analysis methods. For example, SHAP (SHapley Additive exPlanations) is a game theory-based method to measure each prediction's additive feature importance score (Sappagh et al. 2021).

(B) *Visualization-based XAI*: A survey of these techniques can be found in Cortez and Embrechts (2013). Lamy et al. (Lamy et al. 2018) proposed a case visualization technique to explain a CDSS for breast cancer. Their visual interface displays quantitative and qualitative similarities between the query and similar cases. Other visualization techniques, such as partial dependency plots, are used to explain model decisions (Markus et al. 2020).

(C) *Simplification-based XAI*: A new simplified, less complex model is constructed based on the given opaque model. Local interpretable model-agnostic explanations (LIME) and Anchor are the most well-known methods in this category (Markus et al. 2020). LIME explains the opaque model by building local linear models around its predictions. Rule extraction is another technique in this category. Su et al. (2016) proposed an approach to extract rules in conjunctive normal or disjunctive normal forms to map from a complex model to a human-interpretable model.

(D) *Example-based XAI*

(E) : It extracts data samples related to the decision made by the opaque model. These samples show the inner relationships learned by the model being analyzed. CBR is an example of this category (El-Sappagh et al. 2015).

4.4.2.2 Model-specific techniques These XAI techniques are for conventional ML methods (e.g., ensembles, SVMs, etc.) and DL methods (CNN, RNN, etc.) (Arrieta et al. 2020), as discussed below.

Post-hoc techniques for conventional ML Techniques like tree ensembles and SVMs rely on sophisticated learning algorithms, requiring an additional layer of explanation. For example, Deng et al. (2019) proposed a simplified tree ensemble learner. Feature relevance techniques are used to explain the tree ensemble. Auret and Aldrich (2012) analyzed the role of feature importance in understanding the underlying relationships of complex RF models. Other ensembles like bagging, boosting, and stacking got the attention for extending their explainability features. Rajani et al. (2018) proposed stacking with auxiliary features (SWAF) to provide XAI capabilities for a stacking ensemble based on feature relevance techniques. Some post-hoc XAI techniques have been proposed for SVM models (Arrieta et al. 2020). Some studies created rule-based models from SVM support vectors (Barakat and Bradley 2007). Others provide users with SVM visualization tools (Üstün et al. 2007). Üstün et al. (2007) proposed a visualization method to extract information from the kernel matrix.

Post-hoc techniques for neural networks DL techniques-based taxonomy has been proposed in Arrieta et al. (2020), and instance-based taxonomy has been proposed in Huang et al. (2020). In this subsection, the study highlights the popular techniques for multilayer perceptron (MLP), convolutional neural network (CNN), and recurrent neural network (RNN).

(A) *MLP*: A few studies address the challenge of explaining MLP. Feature relevance techniques have been explored to interpret MLP models. For example, Zilke et al. (2016) proposed *DeepRED* for simplifying the model by extracting a list of representative rules from the trained model. Shrikumar et al. (2016) proposed *DeepLIFT* to compute

model importance levels. They compared the activation of a neuron to the reference activation and put the importance level based on the difference. Some authors (Kindermans, et al. 2017) studied the theoretical soundness of these XAI techniques and discovered that explanations were not theoretically correct. Accordingly, they proposed two more theoretically grounded methods: *PatternNet* and *PatternAttribution*.

(B) *CNN*: state-of-the-art models for computer vision-based tasks, such as image classification and object detection. XAI techniques for these models are based on visual data. Explainability for images highlights the pixels or regions in the image that are linked to the prediction. Class activation map (CAM) is a CNN XAI technique where a convolution layer replaces the top fully connected layers to find the spatial distribution of critical regions for the predicted category (Zhang et al. 2021). Other techniques extend the functionality of CAM, such as Gradient-based Localization-CAM, adversarial complementary learning for weakly supervised object localization (Zhang et al. 2018), and guided attention inference networks (Li et al. 2019a). Zeiler et al. (2011) map back the output in the input space to discover which part of the image was essential for generating the output. They used *Deconvnet* neural network to reconstruct the maximum activations by using the feature map of a selected layer. This reconstruction gives an idea about which parts are more important in the image. The authors visualized the most robust activations using saliency maps. Bach et al. (2015) visualized the predictions of every input image pixel in the form of a heatmap using the layer-wise relevance propagation (LRP) algorithm. Xu et al. (2015) utilized textual explanations related to visual contents in an image. They combined the CNN feature extractor with the RNN attention model to learn textual descriptions of images. An alternative is to mix multiple methods like visualization and feature relevance to provide a richer explanation of the network (Olah et al. 2018).

(C) *RNN*: XAI techniques are used to interpret the sequential data (e.g., natural language processing and time series analysis) used in RNN networks. Most studies used feature-relevance techniques to understand what an RNN has learned. For example, Arras et al. (2017) extended the LRP to work with LSTM by proposing a specific propagation rule which works with multiplicative connections. Kwon et al. (2019) proposed RetainVis, a visual analytics tool for interpretable and interactive RNNs. This tool provides an interpretation of time series data collected from EHR systems.

4.4.3 Hybrid models

Integrating background knowledge as logical statements or constraints with data-driven models improves the robustness and explainability of models (Markus et al. 2020; Arrieta et al. 2020). In addition, data fusion and formulation of conventional ML techniques can extend the XAI features. For example, Papernot et al. (2018) proposed the deep k-nearest neighbors, where the classical ML model is extended with its enhanced deep counterpart. The neighbors are used to create human interpretable explanations. Moreover, building twin models by combining black-box and white-box models (e.g., DL with CBR) improves explainability while maintaining accuracy (Keane and Kenny 2019).

Many challenges must be handled to build trusted XAI features. Deep explainability should integrate (1) fuzzy reasoning features to handle the vagueness of the medical domain; (2) semantic reasoning features to support the understanding of semantic concepts such as comorbidities and medications; (3) knowledge-based reasoning to increase

physician confidence in the explanations of data-driven models; and (4) explainers which use understandable and deep features of the input data based on user experience (Arrieta et al. 2020; Guidotti et al. 2018; Ribeiro et al. 2016). Medical image explainability is challenging because images come in different formats with high dimensionalities, such as 3D. In addition, the ground truth images may not be correct. These images require specialized knowledge to be understood. As a result, it is not guaranteed that the ML model can capture the proper features. Furthermore, semantic knowledge and data privacy are critical issues when building XAI models. For example, Blanco-Justicia et al. (2020) used shallow DTs trained on disjoint subsets of the training datasets as surrogate models. They used a micro aggregation clustering technique to create subsets of training data to obtain representative trees and ensured the privacy of subjects considered for training. The study integrated ontologies to manage categorical features in a semantically consistent way. From the social science perspective, Miller (2019) observed that people prefer contrastive XAI, i.e., why the algorithm did not take a different decision, and XAI is social, i.e., it should be part of a broader conversation between explainer and explainee. XAI models should be able to provide on-demand and customizable explainability based on heterogeneous formats. For example, in the medical domain, XAI features should include diverse data types like neuroimages, speech, sequence, text, and tabular data. Joshi et al. (2021) surveyed the explainability of multimodal data in DL literature. Analyzing time-series data is crucial for chronic disease management, but interpreting ML models that learn time-series data is not easy. There is a shortage in the literature for time-series XAI tools and techniques. Saluja et al. (2021) used regular LIME and SHAP to explain ML models that learn time-series data. Recently, Rojat et al. (2021) provided a comprehensive survey of XAI techniques and their evaluation metrics for time series data.

4.4.4 XAI tools

There exist many XAI tools, including (1) Python tools like *Interpret*,[24] *ethic*,[25] *explain*,[26] *IML*,[27] *explainable_ai_sdk*,[28] *DeepExplain*,[29] *DrWhy*,[30] *alibi*,[31] *ELI5*,[32] *Skater*,[33] *FAT Forensics*,[34] *ExplainExplore*,[35] and *AIX360*[36]; (2) R tools like *ALEPlot*,[37] *forestmodel*,[38] *iBreakDown*,[39] *shapper*,[40] *fastshap*,[41] and *EIX*.[42] Maksymiuk et al. (2020) surveyed and

[24] https://github.com/interpretml/interpret
[25] https://github.com/XAI-ANITI/ethik
[26] https://github.com/explainX/explainx
[27] https://github.com/christophM/iml
[28] https://github.com/GoogleCloudPlatform/explainable_ai_sdk
[29] https://github.com/marcoancona/DeepExplain
[30] https://github.com/ModelOriented/DrWhy
[31] https://docs.seldon.io/projects/alibi/en/stable/index.html
[32] https://github.com/eli5-org/eli5
[33] https://github.com/oracle/Skater
[34] https://fat-forensics.org/
[35] https://explaining.ml/
[36] https://www.ibm.com/watson/explainable-ai
[37] http://xai-tools.drwhy.ai/ALEplot.html
[38] http://xai-tools.drwhy.ai/forestmodel.html
[39] http://xai-tools.drwhy.ai/iBreakDown.html
[40] http://xai-tools.drwhy.ai/shapper.html
[41] http://xai-tools.drwhy.ai/fastshap.html
[42] http://xai-tools.drwhy.ai/EIX.html

Table 2 Summary of AD surveys in ML and DL

t

Ref	Year	Period	Papers	Topic	Used data	Suggested challenges	Summary
Rathore et al. (2017)	2017	1985–2016	81	AD and MCI classification (AD/CN, MCI/CN, sMCI/pMCI, AD/MCI/pMCI /sMCI/CN) using ML and DL	sMRI, fMRI, DTI, FDG-PET, amyloid-PET	- Limited generalization ability exists in all AD studies, - Small sample datasets are used to build and test models, - Reproducibility of results using real-world data is required, - Multimodality data fusion needs to be explored	The study surveyed the feature extraction techniques for every image type. Different fusion strategies have been explored. The study asserted that the neuroimaging-based classification frameworks show promise for personalized diagnosis and progression detection
Zhang et al. (2017)	2017	2010–2015	38	AD diagnosis and progression detection (AD/MCI and MCI to AD) using regular ML models as LR and SVM	Big multimodal data	- Small sample datasets problems. Few studies used large sets such as the ZARADEMP cohort or EHR data. Large and multi-site data repositories are needed - Multimodal data fusion of time series data from EHRs is not explored, - Testing models using external resources is critical (i.e., external validation) - The nature of research datasets is different from that of real-world data (i.e., EHRs) in its completeness and patients' representation - Models developed in research data may miss confounding factors. It is unclear to what extent the results collected under an "ideal" situation will be true for real-life patient data - Reproducibility of model results in a real environment is challenging	This systematic review analyzed the role of big data in the AD domain, such as diagnosing AD or MCI, MCI progression to AD, AD risks, clinical care for AD patients, and the relationship between cognition and AD. Many data sources have been investigated, such as ADNI, EHR, and MEDLINE

Table 2 (continued)

Ref	Year	Period	Papers	Topic	Used data	Suggested challenges	Summary
Pellegrini et al. (2018)	2018	2006–2016	111	AD diagnosis and progression detection (CN/MCI/AD) using ML models as SVMs	sMRI, CT, PET	- No sufficiently large neuroimaging data sets exist - All papers concentrated on the ML points of view (e.g., algorithm, parameter setting, training protocol) and neglected important clinically relevant information (e.g., patient demographics, clinical covariates, data acquisition protocols)	The study compared the accuracy of different ML techniques to differentiate CN from MCI from AD and predict the risk of AD progression. Authors investigated the performance metrics of individual ML models by task, disease of interest, imaging sequence, and feature list. The study investigated the role of ground truth on model results. The study also investigated the role of external testing on model generalizability
Forouzannezhad et al. (2018)	2019	–	–	Progression detection (CN to MCI, MCI to AD) using ML and DL	fMRI	- Multimodal neuroimaging fusion - Exploring the relationship of functional network abnormalities and pathological and biological changes of AD - Lack of longitudinal data to investigate AD progression - A need for advanced techniques in fMRI acquisition and analysis - Using fMRI for differing AD from other neurodegenerative diseases - Using fMRI to investigate the correlation of the neurovascular BOLD signal with AD, age, and medication	The study surveyed and compared different techniques (e.g., ICA and graph-based methods) and application methods for data preparation and functional connectivity investigations of fMRI data. The disruption of brain connectivity patterns of MCI and AD can be used to measure cognitive decline and help to predict AD

Table 2 (continued)

Ref	Year	Period	Papers	Topic	Used data	Suggested challenges	Summary
Ahmed et al. (2019a)	2019	2016–2018	40	AD diagnosis using DL models	sMRI, fMRI, CT, PET, SPECT	- Labeling of datasets for supervised ML models is challenging - The radiotracers used in image collection may have some problems - Imaging data have noises that reduce the model accuracies - AD diagnosis should depend on heterogeneous data from multiple sources - The ground truth of neuroimaging data is not perfect - For XAI, DL architectures are needed to visualize and quantify blood flow, and image interpretation - Multimodal longitudinal molecular imaging such as PET/MRI, SPECT/MRI, PET/CT, or SPECT/CT is expected to improve model performance, but these data are not available	The study comprehensively surveyed recent AD diagnostic methods using neuroimages and ML and DL models. The study demonstrated the role of dataset sizes, FS and optimization, and multimodal neuroimaging fusion on model performance
Martí-Juan et al. (2020)	2020	2007–2019	104	AD progression detection (AD/CN, MCI/CN, sMCI/pMCI, AD/pMCI/sMCI/CN) using ML, DL, and statistical learning	Longitudinal neuroimages	- Most papers did not mention missing data imputation - Temporal alignment problem because different patients can be at different stages of AD at the same acquisition time - Reproducibility and interpretability challenges	The study investigated the role of neuroimaging data and their multimodal fusion in predicting AD progression using different techniques, including multitasking learning, SVM, LR, RF, etc. The study survived the longitudinal studies with a different number of time series steps. Authors highlighted the used cross-validation techniques and the sizes of training data

Table 2 (continued)

Ref	Year	Period	Papers	Topic	Used data	Suggested challenges	Summary
Binth et al. (2020)	2020	2015–2019	21	AD diagnosis based on DL such as LSTM, SNN, DNN, SAE, and RNN	MRI	- Unsupervised deep learning architectures can be used to label neuroimaging data and solve the problem of biased labels - Designing bias-free neuroimaging datasets is challenging - Adversarial noises added with neuroimages reduce the model performance, so the cancellation of adversarial errors is a challenge - DNN models need large datasets, so GAN could be used to generate synthetic neuroimages to be used by DL models such as CNN - No standardized method for image acquisition causes a difference in images pertaining to different datasets. Transfer learning should be used to solve this problem - Explainability of DL black-box models needs further research	Authors compared the performance of DL studies to detect AD using fMRI and sMRI. The results suggested CNN as the best method for AD detection. The study surveyed the feature extraction and data preprocessing techniques for DL models as filtering, normalization, correction, trimming, etc

Table 2 (continued)

Ref	Year	Period	Papers	Topic	Used data	Suggested challenges	Summary
Graham et al. (2019)	2020	–	21	AD diagnosis and progression detection using DL, ML, and ensemble classifiers	Socio-demographics, clinical, psychometric, neuroimaging, neurophysiology, EHR, sensors, genomics	- Personalized decisions require the fusion of comprehensive and longitudinal multivariate data from heterogeneous sources. These data are not available, and the ML model should be complex enough to learn these data - To improve the ML model, FS optimization is a critical step, but continually fine-tuning the algorithm raises the likelihood of overfitting because the model is too customized for a particular training data - ML models are suffering from biases because the used datasets are small and labeled - Large and unlabeled datasets require a shift toward semi-supervised or unsupervised learning techniques, which are harder to train - Ethical and social implications of using AI for AD management need to weigh benefits against potential risks (e.g., bias and accountability) to patients - AI models generalizability must be assured before deployment - Model transparency and explainability are required to improve user trust, - Because the ML model could be deployed on mobile devices, privacy issues should be considered - AI model must follow evidence-based practices to diagnose AD, so the AI model must meet clinical standards. However, the threshold of proof in AI models is not yet established	The study surveyed the ML models that integrated biological, psychological, and social factors for the diagnosis and prognosis detection of AD. The studied techniques include supervised and unsupervised ML, deep learning, and NLP. The study highlighted the effect of CV and dataset size on the model performance. All studies asserted that AI would support the decision-making capabilities of the physician, but it will not replace their expertise

Table 2 (continued)

Ref	Year	Period	Papers	Topic	Used data	Suggested challenges	Summary
Ebrahimighahnavieh et al. (2020)	2020	2013–2018	100	AD diagnosis (AD/MCI and MCI/NC) using DL as CNN, RNN, AE, RBM, and DNN	Demographic, genetic, and neuroimages	- Low quality of neuroimaging acquisition and errors in data preprocessing and brain segmentation - Comprehensive dataset unavailability, i.e., a dataset including many subjects and biomarkers - Low variance between different classes of AD stages. Normally AD signs are very similar to the normal healthy brain of older people. The boundaries between AD and MCI, and MCI and NC are vague - No human in the loop, especially for neuroimaging processing such as identifying regions of interest - Neuroimages like MRI and PET are more complex compared to other images - Clear explanations of DL models are required - Incomplete datasets in multimodal studies must be solved - Best data fusion that combines the most discriminative features is required	The study reviewed AD detection literature using DL techniques. Authors studied AD biomarkers (e.g., demographic, genetic, and brain image), the data preprocessing steps, and neuroimaging processing methods. These data were studied from single and multi-modalities sources and from cross-sectional and longitudinal data
Tanveer et al. (2020)	2020	2005–2019	165	AD diagnosis with ensembles, DL, ANN, and SVM	MRI, PET, DTI	- Novel learning techniques should be developed for small datasets - Multimodal clinical data can be utilized in AD management - AD data are always noisy, so noise insensitive techniques must be applied for AD management - Many ML models, such as SVM and ANN, have not been fully investigated in the AD domain. Ensemble of SVMs and SVMs with new kernels need to study - Transfer learning, feature selection, and multi-kernel learning methods need to be explored on multimodal data	They surveyed different ML and DL techniques for AD diagnosis using different methodologies as multitask learning, transfer learning, multi-kernel learning, many FS techniques. Authors concentrated on the used CV methods, especially the tenfold CV

compared the existing R packages for XAI. However, these techniques are not complete enough to provide XAI features to deployed models in sensitive domains like medicine. Model explainability is highly correlated with other TAI requirements. For example, biased models always have biased explanations (Jain et al. 2020). Measuring the robustness and reliability of generated explanations is a critical future direction. One example of measuring the robustness of XAI methods is by using adversarial-based interpretation, where a small perturbation in data could significantly impact the result (Liang et al. 2021). In addition, there is a shortage in the formal evaluation metrics for XAI features. Quality evaluation metrics should support the comparison between different XAI techniques, but there is no underground truth for the comparison. No standard evaluation methods could determine when an AI system is explainable for specific users (Markus et al. 2020). Combining different XAI techniques regarding different data formats (e.g., neuroimages, text, time series, and structured data) can increase explainers' complexity, conflicts, and uncertainties. There are no robust and globally accepted quantitative measures for fidelity, stability, certainty, completeness, context, portability, interactiveness, personalization, actionability, accuracy, parsimony, usability, and user understandability of model explanations. In addition, combining different XAI techniques in a user interface is a required research direction. For a comprehensive survey of qualitative evaluation measures, readers are referred to Mohseni et al. (2018).

5 Trustworthy AI in the AD domain

Many surveys have studied different applications of AI in the AD domain. Due to space restrictions, the study has collected surveys for AD diagnosis and progression detection since 2017 (see Table 2). As it can be noticed, all studies concentrated mainly on neuroimaging data and neglected other types of cost-effective data like symptoms, lab tests, comorbidities, and medications. All studies focused on the ML and DL issues like performance, model architectures, hyperparameters optimization, and FS. These are critical issues to consider, but they are not the primary source of user trust. Only some studies have considered external datasets. In addition, cross-validation and data-splitting methods were not always carried out properly. Data preprocessing techniques have shortages like not handling missing values in a medically sensible way. In addition, outlier detection and handling of imbalanced data are mostly neglected. In summary, current studies do not deal with trustworthy issues in detail. However, most studies mention trustworthy issues as future directions, such as XAI, reproducibility, external testing, handling missing values, generalizability, small datasets, EHR integration, etc. This inspires us to work deeply on this topic and evaluate the literature studies regarding TAI measures.

There is no survey in the literature which measures the level of trustworthiness of the implemented model and the level of applicability of this model in a real environment. It is worth noting that Wen et al. (2020) analyzed 30 studies before starting their implementation to highlight data leakage problems in DL models. They defined five sources of data leakage and concluded that more than half of the studies suffered from at least one of these problems. As noted in the literature, although there are so many studies for AD diagnosis and progression detection using ML and DL techniques, there is no single CDSS to assist physicians or patients in disease management. This indicates that there is a problem with the level of maturity of these studies, and as a result, they failed to gain the needed level of

community trust. Knowing the reasons for these limitations is critical to improving state of the art by highlighting the real challenges that research should concentrate on in the future. Our study focuses on this point and provides readers with a comprehensive study of AD literature by concentrating on the TAI metrics and guidelines. The study concentrates only on supervised problems, mainly classification tasks. In the following sections, we survey the AD literature that does not consider trustworthy, and then we pay attention to models that implemented some TAI requirements such as XAI, robustness, reproducibility, or fairness. Note that all selected models were built with the model validation and optimization steps that are required by TAI.

5.1 Studies that do not explicitly consider trustworthy AI

This section surveys AD studies that concentrate on model performance and reproducibility, which are sub-requirements of robustness, as discussed in Sect. 4.2. Regarding model performance, the study compares the reported results with respect to data preparation steps, validation techniques, statistical hypothesis testing techniques, and hyperparameter optimization techniques. In addition, the study checks if it was done either bias analysis of the dataset or ground truth was verified, or results were presented with confidence intervals. We evaluate dataset and cohort sizes, the feature engineering steps, whether the study evaluated its model using multiple datasets, whether the dataset and code were publicly available, and whether the model's sensitivity to slight changes in input data has been evaluated. In the reproducibility column, bold text indicates the handled metrics. Results are formatted as (–/–/–/–/–) for (Accuracy/Precision/Recall/F-score/AUC).

5.1.1 Classical machine learning models

This subsection considers only the papers related to classical ML algorithms such as SVMs, DTs, or logistic regression (LR). All studies in this section did not use time series or multimodal data. Table 3 compares 18 chronically sorted studies from 2017 and 2021. Column titles and acronyms are described in the caption of Table 3. From this table, we figure out why AI does not have any real application in the AD domain. The most important issues that could be considered as efforts in TAI are cross-validation and reproducibility. Most studies used ADNI datasets (ranging from 40 to 41,350 examples) (Lu et al. 2017a; Divya et al 2021; Poloni et al. 2021; Eke et al. 2021; Park et al. 2020; Richhariya et al. 2020; Zhu et al. 2019). In addition, most studies relied on an FS step to select a small number of features. However, they did not consult domain experts to evaluate the medical relevance of the collected features. In addition, ADNI collected both longitudinal and cross-sectional data, but only (Sørensen et al. 2017) mentioned the used category. Note that longitudinal data measures the gradual progression of the disease and needs special manipulation to produce medically relevant results. Most studies provided not sufficient dataset description, thus making their results irreproducible. Moreover, most studies did not provide access to their used datasets or detailed descriptions of the used feature sets. Most studies did not consider the physician's opinion (Park, et al. 2020; Poloni et al. 2021; Eke et al. 2021; Zhu et al. 2019; Li et al. 2019b; Gómez-Sancho et al. 2018). As can be noticed in Table 3, most studies depended only on MRI data which usually achieved not satisfactory results (Poloni et al. 2021; Wang et al. 2020b; Richhariya et al. 2020; Vaithinathan and Parthiban 2019). MRI is handled by a standard pipeline which includes the following steps

Table 3 Review of studies based on classical ML models

Ref. (year)	Data set	Subjects	Domain expert	Modality	Task (purpose)	Data preparation	Feature eng
Divya and Shantha Selva Kumari (2021)	ADNI	AD (171), MCI (558), CN (347)	NO	sMRI	AD detection (classification)	Volumetric segmentation, missing values, scaling	GA and RFE feature selection
Poloni et al. (2021)	ADNI	AD (209), MCI (251), CN (302)	YES	MRI	AD detection (classification)	Noise reduction using NLMT, bias field correction by N4-ITK, intensity standardization, registration by NAC T1-w, Hemispheres' extraction	Asymmetry index-based FS, directional filtering, ANOVA features
Eke et al. (2021)	ADNI	CN (112), MCI (136), AD (108)	YES	Blood plasma proteins	AD detection (classification)	Standardization	CFS feature selection
Park et al. (2020)	KNHISD	AD (614), MCI (2026), CN (38,710)	YES	Administrative health data	AD progression (classification)	Data alignment, coding, data filtering, missing values	FS by variance threshold
Wang et al. (2020b)	ADNI, PUTHC	***PUTHC***: AD (131), CN (131), ***ADNI***: AD (67), CN (105)	NO	MRI	AD detection (classification)	MRI segmentation, ROIs masking by WFU Pick- Atlas	Feature extraction and SPCA
Richhariya et al. (2020)	ADNI	AD (50), MCI (50), CN (50)	NO	sMRI	AD detection (classification)	Image registration, segmentation (GM, WM, CSF), normalization, smoothing by GKWFW	FS by USVM-RFE + PCA
Carvalho et al. (2020)	CAD, CRASI	CAD (319), CRASI (124)	YES	NSB + CSs	AD and MCI detection (classification)	Missing values + balancing by SMOTE	Discretization
Durongbhan et al. (2019)	STHNT	CN (20), AD (20)	NO	EEG	AD detection (classification)	3 segments data augmentation	Feature extraction
Vaithinathan and Parthiban (2019)	ADNI	AD (189), NC (227), pMCI (165), sMCI (231)	NO	MRI	AD progression (classification)	Realignment, registration, normalization, and segmentation to GM and WM	FS with fisher, elastic net, SVM-RFE + bootstrap

Table 3 (continued)

Ref. (year)	Data set	Subjects	Domain expert	Modality	Task (purpose)	Data preparation	Feature eng
Baskar et al. (2019)	ADNI, Bordeaux-3	AD (137), MCI (210), CN (162)	NO	sMRI	AD detection (classification)	Feature extraction for texture and shape features from HC and PCC using AAL	Multiple-criterion feature selection
Zhu et al. (2019)	ADNI	AD (186), MCI (393), CN (226)	YES	MRI	AD progression (classification)	Image correction, skull-stripping extraction, segmentation	FS by TSSRFS
Li et al. (2019b)	ADNI, Tongji	**ADNI**: CN (175), AD (117), **Tongji**: CN (14), AD (12)	YES	fMRI	AD detection (classification)	Head motions correct, normalize, smooth, brain network modeling by AAL atlas	Singular value decomposition
Bucholc et al. (2019)	ADNI	CN (178), MCI (263), AD (47)	YES	CFA	AD detection (classification + regression)	Normalization	Recursive feature elimination-
Gómez-Sancho et al. (2018)	ADNI	AD (200), NC(231), pMCI(164), sMCI(100)	YES	MRI	AD progression (classification)	Feature extraction for HC and entorhinal cortex volumes and region features + VBM	PCA voxel
Zeng et al. (2018)	ADNI	AD (92), sMCI (82), pMCI (95), CN (92)	NO	MRI	AD progression (classification)	MRI skull stripping, registration with 10 mm level B-spline FFD, segmentation and normalization, and smoothing	Feature extraction + PCA
Tong et al. (2017)	ADNI	CN (229), sMCI (129), pMCI (171), uMCI (98), AD (191)	NO	MRI	AD progression (classification)	MRI registration, removing normal aging effect, data normalization	Voxel-based feature selection

Table 3 (continued)

Ref. (year)	Data set	Subjects	Domain expert	Modality	Task (purpose)	Data preparation	Feature eng
Sørensen et al. (2017)	ADNI, AIBL, CAD-Dementia	***ADNI/ AIBL***: CN (182/88) MCI(245/29) AD(117/28), ***CAD-Dementia***: 354	NO	sMRI	AD detection (classification)	z-score normalization, HC segmentation	Manual FS of HC, cortical thickness, and volumetry
Lu et al. (2017a)	ADNI	CN (152), MCI (120)	NO	FDG-PET	MCI detection (classification)	Extract special features of voxel volume (sub)cortical regions by normalization + AAL registration	Data transformation using incomplete random forest

Ref. (year)	Algorithm	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Divya and Shantha Selva Kumari (2021)	RBF-based SVM	Repeated stratified 10-CV (-)	Grid search	*OH*, data and code not available, no CI, no SEN, no validation, no SST, no BA, no GTV	CN vs. AD CN vs. MCI MCI vs. AD	96.8/-/92.8/-/- 89.4/-/95.2/-/- 90.4/-/69.6/-/-	-/-/-/-/- -/-/-/-/- -/-/-/-/-
Poloni et al. (2021)	Polynomial kernel SVM	Nested 10-CV (–)	Grid search	*OH*, data and code not available, CI, no SEN, no validation, SST, no BA, no GTV	CN vs. AD CN vs. MCI	82.6/-/-/78/90 69.4/-/-/66/76	-/-/-/-/- -/-/-/-/-
Eke et al. (2021)	k kernel SVM	10 times repeated 10-CV (I)	–	Data not available, *code available*, no SST, *I validation*, no CI, no SEN, no OH, no BA, no GTV	CN vs. AD CN vs. MCI	-/-/85/-/89 -/-/80/-/80	-/-/-/-/- -/-/-/-/-

Table 3 (continued)

Ref. (year)	Algorithm	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Park et al. (2020)	RF	Nested stratified 5-CV (I)	Grid search	Data and code available, no SST, *I validation*, *NCV*, no CI, no SEN, no OH, no BA, no GTV	Probable AD AD/non-AD	82.3/-/79.5/-/89.8	-/-/-/-/-
					Definite AD (AD/non-AD)	78.8/-/72.3/-/85	-/-/-/-/-
Wang et al. (2020b)	EN-ROC	–	–	No data nor code, no validation, no SST, small dataset, no SEN, no CI, no OH, no BA, no GTV	CN vs. AD	-/-/-/-/-	86.4/-/-/-/-
Richhariya et al. (2020)	Linear kernel SVM	5-CV (-)	MOSEK toolbox	No data nor code, no validation, no SST, small dataset, no SEN, no CI, no BA, no GTV	CN vs AD	100/-/-/-/-	-/-/-/-/-
					CN vs MCI	90/-/-/-/-	-/-/-/-/-
					MCI vs AD	73.7/-/-/-/-	-/-/-/-/-
Carvalho et al. (2020)	SVM, A1DE	10-CV + LOOCV (I/E)	AutoWEKA	Data and code were available, *OH*, *I/E validation*, no SST, small dataset, no SEN, no BA, no GTV	AD vs. not AD	85/-/-/90/89	55/-/-/-/-
					MCI vs. not MCI	95/-/-/95/97	100/-/-/-/-
Durongbhan et al. (2019)	KNN(K = 2)	10-CV (-)	–	Data and code were available, no OH, no CI, no validation, small dataset size, no SEN, no BA, no GTV	CN vs. AD	99/-/-/-/-	-

Table 3 (continued)

Ref. (year)	Algorithm	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Vaithinathan and Parthiban (2019)	KNN	10-CV (-)	Grid search	*OH*, data and code not available, no CI, no SEN, no validation, no BA, no GTV, no SST	CN vs. AD	87.4/-/89.6/-/-	-/-/-/-/-
					MCI vs. AD	63.4/-/57.3/-/-	-/-/-/-/-
					CN vs. MCI	64.7/-/45.6/-/-	-/-/-/-/-
					sMCI vs. pMCI	66.4/-/60.2/-/-	-/-/-/-/-
Baskar et al. (2019)	KFCM based BPANN	15 rounds of 10-CV (I)	–	No data nor code available, SST, no OH, *I validation*, no SEN, no CI, no BA, GTV	CN vs. AD	97.6/-/-/-/-	-/-/-/-/-
					CN vs. MCI	95.4/-/-/-/-	-/-/-/-/-
					MCI vs. AD	96.4/-/-/-/-	-/-/-/-/-
Zhu et al. (2019)	SVM	10 times 10-CV (-)	Grid search	No data nor code available, *SST*, *OH*, no validation, no SEN, *CI*, no BA, no GTV	AD vs. CN	90.3/-/91.5/-/91.2	-/-/-/-/-
					MCI vs. CN	72.2/-/68.8/-/79	-/-/-/-/-
					pMCI vs. sMCI	71.3/-/68.1/-/78.1	-/-/-/-/-
					AD vs. CN vs. MCI	63.9/-/-/-/-	-/-/-/-/-
					AD vs. CN vs. pMCI vs. sMCI	59.3/-/-/-/-	-/-/-/-/-
Li et al. (2019b)	Discriminant Analysis Classifier	Random division	–	No data and code available, no SST, no CI, no OH, no validation, no SEN, no BA, no GTV	CN vs. AD	84.6/-/92/-/80	-/-/-/-/-
Bucholc et al. (2019)	SVM + KRR	90:10 train:test, LOOCV (I)	Grid search	No data and code available, no SST, *CI*, no OH, *I validation*, no SEN, no BA, no GTV	CN vs. MCI vs. AD	-/-/-/-/-	83/-/89.7/-/94.9
					CDR prediction	$R^2 = 0.832$	

Table 3 (continued)

Ref. (year)	Algorithm	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Gómez-Sancho et al. (2018)	Regularized LR	Repeated nested 10-CV (-)	Grid search	OH, CI, no data nor code available, no validation, no SEN, no BA, no GTV	sMCI vs. pMCI	71.7/-/84.1/-/79.6	-/-/-/-/-
Zeng et al. (2018)	SVM	10-CV (-)	SDPSO	No data nor code available, *OH*, no SST, no validation, small dataset, no SEN, no CI, no BA, no GTV	sMCI vs. pMCI	69.2/-/-/-/-	-/-/-/-/-
					CN vs. AD	81.3/-/-/-/-	-/-/-/-/-
					CN vs. sMCI	76.9/-/-/-/-	-/-/-/-/-
					CN vs. pMCI	85.7/-/-/-/-	-/-/-/-/-
					sMCI vs. AD	71.2/-/-/-/-	-/-/-/-/-
					pMCI vs. AD	57.1/-/-/-/-	-/-/-/-/-
Tong et al. (2017)	SVM, RF	10-CV (-)	Grid search	*Data and code available*, *Student's t-test based SST*, no CI, *OH*, no validation, no SEN, no BA, no GTV	sMCI vs. pMCI	81/-/-/-/87	-/-/-/-/-
					sMCI vs. pMCI	84/-/-/-/92	-/-/-/-/-
Sørensen et al. (2017)	LDA	Stratified 10-CV (E)	Grid search	No data and code available, *McNemar's test based SST*, no CI, no OH, *E validation*, no SEN, no BA, no GTV	CN vs. MCI vs. AD	62.7/-/-/-/-	63/-/-/-/78.5

Table 3 (continued)

Ref. (year)	Algorithm	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Lu et al. (2017a)	SVM	Stratified inner, 3-CV (-)	Robust optimization	No data and code available, no SST, no CI, no OH, no validation, no SEN, no BA, no GTV	CN vs. MCI	92.1/-/88.5/-/-	-/-/-/-/-

Results (accuracy/precision/recall/f-score/AUC); *Domain expert*: did you include domain expert in model evaluation and data preparation steps? *CV(I/E)*: The cross-validation technique (internal validation/external validation); *STHNT*: Sheffield Teaching Hospitals NHS Trust memory clinic; *SEN*: the sensitivity of the model to little changes in input data has been evaluated; *CAD*: Center for Alzheimer's Disease in Brazil; *CRASI*: Center of Reference in Attention to Health of the Elderly in Brazil; *NSB*: Neuropsychological battery features; *CS*: Cognitive score; *CI*: Confidence interval was given; *OH*: Optimization hyperparameters were given; *A1DE*: Averaged One Dependence Estimators; *PCA*: principal component analysis; *uMCI*: unknown MCI; *PUTHC*: Peking University Third Hospital of China; *SPCA*: Sparse PCA; *EN-ROC*: Modified elastic net logistic regression; *KNHISD*: Korean National Health Insurance Service Database; *GA*: genetic algorithms; *GM*: Gray matter; *WM*: White matter; *BA*: Whether the study made bias and imbalance analysis of the dataset; *HC*: Hippocampus, *PCC*: Posterior Cingulate Cortex; *GTV*: Whether ground truth has been verified; *KFCM*: Kernel fuzzy c-means clustering; *BPANN*: Back-propagation artificial neural network; *AAL*: Automated Anatomical Labeling atlas; SST: The study performed statistical significance testing for the results; *NLMT*: Non-Local Means technique; *VBM*: Voxel-based morphometry; *USVM-RFE*: Universum SVM based recursive feature elimination; *GKWFW*: Gaussian kernel with full width at half maximum (FWHM) of 8 mm; *FS*: Feature selection; *AIBL*: Australian Imaging Biomarkers and Lifestyle flagship study of ageing; *LDA*: Linear discriminant analysis; TSSRFS: Task-specific relations and self-representation base feature selection; *IPPS*: Skull stripping, spatial standardization and segmentation, modulation, smoothing, registration, selecting the gray matter volume of the 90 auto-labeled regions of interest (ROI) as features; *CFS*: Correlation-based feature subset selection; *CFA*: (FAQ, ADAS13, MoCA, MMSE); *KRR*: Kernel Ridge Regression

(Forouzannezhad, et al. 2018): (1) skull stripping to remove non-brain tissue; (2) standardization and segmentation to extract the grey matter, white matter, and cerebrospinal fluid regions; (3) modulation to extract volume feature; (4) smoothing to remove noise; and (4) registration with a template such as Automated Anatomical Labeling (AAL). In addition, the image processing techniques applied to these images do not make sense from the medical point of view, and the resulting features may not be relevant to the physician from the explainability point of view.

Some studies reported that they did a k-fold CV after dividing the datasets into training and testing. However, they reported the CV results only, and very few papers reported the testing results (i.e., internal validation) (Sørensen et al. 2017; Wang et al. 2020b) (Carvalho et al. 2020). The major problem is that datasets were separated after preprocessing (Eke et al. 2021), which could result in a data leakage problem. Some studies, like (Park et al. 2020), used stratified nested CV, and some studies repeated the CV to confirm their results (Divya et al. 2021; Baskar et al. 2019)). Bucholc et al. (2019) implemented the validation process by keeping 10% of the data untouched for the model validation from the very beginning. The training data, i.e., 90% of data, were used in the preprocessing, FS, SMOTE balancing, and hyperparameter optimization steps using the LOOCV technique. The resulting features were masked on the test set. Authors compared many FS techniques using 10-times repeated 10-CV by combining data modalities, and they tuned ML models for predicting AD. This study produced a classifier to determine the patient class and a regressor to predict the disease severity by the clinical dementia rating (CDR) score. However, the study did not repeat the 90:10 split many times and did not use a suitable statistically significant test to compare the generated ML models. Some studies referred to the concept of robustness (Lu et al. 2017a; Li et al. 2019b), but they did not handle any of the TAI requirements for the robustness dimension. As far as we know, no study performed external validation based on a different dataset from the one used in model building, even if external validation is a crucial step to report the model generalization performance confidently (Birkenbihl et al. 2020). The grid search optimization technique is commonly used for hyperparameter optimization (Park, et al. 2020; Bucholc et al. 2019). In addition, the classification task is usually carried out as a binary classification task. Only some authors (Zhu et al. 2019; Sørensen et al. 2017) formulated the classification problem as a multiclass classification task, but as expected, they achieved lower performance than the results reported by binary classification tasks. Most studies reported Accuracy and AUC performance metrics, but authors did not consult domain experts to determine the relevant metrics a priori. In addition, if data were biased or imbalanced, the usual metrics are not representative enough. Regarding the reproducibility requirement, a few studies (Poloni et al. 2021; Zhu et al. 2019; Gómez-Sancho et al. 2018) reported results along with confidence intervals and/or statistical hypothesis testing (e.g., the McNemar's test is applied in Sørensen et al. (2017), and the Student's t-test is applied in Tong et al. (2017)). On the contrary, it is common reporting the method for optimization and the optimized hyperparameters (Divya and R. Shantha Selva Kumari 2021; Park, et al. 2020; Zhu et al. 2019; Carvalho et al. 2020; Tong et al. 2017). Moreover, no study performed any analysis to evaluate the sensitivity of the proposed model to slight changes in the input data; no study performed bias and imbalance analysis of the dataset, and no study verified the ground truth of the used dataset.

It is worth noting that most studies were based on the ADNI dataset, which requires authors to deal with many challenging TAI issues. Nevertheless, most of them are simply ignored. For example, ADNI owners did not provide a clear evaluation from the medical point of view for the completeness and representativeness of the data. ADNI is also a

multimodal time-series dataset. However, the number of time steps in the dataset is insufficient to perform accurate longitudinal analysis and track the disease's progress. In addition, the dataset is very sparse, and many missing values are hard to fill in using existing preprocessing techniques. In addition, ADNI did not provide any explanation about the ground truth of the given labels. Physicians assign patient diagnosis and progression detection labels manually based on the patient's MMSE measurement (Sørensen et al. 2017). As a result, given labels may not be accurate enough. Surprisingly, the ADNI dataset is biased in race, ethnicity, language, and marriage status. For example, the *Hisp/Latino* ethnicity represents more than 97% of the patients. The *race* of patients is distributed as 91.7% for white, 4.9% for black, 2.0% for Asian, 0.18% for Indian/Alaskan, and 1.22% for more than one race. As a result, recommendations reported in most studies may not be suitable for underrepresented races since it has been reported that the brain shape is variant across ethnicities (Bin Bae et al. 2020). ADNI includes mainly *married* people (75.9%).

Furthermore, ADNI provided neither a detailed datasheet for describing how data were collected, details about the data preparation steps, nor data privacy and security settings. In addition, comparing results and conclusions drawn in different studies is not straightforward because they usually consider different criteria for selecting their cohorts regarding the patient's cognitive scores like MMSE and CDR, demographics like age, and symptoms like depression. For example, Zhu et al. (2019) selected CN patients with CDR of 0 and MMSE between 24 and 30, but El-Sappagh et al. (2020) selected CN patients with CDR of 0.384 and MMSE between 27.84 and 30.12. No standardization and not considering domain experts in the ML pipeline complicate the reproducibility and general acceptance of results. Some studies merge datasets from multiple data sources (Wang et al. 2020b) (Li et al. 2019b). Sørensen et al. (2017) trained a linear discriminant analysis (LDA) algorithm to detect AD patients based on MRI data from ADNI and AIBL datasets. Then they externally tested their model based on CADDementia testing data. Their results are significantly better than other studies. The learning curve analysis asserted that no more data nor complex algorithms could improve the performance. The most relevant features were cortical thickness, volumetric, and hippocampus shape and texture. However, the combined ADNI and AIBL data might not come from the same distribution because the used raw ADNI data were 1.5T T1-weighted images and raw AIBL data were 3T T1-weighted images. In addition, testing CADDementia data were 3T T1-weighted scans. Authors did not consult domain experts to assure the possible effect of this difference and did not consider the potential effect in model training. *Note that these data can be used to evaluate model reproducibility by training with one dataset and validate with the other, but in these studies the data are combined and then used to train and validate models.* In addition, comparing ADNI, AIBL, and CADDementia datasets, approximately 50% of CADDementia CN group are controls with subjective complaints, but they are only 20% in cse of ADNI and AIBL. CN group in ADNI has no memory complaints but approximately 50% of CN in AIBL had subjective memory complaints. ADNI and AIBL have only mild AD patients at baseline, but CADDementia did not restrict the severity (Sørensen et al. 2017).

5.1.2 Deep learning models

Even though conventional ML models usually have a feature engineering step, the extracted features may not capture the full characteristics of brain atrophy (Cui and Liu 2019). On the contrary, DL models can implicitly learn deep representations from unstructured data like images and text. Table 4 provides a comparison between 17 studies where DL models

were applied for AD diagnosis and progression detection. We consider again the same metrics that we considered in Table 3. These studies used neither ensemble models, multimodality data, nor time-series data. They are mainly CNN models supported by neuroimaging (especially sMRI and fMRI) (Wen et al. 2020; Chen and Xia 2021; Buvaneswari and Gayathri 2021; Duc et al. 2020; Lian et al. 2020a, b; Abrol et al. 2020; Chitradevi and Prabha 2020; Basheera and Satya Sai Ram 2020; Ju et al. 2017; Li et al. 2019c; Basaia, et al. 2018). CNN models are recognized because of their ability to automatically learn complex, deep, and spatial representations from images. Normally, CNNs require minimal preprocessing because they can automatically extract low-to-high level features. As seen in Table 4, all studies took different steps to prepare neuroimages but without systematically assessing their impact (Basheera and Sai Ram 2019). This is most likely because datasets are relatively small, and thus CNNs are hard to train. Buvaneswari and Gayathri (2021) applied the CNN model (SegNet) for MRI segmentation to extract the brain's morphological change features and then another CNN model (ResNet-101) for classification. Because the same data were used for feature extraction and classification, this model is expected to be biased. Hedayati et al. (2020) used an ensemble of 11 CNN-based autoencoders to extract features from the gray matter of 3D MRI images and then stacked the selected features to be exploited by another 3D CNN model. Data are used with a tenfold CV, but no additional validation is reported. Other studies (Chen and Xia 2021) used a DL model (e.g., ResNet10) for representation learning and a conventional ML model (e.g., sparse regression module) for AD prediction.

Most studies are based on the ADNI dataset because it is the largest dataset in the AD domain (Jo et al. 2020). Compared to the conventional ML studies shown in Table 3, the used datasets for DL models are more prominent in the number of cases. Nine studies considered domain expert opinion in the ML pipeline (Duc et al. 2020; Lian et al. 2020a; Jo et al. 2020). On the one hand, some studies depended on a single dataset which was randomly divided into train, validation, and testing sets (Buvaneswari and Gayathri 2021; Basheera and Satya Sai Ram 2020; Pan et al. 2021). As a result, they reported internal validation only. The splitting process is repeated multiple times, and average results are reported (Pan et al. 2021). On the other hand, some studies depended on one dataset but performed CV (Li et al. 2019c; Chen and Xia 2021; Duc et al. 2020). Some studies used repeated stratified CV to remove any bias in the reported results (Jo et al. 2020; Abrol et al. 2020; Ju et al. 2017). These studies did not perform either internal or external validation, and they reported only the CV results. In addition, other studies used more than one dataset (Lian et al. 2020b; Li et al. 2019c; Basaia, et al. 2018). Datasets can be combined and then split randomly for model training and testing (Basaia, et al. 2018). In addition, each dataset can be divided into train/test/validate and used to train and validate the model separately (Yue et al. 2019). Some datasets are kept for model testing only (Lian et al. 2020a, b; Li et al. 2019c). However, different datasets usually rely on different criteria for subject labeling and are likely to have different label distributions (e.g., OASIS, ADNI, and AIBL, which are the most popular datasets in the AD domain (Wen et al. 2020)). As a result, the case of building a model with one dataset and using another dataset to check the model generalizability is unfair (e.g., Li et al. (2019c) checked model reproducibility by training a DL model based on the ADNI-1 cohort and evaluated its prognostic performance based on the ADNI-GO&2 and AIBL cohorts. In addition, the impact of images generated by different scanners, i.e., 1.5T and 3T scanners).

The most rigorous study in Table 4 is Wen et al. (2020), where authors used three datasets (ADNI, AIBL, and OASIS) to validate their models. They implemented many CNN architectures to explore the role of different MRI features and preprocessing steps, i.e.,

Table 4 Review of studies based on DL models

Ref. (Year)	Data set	Subjects	Dom. expert	Modality	Task (purpose)	Data preparation	DL model
Hedayati et al. (2020)	ADNI	CN (100), MCI (100), AD (100)	NO	3D MRI	AD detection (classification)	Skull Stripping & Segmentation of GM, GM feature extraction by combining vectors of an ensemble of 11 CNN-based autoencoders	3D CNN: 3 conv blocks [conv + max-pooling] + flatten + 2 FC layers + output layer
Pan et al. (2021)	ADNI	AD (237), CN (242), pMCI (166), sMCI (360)	NO	3D FDG-PET	AD progression (classification)	Spatial normalization, intensity normalization, smoothing	CNN (MiSePyNet): 3 CNNs models for axial, coronal, sagittal data, and each model has 3 2D CNNs, Concat, 2 FC layers, output layer
Chen and Xia (2021)	ADNI	AD (347), CN (417), pMCI (179), sMCI (305)	NO	MRI	AD detection (classification)	Deep feature extraction (3D ResNet10) for local-to-global structural information from 62 cortical regions	Sparse regression module based on the features extracted from ResNet
Buvaneswari and Gayathri (2021)	ADNI	CN (80), MCI (80), AD (80)	NO	sMRI	AD detection (classification)	Skull stripping, SegNet-based Segmentation for GM, WM, CSF, HC, and cortical thickness, augmentation	CNN (ResNet-101): regular ResNet with extra 1 × 1 conv shorting pass layer

Table 4 (continued)

Ref. (Year)	Data set	Subjects	Dom. expert	Modality	Task (purpose)	Data preparation	DL model
Li et al. (2021)	ADNI	AD (194), pMCI (164), sMCI (233), CN (216)	NO	3D sMRI	AD progression (classification)	Normalization, skull-stripping, and cerebellum removal, registration, segmentation to create bilateral HC binary masks	Cascaded 2D CNN on multi-channel 3D CNNs: 3 CNNs with (4 Conv and 3 pooling layers + 3 FC layers) + 3 flatten + 2 2D CNNs + concatenate + 1 FC + SofMax output layer
Duc et al. (2020)	CUNDRC	CN (198), AD (133)	YES	rs-fMRI	AD detection (classification)	Motion correction, slice timing correction, spatial smoothing, Gaussian kernels, temporal filtering, co-registration, 30 ICA, gICA-based 3D special maps	3D CNN (modified VGG-Net) 5 Conv blocks with BN layers in Conv block and ReLU activation + 2 FC layer + output layer
Puente-Castro et al. (2020)	ADNI OASIS	ADNI (1743), OASIS (436)	NO	MRI, sex, age	AD progression (classification)	Volume extraction, data balancing using weights	Hybrid model (ResNet-SVM)
Abrol et al. (2020)	ADNI	CN (237),sMCI (245), pMCI (157), AD (189)	NO	MRI	AD progression (classification)	Segmentation of GM areas, normalization, smoothing by 3D Gaussian kernel	CNN (ResNet with D = 3), D is the number of residual blocks

Table 4 (continued)

Ref. (Year)	Data set	Subjects	Dom. expert	Modality	Task (purpose)	Data preparation	DL model
Wen et al. (2020)	ADNI, AIBL, OASIS	**ADNI:** CN(330), MCI (787), sMCI(298), pMCI(295), AD(336), **AIBL:** CN(429), MCI (93), sMCI(13), pMCI (20), AD (76), **OASIS:** CN (76), AD (78)	YES	3D MRI	AD detection (classification)	Convert to BIDS, bias field correction, (non)linear registration, cropping, intensity rescaling, skull stripping	Soft voting of 4 CNNs for 2D slice-level, 3D ROI-based, 3D patch-level, and 3D subject-level. Each CNN has 4 conv blocks + 2 FC layers + BN and dropout
Chitradevi and Prabha (2020)	CHC	CN (100), AD (100)	YES	MRI	AD detection (classification)	Contrast enhancement, skull stripping, GWO based segmentation for GM, WM, CC, HC	CNN (AlexNet)
Basheera and M. Satya Sai Ram (2020)	ADNI	AD (65), CN (28), MCI (32)	NO	MRI	AD detection (classification)	Slicing, Gaussian filtering, skull stripping, segmentation by ICA, volume extraction of GM	CNN with 5 conv layers + max pooling + 6 FC layers + output layer
Lian et al. (2020b)	ADNI, AIBL	**ADNI-1**: CN(229), sMCI(226), pMCI(167), AD(199), **ADNI-2**: CN(201), sMCI(239), pMCI(38), AD(159), **AIBL**: CN(447), AD(72)	YES	3D sMRI	AD progression (classification)	N3 correction, skull and dura stripping, linear align to Colin27 template, crop to size 144 × 184 × 152	Fully CNN [12 conv + 0 padding + BN + ReLU + 1 GAP + 1 FC] + multi-branch hybrid CNN (HybNet) [global branch, local branch, fusion branch] + output layer

Table 4 (continued)

Ref. (Year)	Data set	Subjects	Dom. expert	Modality	Task (purpose)	Data preparation	DL model
Ju et al. (2017)	ADNI	CN (79), MCI (91)	YES	rs-fMRI	MCI detection (classification)	Motion correction, slice timing correction, skull stripping, normalization, smoothing, filtering, AAL-based 90 regions registration	Pearson's correlation matrix + FS + 2 hidden layers stacked Autoencoder
Li et al. (2019c)	ADNI, AIBL	**ADNI**: CN (539), MCI (822), AD (350), **AIBL**: CN (328), MCI (40), AD (67)	YES	MRI	AD progression (classification)	HC extraction (registration, segmentation)	2 CNNs (ResNet) for left and right HC + fusion by 1 concat + dropout layer + 1 FC layer + output layer; LASSO regularized Cox
Basaia et al. (2018)	ADNI, MILAN	**ADNI:** CN (407), AD (418), pMCI (280), sMCI (533); **MILAN:** CN (55), MCI (23), pMCI (27), AD (124)	YES	3D MRI	AD progression (classification)	Registration, segmentation to produce GM, WM, CSF, normalization + data augmentation	3D CNN with 12 repeated blocks of Conv layers + ReLU + 1 FC + output layer
Basheera and Sai Ram (2019)	ADNI, PMCC	**ADNI:** AD (635), MCI (548), CN (637), **PMCC**: AD (9), MCI (5), CN (7)	NO	MRI	AD detection (classification)	Gaussian filter enhanced voxels, skull stripping removed irrelevant tissues, segmentation of GM volume by ICA	CNN with 5 Conv layers and Max pooling layers + 6 FC layers + output layer

Table 4 (continued)

Ref. (Year)	Data set	Subjects	Dom. expert	Modality	Task (purpose)	Data preparation	DL model
Yue et al. (2019)	ADNI-1, ADNI-2	**ADNI-1:** NC (785), MCI (542), AD (335), **ADNI-2:** NC (1106), early MCI (1320), late MCI (987), AD (305)	NO	3D sMRI	AD detection (classification)	Normalization, segmentation of GM, WM, CSF, smoothing GM, use AAL to segment 90 ROI	CNN with 3 conv layers with 3 pooling layers + 2 FC layers + output layer

Ref. (Year)	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Hedayati et al. (2020)	10 times repeated 10-CV (-)	Adam optimizer, 200 epochs	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. AD MCI vs. AD CN vs. MCI	95.0/100/92/-/- 90.0/85/94.4/-/- 92.5/100/85.7/-/-	-/-/-/-/- -/-/-/-/- -/-/-/-/-
Pan et al. (2021)	10 times stratified (train/validation/ test) (60/20/20) (I)	SGD, 40 epochs	No data and code available, **SST**, no CI, **not small dataset**, **I validation**, no SEN, no BA, no GTV	sMCI vs. pMCI CN vs. AD	-/-/-/-/- -/-/-/-/-	83.1/-/72.1/-/86.8 93.1/-/90.3/-/97.1

Table 4 (continued)

Ref. (Year)	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Chen and Xia (2021)	5-CV (-)	Adam optimizer, 50 epochs	No data and code available, no SST, no CI, **not small dataset**, no validation, no SEN, no BA, no GTV	CN vs. AD	95.3/-/91.2/-/-	-/-/-/-/-
				pMCI vs. sMCI	77.6/-/71.6/-/-	-/-/-/-/-
Buvaneswari and Gayathri (2021)	Age/gender balanced train/test split (-)	–	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	Cn vs. MCI vs. AD	96.3/-/96.7/-/-	-/-/-/-/-
Li et al. (2021)	5-CV (-)	Adadelta optimizer	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. AD	85.9/-/81.5/-/88.4	-/-/-/-/-
				CN vs. MCI	73.3/-/78.1/-/74.6	-/-/-/-/-
				sMCI vs. pMCI	71.0/-/59.8/-/71.9	-/-/-/-/-

Table 4 (continued)

Ref. (Year)	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Duc et al. (2020)	Stratified 10-CV (-)	Adam optimizer, 50 epochs	No data and code available, no SST, **CI**, small dataset, no validation, no SEN, no BA, no GTV	CN vs. AD	85.3/-/67.2/-/-	85.3/-/-/-/-
Puente-Castro et al. (2020)	LOOCV, and 50-times repeated train:test	Random Search	**Data and code available**, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. sMCI vs. pMCI vs. AD (OASIS, ADNI)	86.8/32.1/37.8/33.5/- 78.6/68.9/58.3/60.3/-	-/-/-/-/- -/-/-/-/-
Abrol et al. (2020)	10 times repeated, stratified 5-CV (I)	Adam Optimizer, 100 epochs	No data and code available, **ANOVA SST**, no CI, **not small dataset**, **I validation**, no SEN, no BA, no GTV	CN vs. AD CN vs. pMCI sMCI vs. AD sMCI vs. pMCI CN vs. sMCI vs. pMCI vs. AD	91.0/-/-/94 89.3/-/-/90 88.1/-/-/90 77.8/-/-/-78 53.8/-/-/-	89.3/-/-/- 86.5/-/-/- 87.5/-/-/- 75.1/-/-/- 51.41/-/-/-

Table 4 (continued)

Ref. (Year)	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Wen et al. (2020)	Train/ validation/ test, then 5-CV using train/ validation (I/E)	–	No data but **code available**, no SST, **CI**, **not small dataset**, **I/E validation**, no SEN, **BA**, no GTV	CN vs. AD (balanced accuracy)	88.0/-/-/-/-	89.0/-/-/-/-
				sMCI vs. pMCI (balanced accuracy)	77.0/-/-/-/-	74.0/-/-/-/-
Chitradevi and Prabha (2020)	–	–	No data but **code available**, no SST, no CI, small dataset, no validation, no SEN, no BA, **GTV**	CN vs. AD	95/-/95/-/-	-/-/-/-/-
Basheera and Satya Sai Ram (2020)	80: 20 Random splitting (I)	Adam Optimizer, 200 epochs	No data and code available, no SST, **CI**, small dataset, **I validation**, no SEN, no BA, no GTV	CN vs. AD	100/-/100/-/-	-/-/-/-/-
				MCI vs. AD	96.2/-/93.0/-/-	-/-/-/-/-
				CN vs. MCI	98.0/-/96.0/-/-	98/-/-/-/-
				CN vs. MCI vs. AD	86.7/-/89.6/-/-	-/-/-/-/-

Table 4 (continued)

Ref. (Year)	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Lian et al. (2020b)	Train by ADNI-1 and validate by ADNI-2 and AIBL (E)	Adam optimizer	No data and code available, no SST, no CI, **not small dataset**, **E validation**, no SEN, no BA, no GTV	CN vs. AD (ADNI2 + AIBL)	-/-/-/-/-	91.9/-/88.7/-/96.5
					-/-/-/-/-	89.8/-/87.3/-/94.6
				sMCI vs. pMCI	-/-/-/-/-	82.7/-/57.9/-/-79.3
Ju et al. (2017)	10 times repeated 10-CV (-)	L-BFGS, 50 epochs	No data and code available, no SST, **CI**, small dataset, no validation, no SEN, no BA, no GTV	CN vs. MCI	86. 5/-/-/-/91.6	-/-/-/-/-
Li et al. (2019c)	10-CV (E)	SGD, 100,000 epochs	No data and code available, **WIL based SST**, **not small dataset**, no CI, **E validation**, no SEN, no BA, no GTV	CN vs. AD (ADNI)	-/-/-/-/-	90/-/-/-/95.6
				CN vs. AD (AIBL)	-/-/-/-/-	92-/-/-/-/95.8

Table 4 (continued)

Ref. (Year)	CV (I/E)	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Basaia et al. (2018)	Random split of 90–10; 10-CV on the 90% and test on 10% (I)	–	No data and code available, no SST, no CI, **not small dataset**, **I validation**, no SEN, no BA, no GTV	CN vs. AD (ADNI, ADNI + Milan)	-/-/-/-/- -/-/-/-/-	99.2/-/98.9/-/- 98.2/-/98.1/-/-
				CN vs, pMCI (ADNI,ADNI + Milan)	-/-/-/-/- -/-/-/-/-	87.1/-/87.8/-/- 87.7/-/87.3/-/-
				CN vs. sMCI (ADNI, ADNI + Milan)	-/-/-/-/- -/-/-/-/-	76.1/-/75.1/-/- 76.4/-/75.1/-/-
				pMCI vs. AD (ADNI, ADNI + Milan)	-/-/-/-/- -/-/-/-/-	75.4/-/74.5/-/- 75.8/-/74.8/-/-
				sMCI vs. AD (ADNI, ADNI + Milan)	-/-/-/-/- -/-/-/-/-	85.9/-/83.6/-/- 86.3/-/84.0/-/-
				sMCI vs. pMCI (ADNI,ADNI + Milan)	-/-/-/-/- -/-/-/-/-	75.1/-/74.8/-/- 74.9/-/75.8/-/-
Basheera and Sai Ram (2019)	Random split in 80:20 for train:test (E)	Adam optimizer, 200 epochs	No data and code available, no SST, no CI, small testing dataset, **E validation**, no SEN, no BA, no GTV	CN vs. MCI vs. AD	86.7/-/89.6/-/88.5	90.5/ 92.6/ 86.7/-/-
				CN vs. AD	100/-/100/-/100	-/-/-/-/-
				AD vs. MCI	96.2/-/93.0/-/98.7	-/-/-/-/-
				CN vs. MCI	98.0/-/96.0/-/99.9	-/-/-/-/-

Table 4 (continued)

Ref. (Year)	CV (I/E)	Hyper. Opt	Reproducibility	Target		CV results	Testing results
Yue et al. (2019)	Random split in 80:20 for train:test, 5-CV of train (I)	–	No data and code available, **Pearson, Kendall, Spearman SST**, **CI**, **not small dataset**, no validation, no SEN, no BA, no GTV	ADNI 1	AD vs. MCI	97.8/-/-/-/-	-/-/-/-/-
					CN vs. AD	99.7/-/-/-/-	-/-/-/-/-
					CN vs. MCI	98.9/-/-/-/-	-/-/-/-/-
				ADNI 2	CN vs. AD	98.6/-/-/-/-	-/-/-/-/-
					LMCI vs. AD	99.8/-/-/-/-	-/-/-/-/-
					EMCI vs. AD	99.8/-/-/-/-	-/-/-/-/-
					CN vs. EMCI	99.6/-/-/-/-	-/-/-/-/-
					EMCI vs LMCI	98.0/-/-/-/-	-/-/-/-/-

FC: Fully connected layer; *GAP*: Global average pooling; *WIL*: Wilcoxon rank sum tests; *GM*: Gray matter; *WM*: White matter; *CC*: Corpus callosum; *HC*: Hippocampus; ICA: Independent component analysis; *SGD*: Stochastic gradient descent algorithm; *GWO*: Grey wolf optimization; *CHC*: Chettinad Health City, Chennai; *BN*: Batch normalization; *MILAN*: Dataset recruited from the Department of Neurology, Scientific Institute and University Vita-Salute San Raffaele, Milan; *ICA*: Independent component analysis; *CNN*: Convolution neural network; *PMCC*: Poorvi MRI Center, Chirala; *OASIS*: Open Access Series of Imaging Studies; *BIDS*: Brain imaging data structure; AC-PC: anterior commissure (AC)-posterior commissure (PC); *ISDL*: Iterative sparse and DL; *CUNDRC*: Chosun University National Dementia Research Center (Gwangju, South Korea); *BFN*: Brain functional network; *FCO*: Functional Connectivity; *L-BFGS*: Limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm; *MiSePyNet*: Multi-view Separable Pyramid Network

2D slice-level, 3D patch-level, ROI-based, and 3D subject-level features. To prevent possible data leakage, the ADNI data were split into training/validation/test sets at the very beginning of the pipeline, and only the training/validation sets were used for model selection. The test set intended for internal validation remained unmodified until the very end of model evaluation. Train/validation sets are used to train and validate the models with a tenfold CV. The authors noted that the distributions of all considered datasets were not significantly different. AIBL and OASIS datasets were used for external validation.

All studies except (Wen et al. 2020; Chitradevi and Prabha 2020) did not share the datasets and code, but all studies described the used DL architecture and utilized hyperparameters such as the number of epochs, learning rate, etc. Half of the studies used relatively large datasets compared to studies on conventional ML models in Table 3 (Wen et al. 2020; Chen and Xia 2021; Lian et al. 2020b; Li et al. 2019c; Basaia, et al. 2018; Pan et al. 2021; Yue et al. 2019). However, the rest of studies used small datasets which could cause model overfitting (Buvaneswari and Gayathri 2021) (Duc et al. 2020; Lian et al. 2020a; Chitradevi and Prabha 2020; Basheera and Satya Sai Ram 2020; Ju et al. 2017; Jo et al. 2020). Basheera et al. (2019) used a large dataset for model validation and used a small but independent dataset to perform external validation. Most studies did not use any statistical significance tests to measure and compare results (Wen et al. 2020; Chen and Xia 2021; Buvaneswari and Gayathri 2021; Duc et al. 2020; Chitradevi and Prabha 2020; Basheera and M. Satya Sai Ram 2020; Lian et al. 2020b; Ju et al. 2017) (Basaia et al. 2018; Basheera and Sai Ram 2019; Jo et al. 2020). As a result, some authors could argue that their implemented models were better than other models in the literature, but this conclusion is misleading because it was not supported by significant statistical evidence. Some studies applied statistical tests like the Student t-test (Lian et al. 2020a), ANOVA (Abrol et al. 2020), and Wilcoxon rank-sum tests (Li et al. 2019c). However, the authors did not justify why they selected these specific tests. Only five studies reported the confidence intervals for their results which measure the model stability (Wen et al. 2020; Duc et al. 2020; Basheera and Satya Sai Ram 2020; Ju et al. 2017; Yue et al. 2019). No studies did any sensitivity analysis of models against small changes in the input data. No studies statistically did bias and imbalance analysis of the selected datasets. Finally, no studies, except (Chitradevi and Prabha 2020), checked the correctness of the ground truth of the used datasets. In addition, Wen et al. (2020) mentioned that the classification performance of DL models in different studies is not easy to compare because different studies consider different MRI sections, different MRI preprocessing steps, or different validation procedures. They surveyed 30 research studies and concluded that more than 15 papers have suffered from data leakage and reported biased results. In addition, the reported results were hardly reproducible because their codes were not publicly accessible and because implementation details were missing. Moreover, many studies reported biased performance because of insufficient or unclear validation or model selection procedures.

5.1.3 Ensemble models

As seen in previous sections, neither conventional nor DL models achieved stable and medically relevant results for AD. Using an ensemble of multiple diverse classifiers (i.e., a multi-classifier system) with different validation settings can yield more stable and robust models and, as a result, enhance classification performance (Lu et al. 2018a; Lei et al. 2019). This is because different and complementary base classifiers can compensate for the errors of a single classifier. Nevertheless, it depends on the level of diversity and

independence of the base classifiers. In the AD domain, Armañanzas et al. (2017) concluded that there was no remarkable statistical difference in performance between ensemble and base classifiers. Still, ensemble models had better average behavior due to stability gained by combining individual outputs. Ensembles can use different ML models (e.g., DTs, SVM, logistic regression, or even other ensembles as RF) as base learners. In addition, DL models can be used as base classifiers in bagging, boosting, and stacking ensembles (Li et al. 2018). Ensemble models could mitigate challenging issues such as class imbalance, data bias, model overfitting, concept drift, and the curse of dimensionality (Sagi and Rokach 2018; Syed et al. 2020). In addition, ensembles can improve the robustness of DL models against adversarial attacks (Pang et al. 2019). Moreover, dynamic ensemble classifiers are used to extend the static selection (SS) of base classifiers to dynamic selection (DS), where the competence and selection of base classifiers are estimated on the fly according to each new sample to be classified (Cruz et al. 2018). Dynamic ensemble selection (DES) differs from static ensemble selection (SES) in the base classifiers' selection phase, either static or dynamic.

Table 5 provides readers with a comparison between 13 ensemble models in the AD domain. None of the studies considered domain experts. Some studies built ensembles based on conventional ML models as base classifiers (Lei et al. 2019; Armañanzas et al. 2017; Syed et al. 2020; Muhammed Niyas and Thiyagarajan 2021; Pan et al. 2019; Dimitriadis and Liparas 2018; Nanni et al. 2018; Jin and Deng 2018). Other studies used DL models as base classifiers (Lu et al. 2018a; An et al. 2019; Choi and Lee 2020; Wang et al. 2019c; Ahmed et al. 2019b). Diversity was implemented with different techniques: diverse base models (Syed et al. 2020; Choi and Lee 2020; Ahmed et al. 2019b), diverse input data (Muhammed Niyas and Thiyagarajan 2021; Dimitriadis and Liparas 2018; Jin and Deng 2018; An et al. 2019), and diverse base model and data (Wang et al. 2019c). Muhammed and Thiyagarajan (2021) is the only study that explored the role of DES in improving AD detection. The authors used ADNI's TADPOLE dataset of 1737 subjects and reported an enhanced performance for the META-DES algorithm with RF and bagged DT as base classifiers. An et al. (An et al. 2019) proposed a multilayer deep ensemble model based on a large sample of multimodal data. *First*, 100 diverse features were fused and normalized. *At the first layer (voting layer)*, two stacked autoencoders were used for feature learning and created two feature spaces. Next, three different base classifiers groups used these two feature spaces and the original feature space (using 35 classification algorithms) to generate different diagnoses. To improve diversity, the different base classifiers were trained using different data resampling methods. *At the second layer (stacking layer)*, a deep belief network (DBN) was used to combine base classifiers' predictions in a weighted manner. *At the third layer (optimizing layer)*, the cost-sensitive method's three feed-forward neural networks were parallelly built. The input to these networks is the DBN along with the outputs of the 35 classifiers. Half of the data (50%) were used for FS and then combined with another 30% of data to form the training set. The remaining 20% of the data were kept untouched for testing. However, this division was done after the data normalization process. The resulting model outperformed many other ensembles, but the authors did not report precise results.

It is worth noting that only two studies (Dimitriadis and Liparas 2018) performed the external validation process. Studies sometimes divided the dataset after the preprocessing steps or did not leave a test set at all. Syed et al. (2020) used the recursive feature elimination (RFE) technique for FS, where the study assigned weights (i.e., 0.2 to majority class, 0.8 to minority class) to classes to handle the imbalanced dataset. However, the authors depended on the whole dataset to perform the FS step. In addition, they firstly preprocessed

Table 5 Review of studies based on ensemble models

Ref. (Year)	Data set	Subjects	Modalities	Task	Data preparation + Feature eng	Base classifiers	Diversity source	Fusion method
M. N. K. P. and T. P. (2021)	ADNI	CN (523), MCI (872), AD (342)	MRI, PET, CSF, CS, (age, sex, education)	AD detection (Classification)	Data fusion, manual feature selection, data denoising [missing values, labeling, datatypes]	2 classifiers [RF + BDT]	Diverse ML models	5 diverse data types (Early fusion)
Syed et al. (2020)	Figshare	CN (242), MCI (91)	CSF protein biomarkers	MCI detection (Classification)	Z-score normalization, encoding, SVM-RFE and LR-RFE resulted in 3 features only	2 classifiers [LR + linear SVM]	Diverse ML models	–
An et al. (2019)	NACC UDS	23,165 samples with 100 features	7 groups (SGOD)	AD detection (Classification)	Normalization	35 based classifiers + 3 classifiers [DNN]	Diverse ML models	3 diverse feature spaces
Choi and Lee (2020)	ADNI	AD (175), MCI (305), CN (335)	3D MRI	AD detection (Classification)	Normalization,	32 classifiers [deep CNN as VGG16, GoogleNet, AlexNet]	Multiple MRI differential projections + multiple CNN architectures	–
Wang et al. (2019c)	ADNI	CN (315), MCI (279), AD (221)	MRI	AD detection (Classification)	Grad-warped, intensity corrected, scaled for gradient drift, stripping non-brain tissue, and template alignment	5 classifiers [3D-DenseNets]	3D-DenseNets models with different architectures	Subsets of features (late fusion)
Pan et al. (2019)	ADNI	CN (90), sMCI (44), pMCI (44), AD (94)	ADNI's Post processed FDG-PET	AD detection (Classification)	116 ROI segmentation by AAL + three levels feature extractions	*Level 1*: 7 classifiers [SVM] + *Level 2*: 3 classifiers [SVM]	7 LASSO FS on 7 different feature sets	Region based and connectivity between regions-based features (late fusion)
Ahmed et al. (2019b)	ADNI, GARD	ADNI: CN (129), AD (77), GARD: AD (81), CN (171)	sMRI	AD detection (Classification)	Intensity normalization, left and right HC extraction,	3 classifiers [CNNs] with different architectures	Three feature sets from TVPLH, TVPRH, TVPLHRH	Subsets of features (early fusion)

Table 5 (continued)

Ref. (Year)	Data set	Subjects	Modalities	Task	Data preparation + Feature eng	Base classifiers	Diversity source	Fusion method
Lei et al. (2019)	ADNI	CN (152), sMCI (98), pMCI (104), AD (91)	MRI	AD detection (Regression)	correction, skull stripping, segmentation (GM, WM, CSF), registration (93 ROI), feature extraction (GM volume)	[SVR]	Using CSTGL FS with different time steps and different SVR	–
Dimitriadis and Liparas (2018)	ADNI	**Training**: CN (60), sMCI (60), pMCI (60), AD (60), **Testing**: CN (40), sMCI (40), pMCI (40), AD (40)	MRI, MMSE, age, CSF	AD progression (classification)	Standard MRI preprocessing from ADNI + FS for each base model by Gini impurity index	5 classifiers [RF]	Diverse input data (DID1)	Subsets of features (early and late fusion)
Nanni et al. (2018)	ADNI	CN (60), sMCI (60), pMCI (60), AD (60)	MRI, age, gender, MMSE	AD detection (Classification)	Standard MRI preprocessing from ADNI + FS by FS and MI	4 classifiers [SVM]	Diverse input data (DID2)	Subsets of features (early and late fusion)
Lu et al. (2018a)	ADNI	CN (304), sMCI (409), pMCI (112), AD (226)	MRI, FDG-PET	AD progression (classification)	MRI GM and WM segmentation, voxel patch parcellation, MRI and PET co-registration, normalization	10 classifiers [DNN]	3 vectors of 1488, 705, 343 lengths for multiscale features	–
Jin and Deng (2018)	ADNI	CN (60), sMCI (60), pMCI (60), AD (60)	Preprocessed MRI, age, gender, MMSE	AD progression (classification)	FS by NCA	150 classifiers [decision tree]	Different tree architectures	–
Armañanzas et al. (2017)	NFDC	CN (51), AD (56)	fMRI	AD detection (Classification)	Normalization, motion correction, spatially normalization to MNI atlas, RE univariate feature selections	3 classifiers [SVMs]	Diverse input data	–

Table 5 (continued)

Ref. (Year)	Ensemble technique	CV	Hyp. Opt	Reproducibility	Target	CV results	Testing results
M. N. K. P. and T. P. (2021)	META-DES (DES)	Split 80:20 for train:test + stratified 10-CV on train (I)	Grid search	**Data available** and code not available, no SST, no CI, **not small dataset**, **I validation**, no SEN, no BA, no GTV	CN vs. MCI vs. AD (balanced accuracy)	87/-/-/-/-	82/-/80/-/-
Syed et al. (2020)	Weighted average (SES)	Stratified split 80:20 for train:test + 5-CV on train (I)	–	No data and code available, **Student t-test SST**, **CI**, small dataset, **I validation**, no SEN, no BA, no GTV	CN vs. MCI	-/-/-/-/-	95.5/-/95.7/-/97.9
An et al. (2019)	3 layers of voting, stacking, and optimization	50:30:20 split, then 50 for base classifiers, 50 + 30 for training, 20 for testing (I)	–	No data and code available, no SST, no CI, small dataset, **I validation**, no SEN, no BA, no GTV	–	-/-/-/-/-	Accuracy is 4% better than voting, bagging, stacking, RF, AdaBoostM1, and LogitBoost
Choi and Lee (2020)	Learnable weighted sum of probabilities	10 times repeated random split 60:40 for train:test (I)	SGD, 200 epochs	No data and code available, no SST, **CI**, small dataset, **I validation**, no SEN, no BA, no GTV	CN vs. AD MCI vs. AD CN vs. MCI CN vs. MCI vs. AD	-/-/-/-/- -/-/-/-/- -/-/-/-/- -/-/-/-/-	93.15/-/-/-/- 94.71/-/-/-/- 93.39/-/-/-/- 93.84/-/-/-/-

Table 5 (continued)

Ref. (Year)	Ensemble technique	CV	Hyp. Opt	Reproducibility	Target	CV results	Testing results
Wang et al. (2019c)	Learnable probability- based fusion	10 times repeated 10-CV (-)	–	No data and code available, no SST, **CI**, **not small dataset**, no validation, no SEN, no BA, no GTV	CN vs. MCI vs. AD	-/-/-/-/-	97.5/97.1/97.0/97.1/-
					sMCI vs. AD	-/-/-/-/-	93.6/94. 6/92. 5/-/-
					CN vs. sMCI	-/-/-/-/-	98.4/98. 4/98.3/-/-
					CN vs. AD	-/-/-/-/-	98.8/98.7/98.7/-/-
Pan et al. (2019)	Maximum mean square error (mMsE) of 7 SVMs + majority voting of 3 SVMs	10 times repeated 10-CV (-)	–	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. AD	91.9/-/92.8/-/95.9	-/-/-/-/-
					CN vs. MCI	83.2/-/84.2/-/88.9	-/-/-/-/-
					sMCI vs. pMCI	72.3/-/73.3/-/71.7	-/-/-/-/-
Ahmed et al. (2019b)	Stacking with SoftMax meta classifier	80:20 for train:test (I/E)	Adam optimizer, 20 epochs	No data and code available, no SST, **CI**, small dataset, **I/E validation**, no SEN, no BA, no GTV	CN vs. AD on ADNI	-/-/-/-/-	85.6/85.5/85.5/85.5/-
					CN vs. AD on GARD	-/-/-/-/-	90.1/89.9/90.0/90.0/-
Lei et al. (2019)	Averaging	10-CV (-)	–	No data and code available, no SST, **CI**, small dataset, no validation, no SEN, no BA, no GTV	Mean square error of MMSE = 2.323 ± 0.279 Mean square error of ADAS-Cog = 4.595 ± 0.446		

Table 5 (continued)

Ref. (Year)	Ensemble technique	CV	Hyp. Opt	Reproducibility	Target	CV results	Testing results
Dimitriadis and Liparas (2018)	Majority voting (SES)	Repeated 10-CV (E)	–	No data and code available, no SST, no CI, small dataset, **E validation**, no SEN, no BA, no GTV	CN vs. sMCI vs. pMCI vs. AD	-/-/-/-/-	61.9/60.2/61.9/-/60.5
Nanni et al. (2018)	Static classifier selection (SES)	Split 80:20 for train:test + stratified 4-CV on train (I)	–	**Data and code available**, no SST, no CI, small dataset, **I validation**, no SEN, no BA, no GTV	CN vs. sMCI vs. pMCI vs. AD	-/-/-/-/-	52.9/-/-/-/79.6
Lu et al. (2018a)	Majority voting (SES)	10 times repeated 10-CV (-)	–	No data and code available, no SST, **CI**, small dataset, no validation, no SEN, no BA, no GTV	CN vs. AD sMCI vs. pMCI	-/-/-/-/- -/-/-/-/-	93.6/-/91.5/-/- 81.6/-/73.3/-/-
Jin and Deng (2018)	Boosting decision tree ensemble	10-CV (-)	Grid search	No data and code available, no SST, **CI**, small dataset, no validation, no SEN, no BA, no GTV	CN vs. sMCI vs. pMCI vs. AD	-/-/-/-/-	56.3/-/-/-/-

Table 5 (continued)

Ref. (Year)	Ensemble technique	CV	Hyp. Opt	Reproducibility	Target	CV results	Testing results
Armañanzas et al. (2017)	- (SES)	Inner 5-CV (-)	–	**Data available** and code not available, **signed-rank Wilcoxon test-based SST**, small dataset, no validation, **CI**, no SEN, no BA, no GTV	CN vs. AD	97.14/-/-/-/-	-/-/-/-/-

CV results (Accuracy/precision/recall/f-score/AUC); *LR*: Logistic regression; *RFE*: Recursive feature elimination; *DID1*: (i) cortical thickness, (ii) cortical surface area, (iii) cortical curvature, (iv) grey matter density, (v) the volume of the cortical and subcortical structures, (vi) the shape of the hippocampus and (v) hippocampal subfield volumes; *RF*: Random Forest; *BDT*: Bagged decision tree; *NCA*: Neighborhood component analysis; *DID2*: cortical thickness and subcortical volumes, hippocampal subfields; *FS*: Fisher score,; *MI*: Mutual Information; SGOD: Medical history, Hachinski ischemic score, cerebrovascular disease, Unified Parkinson's Disease Rating Scale, Neuropsychiatric Inventory Questionnaire, Geriatric Depression Scale, Functional Activities Questionnaire; *DNN*: Artificial neural network; *NFDC*: National fMRI Data Center maintained at Dartmouth College (Hanover, NH, USA); *MNI*: Montreal Neurological Institute; RE: Random feature elimination; *BO*: Bayesian optimization; *GARD*: Gwangju Alzheimer's and Related Dementia, Gwangju, South Korea; TVPLH: Three view patches from left HC; TVPRH: Three view patches from right HC; TVPLHRH: Three view patches from merged HCs; CSTGL: Correntropy spatial–temporal group sparse model; FS: Feature selection

the entire dataset and then divided it into training and testing sets. This is the wrong way to implement the ML pipeline because it results in a data leakage problem. In addition, the final model was based only on three features (i.e., Cystatin C, MMP10, and tau), which are not medically relevant. The same data splitting procedure was followed by Muhammed and Thiyagarajan (2021), where data are split into train/test after the preprocessing step. Nanni et al. (2018) did not specify the exact time of splitting the dataset into train and test. Wang et al. (2019c) did not determine if data preprocessing steps were done before or after the tenfold CV.

Other studies applied good practices in the ML pipeline (An et al. 2019). Armañanzas et al. (2017) selected only elderly patients to avoid bias as a result of the age difference. Dimitriadis et al. (2018) divided the dataset from the beginning (i.e., before preprocessing) as training and testing sets and kept the test set untouched for external validation of the trained model. Note that the training and testing data are collected randomly from the same population. Yao et al. (2018) used a real dataset of 160 samples and a simulated dataset of 340 samples for testing, but a small dataset of 240 samples trained the model. Ahmed et al. (2019b) trained the model using the ADNI dataset and tested it using a separate dataset from Korea (GARD) with different distributions for MRI data. ADNI participants were recruited at 57 sites in the United States and Canada with ages between 55 and 90, whereas GARD participants' ages range from 49 to 87. In ADNI, the selection was from diverse races, ethnicities, and age groups. On the other hand, GARD was collected from different races and ethnicities.

A few studies reported the specific hyperparameter optimization methodology applied to optimize base classifiers (Muhammed Niyas and Thiyagarajan 2021; Jin and Deng 2018; Choi and Lee 2020). No studies used multi-objective optimization techniques for minimizing the number of base classifiers while maximizing performance. In addition, no study used AutoML techniques to find optimally and simultaneously the best type of base classifiers, base classifier hyperparameters, number of base classifiers, type of meta-classifier, meta classifier hyperparameters, etc. Even if optimizing ensemble models is a very complex task, all studies in Table 5 used naïve methods for optimization in comparison with those considered in other medical fields (e.g., Singh et al., (Singh and Singh 2020) used the multi-objective optimization NSGA-II technique for optimizing a stacking ensemble model for diabetes prediction).

Many ensemble studies were based on a single modality such as MRI (Li et al. 2021; Lei et al. 2019; Choi and Lee 2020; Wang et al. 2019c; Ahmed et al. 2019b), PET (Pan et al. 2019), and CSF (Syed et al. 2020). Other studies fused several modalities: MRI + PET (Lu et al. 2018a); MRI, MMSE, age, CSF (Dimitriadis and Liparas 2018; Nanni et al. 2018; Jin and Deng 2018); MRI, PET, CSF, CS (age, sex, education) (Muhammed Niyas and Thiyagarajan 2021); Medical history, Hachinski ischemic score, cerebrovascular disease, neuropsychiatric inventory questionnaire, geriatric depression scale, functional activities questionnaire (An et al. 2019). None of these studies discussed the data balancing and missing values problems. Most studies have ignored most reproducibility metrics. For example, none of the studies statistically did bias and imbalance analysis of the selected datasets. No studies checked the correctness of the ground truth of the used datasets. Only two studies used a suitable statistical significance test to measure and compare model performance (Armañanzas et al. 2017; Syed et al. 2020), and only a few studies (Lu et al. 2018a; Lei et al. 2019; Armañanzas et al. 2017; Jin and Deng 2018; Ahmed et al. 2019b) reported the confidence intervals of the results.

Finally, although the surveyed ensemble models have proven their superiority in building computer-aided diagnosis models, important issues are disregarded and require

further research: (1) considering the fusion of heterogeneous modalities including imaging, neuropsychological and clinical data; (2) considering more seriously the hyperparameter optimization of base classifiers; (3) improving model accuracy and diversity by considering either heterogeneous or optimized homogeneous models; (4) exploring the role of dynamic selections, i.e., dynamic classifier selection (DCS) and dynamic ensemble selection (DES), to improve ensemble interpretability and performance; (5) focusing on the role of multimodal data fusion for building a comprehensive model; and (6) exploring the role of AutoML algorithms to optimize the full pipeline from data preprocessing up to the model selection and hyperparameter tuning.

5.1.4 Time-series data-based models

This section investigates the role of time-series data analysis in enhancing model performance. As seen in previous sections, most current AD studies are based on cross-sectional data and on classic supervised ML techniques like SVM and RF (Abuhmed et al. 2021). All these studies were based on the baseline visit data to diagnose AD or to predict its progression after several years (see Tables 3 and 4). Nevertheless, AD progresses over time, so its status at one time is not independent of the status in a previous or next time. Extracted features in ML and DL models based on single visit data have many limitations (Cui and Liu 2019). Because brain changes in AD patients happened gradually over time (Hong et al. 2019), time-series data analysis techniques are more accurate and medically intuitive for analyzing longitudinal data of chronic diseases like AD (see Table 6). Unfortunately, most research has not considered AD data's temporal/sequential nature except for a few studies that utilized hidden Markov models (Williams et al. 2020). DL models can analyze multivariate longitudinal data to learn deep and more relevant representations of the correlation among heterogeneous modalities and the correlation among temporal values. These models are expected to discover unknown and more valuable patterns in the data, which is critical to personalized medicine for AD (El-Sappagh et al. 2020; Li et al. 2020). However, the most significant challenges influencing time-series data analysis in AD are the number of time steps and the amount of missing data. For example, even though ADNI is the largest dataset in the AD domain, it has very few time steps, regularly collected every six months, including missing values. However, DL models such as CNN and RNN (e.g., GRU and LSTM) need datasets with a large number of time steps to learn deep features (Abuhmed et al. 2021), where CNN can learn special features and RNN can extract longitudinal features. Ghazi et al. (2019) proposed a novel LSTM-based algorithm that handles incomplete longitudinal data (up to 74% missing values). Cui et al. (2019) used CNN to extract spatial features from MRI images in a sequence, with bidirectional GRU layers cascaded on the outputs of CNN at multiple time points for extracting the longitudinal features. Then the spatial and longitudinal features are combined to make the prediction. The authors reported an improved performance by adding more time steps, and the models became more robust and stable. Jie et al. (2017) proposed a temporally constrained group sparse linear regression model (e.g., LASSO), which counts for sparse data in longitudinal AD time series. The authors proposed a cost function with two regularization parameters: (1) *group regularization* term to put together weights of the same brain region across different time points; and (2) *smoothness regularization* to reflect the slight changes in data at adjacent time points.

Regarding robustness requirements, as discussed in previous sections, many of the proposed ML pipelines did not handle model validation correctly (Cui and Liu 2019; Dua

Table 6 Review of studies regarding time-series data

Ref. (Year)	Data set	Subjects	TS	Modality (Expert?)	Task (Purpose)	Data preparation	Feature eng	Missing values
Zhu et al. (2021)	ADNI	sMCI (81), pMCI (70)	5	MRI (YES)	AD progression (classification)	AAL-template registration, 7 volume features extraction	Extract partial sequences, extract longitudinal feature representations	–
Er and Goularas (2021)	ADNI	sMCI (169), pMCI (125)	2	sMRI, age, gender, TIV (NO)	AD progression (classification)	Normalized volumes of baseline and M12 MRI, normalized 3D-Jacobian determinant volume, volumes dilation, TIV	FS by SVM-RFE	–
Li et al. (2020)	ADNI	CN (174), MCI (99), AD (116)	61	4D fMRI (NO)	AD detection (Classification)	3D slice timing, head motion correction, normalizing, smoothing, band-pass filtering	Downsampling for imbalanced data	–
Dua et al. (2020)	OASIS-1 OASIS 2	OASIS-1 (416), OASIS 2 (150)	–	MRI (NO)	AD detection (Classification)	GM maps extraction for cross-sectional MRI from OASIS-1 to CNN + missing values imputation for 9 longitudinal features in OASIS-2 for RNNs	–	Mean and median for OASIS-2

Table 6 (continued)

Ref. (Year)	Data set	Subjects	TS	Modality (Expert?)	Task (Purpose)	Data preparation	Feature eng	Missing values
Lei (2020)	ADNI	805 subjects	6	MRI + 4 CSs (YES)	AD progression (regression)	MRI skull strip, segmentation (GM, WM, CSF), registration to Jacob atlas, 93 ROI volumes normalization	FS by CRJL and feature encoding by DPN	CSs missing by regression model
Wang et al. (2019d)	ADNI	CN (48), MCI (95), AD (31)	3	rs-fMRI (NO)	AD progression (classification)	Slice timing correction, head motion estimation, bandpass filtering, regression of nuisance covariates, skull stripping, smoothing,	(1) Extract mean time series of 116 ROI and create BOLD, (2) fMRI time series partitioning in overlapping sliding windows, and (3) extract spatially dependent and temporal dependent patterns	–
Hong et al. (2019)	ADNI	1105 subjects: timesteps of CN (900), MCI (900), AD (900)	20	MRI (YES)	AD progression (classification)	MRI skull strip, registration, segmentation, normalization, smoothing	Data interpolation, normalization [0, 1], serialization, and time step preprocess	Customized method: LI, average, and repeating
Cui and Liu (2019)	ADNI	CN (229), MCI (403), AD (198)	5	MRI (NO)	AD progression (classification)	N3 normalization, skull-stripping, cerebellum removal, segmentation of GM, WM, and CSF	CNN for spatial feature extraction, and RNN for longitudinal feature extraction	RNN masking

Table 6 (continued)

Ref. (Year)	Data set	Subjects	TS	Modality (Expert?)	Task (Purpose)	Data preparation	Feature eng	Missing values
Mehdipour Ghazi, et al. (2019)	ADNI	742 subjects	11	MRI (YES)	AD progression (classification)	Normalization by ICV, outlier filtering	Manual selection of 6 MRI Vs	LSTM
Platero et al. (2019)	ADNI MIRIAD	*ADNI*: CN (114), pMCI (101), sMCI (82), AD (77) *MIRIAD*: CN (23)	4	MRI (NO)	AD progression (classification)	4D registration, 4D segmentation	Manual selection of HC features	–
Wang et al. (2018b)	NACC	5432 subjects	12	78 features, DHP-CGF (NO)	AD progression (classification)	Missing value imputation, normalization, and one-hot encoding for categorical features	Manual group of features selection from the 6 groups in DHPCGF	Mean, median, and mode for numerical, ordinal and nominal
Minhas et al. (2017)	ADNI	*2 years data:* sMCI (65), pMCI (37), *3 years data:* sMCI (65), pMCI (54)	6	MRI, 8 NMs (NO)	AD progression (classification)	Extract volumes, surface area, and cortical thickness MRI images	FS using the two sampled student's t-test, generation of autoregressive parameters by 3 LSLR	LOCF
Lorenzi et al. (2017)	ADNI	*Training set:* sCN (67), cCN (5), pMCI (53), AD (75) *Testing set:* CN (74), cCN (17), sMCI (243), pMCI (106), AD (145)	3	MGB (YES)	AD progression (classification)	Data normalization	Manual selection of MGB list of features	–

Table 6 (continued)

Ref. (Year)	Data set	Subjects	TS	Modality (Expert?)	Task (Purpose)	Data preparation	Feature eng	Missing values
Jie et al. (2017)	ADNI	AD (91), sMCI (98), pMCI (104), CN (152)	4	MRI (NO)	AD progression (classification)	ACPC correction, skull-stripping, Segmentation of GM, WM, CSF, registration, image labeling	Manual selection of GM features	–
Huang et al. (2016)	ADNI	sMCI (61), pMCI (70)	7	MRI (NO)	AD progression (classification)	Remove bias field artifact by N3 method, (spatial) normalization, alignment to template, skull stripping, normalize intensity values	LR-based Voxel selection	Forward filling

Ref. (Year)	Model	CV	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Zhu et al. (2021)	TS-SVM	10-CV (-)	Grid search	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	sMCI vs. pMCI	81.8/-/-/-/-	-/-/-/-/-
Er and Goularas (2021)	Autoencoder-CNN-SVM	10-CV (E)	–	No data and code available, no SST, **CI**, small dataset, **E validation**, no SEN, no BA, no GTV	sMCI vs. pMCI	-/-/-/-/-	87.2/-/92.4/-/-

Table 6 (continued)

Ref. (Year)	Model	CV	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Li et al. (2020)	61 3D CNNs -1 layer LSTM	70:10:20 Train: validation: test (-)	Adam	No data and code available, no SST, **CI**, small dataset, no validation, no SEN, no BA, no GTV	MCI vs. AD	92.1/-/-/-/92	-/-/-/-/-
					CN vs. MCI	88.1/-/-/-/89	-/-/-/-/-
					CN vs. AD	97.4/-/-/-/100	-/-/-/-/-
Dua et al. (2020)	Bagging of 1 CNN, 1 RNN, 1 LSTM	70:30 train: test (-)	Adam	No data and code available, no SST, no CI, **not small dataset**, no validation, no SEN, no BA, no GTV	AD vs. Not AD	92.2/91.9/92.6/91.9/-	-/-/-/-/-
Lei (2020)	Ensemble of SVR	10 times repeated 10-CV (-)	–	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	*M06 predictions*: **MAE** of 1.949, 0.944, 0.124, 4.781, and **R** of 0.687, 0.784, 0.815, 0.690 for MMSE, CDR-SOB, CDR-GLOB, ADAS-Cog. Study reported M12, M18, M24, and M36's results		
Wang et al. (2019d)	STNet (CNN-LSTM)	5-CV (-)	Adam	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. AD	90.3/-/96.7/-/89.8	-/-/-/-/-
					sMCI vs. pMCI	79.4/-/80.9/-/83.2	-/-/-/-/-
					CN vs. MCI vs. AD	71.8/-/-/-/-	-/-/-/-/-
					CN vs. sMCI vs. pMCI vs. AD	66.7/-/-/-/-	-/-/-/-/-
Hong et al. (2019)	LSTM	5-CV	–	No data and code available, no SST, no CI, **not small dataset**, no validation, no SEN, no BA, no GTV	CN vs. AD	-/-/-/-/93.5	-/-/-/-/-
					CN vs. MCI	-/-/-/-/69.7	-/-/-/-/-
					MCI vs. AD	-/-/-/-/79.8	-/-/-/-/-
					CN vs. MCI vs. AD	-/-/-/-/77.7	-/-/-/-/-

Table 6 (continued)

Ref. (Year)	Model	CV	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Cui and Liu 2019)	Cascaded 5 3D CNN + 3 stacked layers of BGRU	5-CV (-)	Adadelta (CNN) + Adam (RNN)	No data and code available, no SST, no CI, **not small dataset**, no validation, no SEN, no BA, no GTV	AD vs. NC sMCI vs. pMCI	91.3/-/86.9/-/93.2 71.7/-/65.3/-/73.0	-/-/-/-/- -/-/-/-/-
Mehdipour Ghazi et al. (2019)	LDA	Random 80:10:10 for train: validate: test (I)	–	No data and code available, **McNemar's based SST**, no CI, small dataset, **I validation**, no SEN, no BA, no GTV	CN vs. MCI CN vs. AD MCI vs. AD CN vs. MCI vs. AD	-/-/-/-/- -/-/-/-/- -/-/-/-/- -/-/-/-/-	-/-/-/-/59.1 -/-/-/-/90.2 -/-/-/-/78.4 -/-/-/-/75.9
Platero et al. (2019)	LME	5000 time repeated: bootstrapping with 75:25 train:test	–	No data and code available, **paired DeLong's test-based SST**, **CI**, small dataset, **E validation**, no SEN, no BA, no GTV	CN vs. AD CN vs. MCI MCI vs. AD	-/-/-/-/94.7 -/-/-/-/80.5 -/-/-/-/72.0	-/-/-/-/- -/-/-/-/- -/-/-/-/-
Wang et al. (2018b)	LSTM with 2 layers	10-CV (-)	Adam	No data and code available, no SST, **CI**, **not small dataset**, no validation, no SEN, no BA, no GTV	MCI vs. AD	99/-/-/-/-	-/-/-/-/-

Table 6 (continued)

Ref. (Year)	Model	CV	Hyper. Opt	Reproducibility	Target	CV results	Testing results
Minhas et al. (2017)	SVM	Stratified 5-CV (E)	–	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	sMCI vs. pMCI (1-year)	-/-/-/-/-	84.3-/-/-/-/88.9
					sMCI vs. pMCI (2-year)	-/-/-/-/-	83.3/-/-/-/88.1
Lorenzi et al. (2017)	Gaussian process	Separate datasets for train and test (E)	–	No data and code available, **ANOVA SST,** no CI, small dataset, **E validation**, no SEN, no BA, no GTV	CV vs. AD	-/-/-/-/-	89/-/83/-/99
					sMCI vs. pMCI	-/-/-/-/-	82/-/65/-/88
					sCN vs. cCN	-/-/-/-/-	83/-/18/-/83
Jie et al. (2017)	tgLASSO	10-ICV (-)	Customized optimization algorithm	No data and code available, **Student paired t-test SST, CI**, small dataset, no validation, no SEN, no BA, no GTV	sMCI vs. pMCI (based on MMSE)	75.7/-/72.9/-/-	-/-/-/-/-
					sMCI vs. pMCI (based on ADAS-Cog)	74.7/-/73.9/-/-	-/-/-/-/-
Huang et al. (2016)	Hierarchical ensemble: Voxel level LR, patch level LR, image level LR	Two NCV loops (10-CV each)	Customized gradient descent algorithm	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	sMCI vs. pMCI	79.4/-/86.5/-/81.2	-/-/-/-/-

Table 6 (continued)

CV results (Accuracy/precision/recall/f-score/AUC); BGRU: Bidirectional gated recurrent unit; *ACPC*: Anterior Commissure (AC)–Posterior Commissure (PC); *NACC*: National Alzheimer's Coordinating Center; *DHPCGF*: Demographics, health history, physical information, elements of the CDR scale, geriatric depression scale, and functional activities questionnaire; *Vs*: Volumes of ventricles, hippocampus, whole brain, fusiform, middle temporal gyrus, and entorhinal cortex; *ICV*: intracranial volume; *LDA*: linear discriminant analysis; *LI*: Linear interpolation; *TS-SVM*: Temporally structured support vector machine; *CSs*: ADAS-Cog, CDR-GLOB, CDR-SOB, and MMSE; *DPN*: Deep polynomial network; *CRJL*: Correntropy regularized joint learning; *NM*: Neuropsychological measure; *LOCF*: Last observation carried forward method; *LSLR*: Least square linear regression; *TIV*: Total intracranial volume is the sum of volumes of GM, WM, and CSF; *MGB*: Volumetric features, glucose metabolism features, brain amyloidosis features, and functional, neuropsychological and cognitive function measures by common scores; *sCN*: Stable CN; *cCN*: Converted CN; *STNet*: Spatial–temporal convolutional-recurrent neural network; *BOLD*: Pair-wise temporal correlation between blood oxygen level-dependent; *LME*: Linear mixed effect; *MIRIAD*: Minimal Interval Resonance Imaging in Alzheimer's Disease; TS: Time steps

et al. 2020; Hong et al. 2019; Li et al. 2020 Wang et al. 2019d, 2018b). Platero et al. (Platero et al. 2019) is the only study that mentioned model robustness and reliability. They performed four experiments: (1) a quality control experiment to assess the robustness of the model on a large set of images; (2) evaluating the reproducibility of the model using a different dataset (MIRIAD); (3) statistical analysis of the atrophy rates of the clinical groups; and (4) experimental verification of the role that longitudinal data of hippocampal markers plays to increase the accuracy of the model. However, the study did not discuss how to handle missing values. In addition, the MIRIAD testing dataset is small. On the other hand, some studies concentrated on other important issues like validation. For example, Er and Goularas (2021) used independent datasets for different phases of the modeling process. They used three different sMRI datasets of 126 cases, 51 cases, and 117 cases for voxel-based morphometric, training, and testing, respectively. However, these datasets are small, and their deep hybrid model of autoencoder-CNN-SVM could overfit. Most studies were based on ADNI, and this dataset has been collected either as longitudinal or cross-sectional. Only a few studies refer explicitly to longitudinal data (Cui and Liu 2019; Dua et al. 2020; Hong et al. 2019). Lorenzi et al. (2017) used separate datasets for training and testing. They made all data preprocessing steps (e.g., normalization) on the training dataset and testing data were prepared according to the biomarkers' distribution of the training set. However, once again these datasets were small and biased. As a result, their model achieved low sensitivity in the classification task of stable CN (sCN) vs. converted CN (cCN). Moreover, the selected cases should be considered as errors in the dataset because medically there is no patient that convert from CN state to AD directly.

Time-series data analysis has been considered in AD literature with two main purposes: (1) for processing longitudinal imaging data of a single visit that is sliced at different time frames where an image like 4D fMRI captures spatial and time-varying information (Li et al. 2020); and (2) for processing longitudinal data collected at different visits (Cui and Liu 2019; Dua et al. 2020; Hong et al. 2019; Zhu et al. 2021; Lei 2020). The latter is the most popular, especially using CNN models to extract spatial features from neuroimages like MRI, which were collected at different time steps. Then the collected features are concatenated, and an RNN model learns the longitudinal features (Cui and Liu 2019; Li et al. 2020). Even if time-series data were short, the analysis of longitudinal changes can bring new knowledge which adds critical value in predicting AD progression. For example, Er and Goularas (2021) used the volumes of only two-time steps (baseline and month 12 visits) from sMRI with a hybrid model of Autoencoder-CNN-SVM to detect AD progression. The model achieved good results compared with other models based only on baseline data. The same model was also used to detect the regions of interest that are statistically significant for MCI-to-AD prediction. About half of the studies in Table 6 used DL models for time-series data analysis as a way to pave the way toward predicting AD progression (Cui and Liu 2019; Dua et al. 2020; Hong et al. 2019; Li et al. 2020; Wang et al. 2019d, 2018b; Er and Goularas 2021; Lei 2020). The rest of studies applied conventional ML models (Mehdipour Ghazi et al. 2019; Jie et al. 2017; Platero et al. 2019; Lorenzi et al. 2017; Zhu et al. 202; Minhas et al. 2017; Huang, et al. 2016). Zhu et al. (2021) proposed the temporally structured SVM (TS-SVM) model to learn MRI time-series data for AD progression detection. The method is based on extracting spatial–temporal features in a partial image sequence. TS-SVM achieved an accuracy of 81.75% in early AD diagnosis using only two follow-up MR scans. Er and Goularas (2021) used SVM based on deep features extracted by an autoencoder-CNN.

Most studies modeled AD progression detection as classification tasks, especially binary classification. Some studies formulated the detection process as a regression task

(Jie et al. 2017; Lei 2020). In (Lei 2020), the authors collected six-time steps of MRI and clinical scores for 805 patients. They predicted AD progression based on four clinical scores, including ADAS-Cog, CDR-GLOB, CDR-SOB, and MMSE, based on MRI longitudinal data. The proposed SVR-based ensemble model used previous clinical scores at MRI time points to predict clinical scores in the next visit. Jie et al. (2017) modeled AD progression detection as a regression task to predict multiple time steps of MMSE and ADAS-Cog clinical measures based on longitudinal MRI data. Most studies were based on MRI images using different strategies for tracking the longitudinal changes between time steps. Some studies tracked changes regarding the whole MRI images (Dua et al. 2020; Minhas et al. 2017)), while others focused on volumes of specific regions of the brain, such as the hippocampus (Platero et al. 2019) and gray matter (Jie et al. 2017), as well as other tracked specific types of measures on all ROI such as volume (Er and Goularas 2021).

Regarding reproducibility issues, it is noticed that none of the studies shared either code or datasets. In addition, only a few studies (Jie et al. 2017; Platero et al. 2019; Lorenzi et al. 2017; Mehdipour Ghazi, et al. 2019) did a statistical analysis of the collected results in terms of a suitable hypothesis testing technique. Only a few studies (Li et al. 2020; Jie et al. 2017; Wang et al. 2018b; Er and Goularas 2021) gave the confidence intervals associated with the reported results. None of the studies paid any attention to data balancing issues or statistical bias and imbalanced data analysis of the selected datasets. Finally, none of the studies checked the correctness of the ground truth in the used datasets.

5.1.5 Multimodal multitask models

The ability to measure dozens of patient features is a crucial step toward personalized medicine for AD (Fisher et al. 2018). Many heterogeneous (bio) markers have shown associations with AD diagnosis and progression detection, including CSF $A\beta_{1\text{-}42}$, CSF tau, CS, FDG-PET, and MRI (Nie et al. 2017). However, AD symptomatology is multimodal because there are correlations and complementary relationships among symptoms of different domains, including physiological, behavioral, and psychological (Wang et al. 2019d). Hence, different modalities are associated with AD information from different perspectives, and their combination can potentially provide a more accurate assessment of disease status and probability of progression (El-Sappagh et al. 2020; Nie et al. 2017). Of course, every single modality has its limitations (Rathore et al. 2017; El-Sappagh et al. 2020). For example, in the case of the sMRI modality, damage to the hippocampus can be determined in AD patients easier than in CN patients. However, this is not an easy task at the early stages of the disease. The low contrast of the anatomical structure in sMRI and the presence of noise or outliers in MRI scans reducer the classification accuracy (Yamanakkanavar et al. 2020). This marks the role of sMRI to be quite blurry in the early disease stages (Alberdi et al. 2016). In addition, brain structural atrophy is not specific to AD only, but it is also related to other diseases as well as normal aging. In addition, recent studies reported that AD is associated with the gray and white matter atrophy. Moreover, changes in brain regions connectivity measured by fMRI precede this structural atrophy (Bai et al. 2009). However, fMRI's identified disruption of functional connectivity cannot be associated with a specific disease (Forouzannezhad, et al. 2018). Fortunately, some of these limitations can be overcome by integrating the MRI modality with other (bio)markers (Lu et al. 2018a; Muhammed Niyas and Thiyagarajan 2021; An et al. 2019). Some studies concluded that sMRI and CSF provide independent biomarkers, and their combination improves AD discrimination (Vemuri and Jr 2010). Most studies in the literature achieved poor performance

for MCI vs. AD determination, probably because the MCI biomarkers are similar to AD. Therefore, multimodal classification had better diagnostic power than single modalities (Lazli and Boukadoum 1894). Accordingly, to be efficient, accurate, reliable, and medically intuitive, AD early diagnosis and progression detection should be supported by multimodal data. Considering all kinds of symptoms and detecting even the smallest changes in the combination of them from the very beginning is the only way to differentiate them from other diseases (Alberdi et al. 2016).

In this section, we discuss the role of multimodality in enhancing model performance. There are two main types of multimodality (El-Sappagh et al. 2020; Ebrahimighahnavieh et al. 2020; Jo et al. 2019): (1) multimodal neuroimaging, which combines different neuroimages to identify structural and molecular/functional biomarkers; and (2) global multimodality, where neuroimaging and non-neuroimaging biomarkers are integrated. There are many methodologies for the fusion of multimodalities, including early fusion (feature concatenation), intermediate fusion, and late or decision fusion. If the collected multimodal data are time series, then it is possible to detect the temporal dependency between different time steps within a single feature and jointly between different features. Studies in the literature modeled AD diagnosis and progression detection with many paradigms, including single modal-single task, single multimodal task, single modal-multitask, and multimodal multitask (MM) (El-Sappagh et al. 2020). The MM is the most medically robust and intuitive method because different modalities are chronically fused and analyzed, and clinical variables are predicted (Abuhmed et al. 2021). This results in an improvement in the physician's level of trust. MM models deal with features from multiple sources, and multiple tasks are jointly optimized. In the rest of this section, we briefly review the literature on AD for multimodal and multitasked studies (see Table 7).

Most studies used multimodal data of MRI and PET. Nevertheless, each study achieved a different degree of success in predicting AD progression or diagnosing the disease. In practice, data from multiple modalities are partially available, and selecting relevant data based on the knowledge of medical experts can improve the performance of the model, as investigated by Ljubic et al. (2020). Other studies addressed the challenge of learning from incomplete modalities (Liu et al. 2021b). The authors proposed an autoencoder-based multiview framework to complete the missing data in the kernel space by considering the association between multiple views and the structural information of the data. Shen et al. (Shen et al. 2020) faced the lack of training data by using a sparse FS method that jointly exploits predictor and auxiliary data. Some studies deal with multimodal time-series data (El-Sappagh 2021; El-Sappagh et al. 2020; Tabarestani et al. 2019; Brand et al. 2020; Lee et al. 2019; Wang et al. 2019e), while others considered only baseline visit data (Fang et al. 2020; Liu et al. 2021b; Shen et al. 2020; Zhang et al. 2020; Forouzannezhad et al. 2019; El-Gamal et al. 2020; Bi et al. 2020; Lei et al. 2020). Most studies were based on the ADNI dataset, but only (Lee et al. 2019; Qiu et al. 2018) confirmed using cross-sectional data. Almost all studies used MRI as one of the modalities in combination with other modalities. The number of modalities varies from 1 modality (Liu et al. 2016), 2 modalities (Fang et al. 2020; Liu et al. 2021b; Liu et al. 2018b; Liu et al. 2018c; Wang et al. 2019e; El-Gamal et al. 2020; Bi et al. 2020; Lei et al. 2020; Altaf et al. 2018; Lu et al. 2018b), 3 modalities (Shen, et al. 2020; Forouzannezhad et al. 2019; Qiu et al. 2018; Peng et al. 2019; Zhang et al. 2019b; El-Sappagh et al. 2022a), 4 modalities (El-Sappagh 2021; Tabarestani et al. 2019; Brand et al. 2020; Lee et al. 2019; Zhang et al. 2020; El-Sappagh et al. 2022b), 5 modalities (Nie et al. 2017), to 12 modalities (El-Sappagh et al. 2020). These modalities were combined with early fusion, intermediate fusion, or late fusion. In addition, models were trained and tunned with different validation techniques: CV (El-Sappagh 2021; Liu

Table 7 Review of studies regarding multimodal multitask problems

Ref. (Year)	Data set	Subjects	Modalities (fusion)	Tasks (purpose)	Data preprocessing	Feature eng	Modals	Tasks
Zhang et al. (2020)	ADNI	CN (40), MCI (42), AD (38)	sMRI, fMRI, PET, DTI (Early fusion)	AD detection (C)	*MRI*: TPM-based segmentation in WM and GM, DARTEL-based registration, normalization; *PET*: 96 images combine, registration, normalization	Extraction of volumes from 90 ROIs of MRIs and average intensity values of 90 ROIs of PETs	4	1
Liu et al. (2021b)	ADNI	CN (90), MCI (185), AD (85)	MRI, FDG-PET (Early fusion)	AD detection (C)	*MRI*: ACPC correction, intensity correction, brain extraction, cerebellum removal, tissues segmentation, registration, ROI labeling; *PET*: MR images alignment	*MRI*: GM extraction for each of 93 ROIs; *PET*: Calculate average PET intensity value of each of the 93 ROIs	2	1

Table 7 (continued)

Ref. (Year)	Data set	Subjects	Modalities (fusion)	Tasks (purpose)	Data preprocessing	Feature eng	Modals	Tasks
El-Sappagh (2021)	ADNI	CN (249), sMCI (363), pMCI (106), AD (318)	Comorbidity, medications, CSs, demographic (Early fusion)	AD progression (C)	Medication encoding by ATC ontology, encoding of disease names, KNN-based missing values, normalization, time series features (average values), data balancing	–	4	1
Shen, et al. (2020)	ADNI	*Dataset1-Dataset2*: CN (99–51), sMCI (59–30),pMCI (55–31), AD (93–50)	MRI, PET, age (Early fusion)	AD progression (C)	Spatial distortion correction, skull-stripping, cerebellum removal, MRI-segmenting to GM, WM, CSF, AAL-registration for 90 regions. PET alignment to MRI	[MRI's 90 Gy matter volumes + PET's 90 mean intensities + age] then FS by	3	1
Brand et al. (2020)	ADNI	CN (143), MCI (190), AD (79)	MRI, SNP genotypes, RAVLTs (Early fusion)	AD progression (C + R)	MRI's standard pipeline of ADNI, all data normalization	GM mean modulated measures of 90 ROIs	4	1

Table 7 (continued)

Ref. (Year)	Data set	Subjects	Modalities (fusion)	Tasks (purpose)	Data preprocessing	Feature eng	Modals	Tasks
El-Sappagh et al. (2020)	ADNI	CN (419), sMCI (473), pMCI (305), AD (339)	MRI, PET, CSs, NB, AD, BN of 7 types (Late fusion)	AD progression (C)	MRI and PET by standard ADNI pipeline, missing values by KNN, normalization, feature reduction by PCA	–	12	5
Forouzannezhad, et al. (2019)	ADNI	CN (248), EMCI (296), LMCI (193), AD (159)	AV-45 PET, MRI, and Demographic information (Early fusion)	AD progression (C)	*MRI*: skull-striping, segmentation, demarcation of 45 ROIs; *PET*: Registration, standardization,	*MRI*: extraction of mean intensity, volume, intensity SD, 9 MV; *PET*: mean intensity of 45 and 68 ROIs + *fusion* + *RF-RFE feature selection*	3	1
El-Gamal et al. (2020)	ADNI	CN (19), MCI (62)	sMRI, ^{11}C PiB-PET (Late fusion)	AD detection (C)	*sMRI:* skull stripping, standardization, brain labeling, *PET*: standardization, wavelet denoising, brain labeling	Feature extraction from labeled ROI regions	2	1

Table 7 (continued)

Ref. (Year)	Data set	Subjects	Modalities (fusion)	Tasks (purpose)	Data preprocessing	Feature eng	Modals	Tasks
Bi et al. (2020)	ADNI	*Training*: CN (35), AD (37) *Testing 1*: EMCI (37), CN (36), *Testing 2*: PD (55), CN (49)	rs-fMRI, SNP (Early fusion)	AD detection (C)	*fMRI*: AAL-based segmentation in 90 ROI, *SNP*: gene grouping, creating digital gene sequence (36 SNPs)	Brain region – gene Pearson correlation matrix	2	1
Fang, et al. (2020)	ADNI	CN (251), EMCI (297), LMCI (196), AD (162)	sMRI, F18-AV45 PET (Early fusion)	AD progression (C)	*MRI*: registration, skull striping, segmentation, and delineating cortical and subcortical regions; *PET*: registration, get mean SUV of 68 ROIs, normalize SUVs	*MRI*: cortical thickness, surface area, cortical volume of 68 ROIs; *PET*: normalized SUVs	2	1
Tabarestani et al. (2019)	ADNI	CN (415), MCI (864), AD (336)	MRI, PET, CSF, CSs, demographic (Late fusion)	AD progression (R)	Standard ADNI pipeline for image processing	MRI: extract features from 7 regions VHWEF; PET: extract SUVRs; 3 CSF features	4	4

Table 7 (continued)

Ref. (Year)	Data set	Subjects	Modalities (fusion)	Tasks (purpose)	Data preprocessing	Feature eng	Modals	Tasks
Lei et al. (2020)	ADNI	CN (44), SMC (44), EMCI (44), LMCI (38)	rs-fMRI, DTI (Early fusion)	AD progression (C)	*fMRI*: Corrections, AAL-template based registration, smoothing, regression covariate, time filtering; DTI: brain tissue extraction, correction, calculate diffusion tensor, fMRI co-registration, AAL-alignment	Brain networks construction by a self-calibration method; FS by joint non- convex multi-task learning model	2	1
Peng et al. (2019)	ADNI	Dataset1: CN (47), MCI (93), AD (49) Dataset2: CN (90), MCI (185), AD (85)	MRI, FDG-PET, Genetics (Late fusion)	AD detection (C)	MRI-PET: ACPC + intensity inhomogeneity corrections, skull stripping, cerebellum removal, MR-segmentation, registration	MRI: GM volumes of 93 ROIs, PET: mean intensity of 93 ROI, Genetics: SNPs; MKL-based feature selection	3	1
Zhang et al. (2019b)	ADNI	CN (101), MCI (200), AD (91)	MRI-PET-CNF (Late fusion)	AD detection (C)	AAL-based segmentation into 116 ROIs	–	3	1

Table 7 (continued)

Ref. (Year)	Data set	Subjects	Modalities (fusion)	Tasks (purpose)	Data preprocessing	Feature eng	Modals	Tasks
Lee et al. (2019)	ADNI	CN (415), sMCI (558), pMCI (307), AD (338)	MRI, demographic, CSF, CSs (Late fusion)	AD progression (C)	Standard ADNI MRI preparation pipeline	*MRI*: manual selection of hippocampal volume, entorhinal cortical thickness	4	1
Wang et al. (2019e)	ADNI	AD (90), pMCI (104), sMCI (98), CN (152)	MRI (-)	AD progression (R)	ACPC correction; N3 intensity correction; skull stripping; segmentation of GM, WM, CSF; 93-ROIs registration;	93-ROIs GM volumes extraction	2	2
Qiu et al. (2018)	NACC	CN (303), MCI (83)	MRI, MMSEs, LM tests (Late fusion)	MCI detection (C)	Select 3 slices from 20 slices, normalization [0,1]	3 MMSE scores (3MS), 4 LM tests (4LM), MRI features	3	1
Liu et al. (2018b)	ADNI	CN (100), sMCI (128), pMCI (76), AD (93)	FDG-PET, MRI (Intermediate fusion)	AD progression (C)	*MRI*: ACPC correction, skull-stripping, cerebellum removal, affine registration, resampling; *PET*: registration, normalization, resampling	–	2	1

Table 7 (continued)

Ref. (Year)	Data set	Subjects	Modalities (fusion)	Tasks (purpose)	Data preprocessing	Feature eng	Modals	Tasks
Altaf et al. (2018)	ADNI	CN (90), MCI (105), AD (92)	MRI, clinical data (FAQ, NPI, GDS) (Early fusion)	AD detection (C)	MRI: segmentation in GM, WM, CSF + normalization	Texture-based features extraction from MRI using GLCM	2	1
Liu et al. (2016)	ADNI	CN (225), MCI (390), AD (173)	MRI (-)	AD progression (R)	Motion correction and averaging, intensity correction, normalization, skull stripping, subcortical and WM segmentation, surface deformation and smooth	Thickness average and standard deviation, surface area, 48 cortical and 44 subcortical volumes; handling missing by average	1	5
Lu et al. (2018b)	ADNI	sCN (360), sMCI (409), pCN (18), pMI (217), AD (238)	MRI, FDG-PET (Late fusion)	AD progression (C)	MRI: ROI-wise segmentation, patch-wise segmentation; FDG-PET: MRI-based co-registration, patch-wise segmentation	MRI's patch volume and PET's mean intensity feature extraction	2	1

Table 7 (continued)

Ref. (Year)	Data set	Subjects	Modalities (fusion)	Tasks (purpose)	Data preprocessing	Feature eng	Modals	Tasks
Liu et al. (2018c)	ADNI	ADNI-1 (797), ADNI-2 (599)	MRI, demographic (Late fusion)	AD progression (C + R)	*MRI*: ACPC correction, N3-based intensity inhomogeneity correction, skull stripping, remove cerebellum	–	2	5
Nie et al. (2017)	ADNI	CN (229), MCI (401), AD (188)	MRI, CSF, FDG-PET, META, PROT (Early fusion)	AD progression (C)	Standard ADNI pipeline for image processing	Missing values by MF	5	2

Ref. (Year)	Model	Time series	CV (validation)	H. Opt	Reproducibility	Target	CV results	Testing results
Zhang et al. (2020)	MCSVM	NO	30 times repeated 5-CV (I)	Manual	No data and code available, **FRT and HPT based SST**, **CI**, small dataset, no validation, no SEN, no BA, no GTV	MCI vs. AD CN vs. AD `CN vs. MCI CN vs. MCI vs. AD	92.6/-/-/-/- 96.7-/-/-/-/- 83.2-/-/-/-/- 89.9/-/-/-/-	89.6/-/-/-/- 94.6/-/-/-/- 81.0/-/-/-/- 88.5/-/-/-/-
Liu et al. (2021b)	Autoencoder-based AEMVC + SVM	NO	30 times repeated 10-CV (-)	–	No data and code available, no SST, **CI**, small dataset, no validation, no SEN, no BA, no GTV	CN vs. AD	83.9/-/-/-/-	-/-/-/-/-

Table 7 (continued)

Ref. (Year)	Model	Time series	CV (validation)	H. Opt	Reproducibility	Target	CV results	Testing results
El-Sappagh (2021)	Random forest	4-time steps	Stratified 90:10 train:test; 10 times 10-CV (I)	Grid search	No data and code available, **Friedman and post-hoc Nemenyi based SST**, **CI**, small dataset, no validation, no SEN, no BA, no GTV	CN vs. sMCI vs. pMCI vs. AD CN vs. MCI vs. AD	90.9/91.3/91.1/90.9/- 92.1/92.3/92.1/92.0/-	90.2/90.1/90.2/90.1/- 91.2/91.2/91.2/91.2/-
Shen, et al. (2020)	Linear SVM	NO	100 times repeated 10-CV (-)	Grid search	No data and code available, no SST, **CI**, small dataset, no validation, no SEN, no BA, no GTV	sMCI vs. pMCI	78.7/-/77.3/-/77.6	-/-/-/-/-
Brand et al. (2020)	JMLRC	4-time steps	5 times repeated 5-CV (-)	Grid search	**Code available**, but no data, no SST, **CI**, small dataset, no validation, no SEN, no BA, no GTV	CN vs. AD (balanced accuracy)	58.4/-/-/57.6/-	-/-/-/-/-
El-Sappagh et al. (2020)	Stacked CNN-BiLSTM	15-time steps	10 times repeated Stratified 60:20:20 train:validate: test (I)	Adam	No data and code available, **Student t-test SST**, **CI**, **not small dataset**, **I validation**, no SEN, no BA, no GTV	CN vs. MCI vs. AD	-/-/-/-/-	92.6/-/98.4/ 92.6/-

Table 7 (continued)

Ref. (Year)	Model	Time series	CV (validation)	H. Opt	Reproducibility	Target	CV results	Testing results
Forouzannezhad, et al. (2019)	Gaussian process	NO	80:20 train:test (I)	–	No data and code available, **Student's t-test based SST**, no CI, small dataset, **I validation**, no SEN, no BA, no GTV	CN vs. EMCI	-/-/-/-	78.8/-/81.3/-/-
						CN vs. LMCI	-/-/-/-	79.8/-/70.2/-/-
						CN vs. AD	-/-/-/-	94.7/-/92.3/-/-
						EMCI vs. LMCI	-/-/-/-	73.2/-/81.2/-/-
						EMCI vs. AD	-/-/-/-	88.1/-/92.8/-/-
						LMCI vs. AD	-/-/-/-	77.1/-/79.9/-/-
El-Gamal et al. (2020)	For each modality, 116 pSVMs for each region then global SVM	NO	LOSO (-)	–	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. MCI (averaged)	97.5/-/96.8/-/-	-/-/-/-/-
Bi et al. (2020)	Cluster evolutionary random forest	NO	50:30:20 Train: validate: test (E)	Manual	No data and code available, no SST, no CI, small dataset, **E validation**, no SEN, no BA, no GTV	CN vs. AD	81.0/-/-/-/-	-/-/-/-/-
						CN vs. EMCI	-/-/-/-/-	80.0/-/-/-/-
						CN vs. PD	-/-/-/-/-	83.0/-/-/-/-
Fang et al. (2020)	GDCA algorithm	NO	80:10:10 train: validate: test (-)	–	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. EMCI	79.3/-/79.3/80.7 /79.6	-/-/-/-/-
						EMCI vs. LMCI	83.3/-/82.4/77.8 /89.5	-/-/-/-/-
Tabarestani et al. (2019)	Gradient Boosting machine	6-time steps	NCV: 10 times repeated 70: 30 train:test + 5-CV (I)	Grid search	**Code available**, but no data, no SST, no CI, **not small dataset**, **I validation**, no SEN, no BA, no GTV	RMSE for MMSE: Time step 1 (1.62), Time step 6 (1.78), Time step 12 (2.24),Time step 24 (2.38), Time step 36 (2.28), Time step 48 (2.19)		

Table 7 (continued)

Ref. (Year)	Model	Time series	CV (validation)	H. Opt	Reproducibility	Target	CV results	Testing results
Lei et al. (2020)	SVM	NO	LOOCV (-)	Grid search	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. SMC	82.9/-/88.6/-/89.8	-/-/-/-/-
						CN vs. EMCI	85.2/-/86.4/-/93.5	-/-/-/-/-
						CN vs. LMCI	87.8/-/84.2/-/98.9	-/-/-/-/-
						SMC vs. EMCI	84.1/-/81.8/-/91.2	-/-/-/-/-
						SMC vs. LMCI	90.2/-/89.5/-/96.7	-/-/-/-/-
						EMCI vs. LMCI	81.7/-/78.9/-/92.1	-/-/-/-/-
Peng et al. (2019)	SVM	NO	10 times repeated 10-CV (E)	Grid search	No data and code available, **Student t-test based SST**, no CI, small dataset, **E validation**, no SEN, no BA, no GTV	CN vs. AD	96.1/-/97.3/-/99.2	-/-/-/-/-
						CN vs. MCI	80.3/-/85.6/-/81.1	-/-/-/-/-
						MCI vs. AD	76.9/-/65.9/-/80.8	-/-/-/-/-
Zhang et al. (2019b)	2 CNNs: VGGNet-19 (MRI) + VGG-Net-19 (PET)	NO	90:5:5 Train: validate: test (-)	–	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. AD	98.5/-/96.6/-/98.6	94.5/-/-/-/-
						CN vs. MCI	85.7/-/90.1/-/88.2	80.2/-/-/-/-
						MCI vs. AD	88.2/-/97.4/-/88.0	80.1/-/-/-/-
Lee et al. (2019)	Ensemble of 4 LSTM models	5-time steps	10 times repeated 5-CV (-)	–	No data and code available, **paired t-test based SST**, **CI**, **not small dataset**, no validation, no SEN, no BA, no GTV	sMCI vs. pMCI	81/-/-/-/86	-/-/-/-/-

Table 7 (continued)

Ref. (Year)	Model	Time series	CV (validation)	H. Opt	Reproducibility	Target	CV results	Testing results
Wang et al. (2019e)	Linear SVM	4-time steps	10 times repeated 10-CV (-)	Line search	No data and code available, **Mc-Nemar test SST**, **CI**, small dataset, no validation, no SEN, no BA, no GTV	RMSE: ADAS (0.76), MMSE (0.79)		
Qiu et al. (2018)	Majority vote: 3 VGG-11 for MRI + MLP for MMSE + MLP for LM	NO	5 times stratified 2:1:1 train: validate: test (I)	SGD	No data and code available, no SST, **CI**, small dataset, **I validation**, no SEN, no BA, no GTV	CN vs. MCI	-/-/-/-/-	90.9/92.6/96.3/94.4/-
Liu et al. (2018b)	Cascaded (27) 3D CNNs – (1) 2D CNN	NO	10-CV (-)	Adadelta	**Code available** but no data, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. AD	93.3/-/92.6/-/95.7	-/-/-/-/-
						CN vs. pMCI	82.9/-/81.1/-/88.4	-/-/-/-/-
						CN vs. sMCI	64.0/-/63.1/-/67.1	-/-/-/-/-
Altaf et al. (2018)	KNN (k = 5)	NO	5-CV (-)	–	No data and code available, no SST, no CI, small dataset, no validation, no SEN, no BA, no GTV	CN vs. MCI vs. AD	79.8/-/92/-/-	-/-/-/-/-
						CN vs. AD	97.8/-/100/-/-	-/-/-/-/-
						MCI vs. AD	85.3/-/75/-/-	-/-/-/-/-
						CN vs. MCI	91.8/-/90/-/-	-/-/-/-/-
Liu et al. (2016)	Multitask sparse group LASSO	NO	50 times repeated random 95:5 train:test then nested 5-CV (I)	––	No data and code available, no SST, **CI**, small dataset, **I validation**, no SEN, no BA, no GTV	*RMSE*: ADAS (6.59), MMSE (2.16), RAVLT total (9.5), RAVLT T30 (3.32), RAVLT RECOG (3.54)		

Table 7 (continued)

Ref. (Year)	Model	Time series	CV (validation)	H. Opt	Reproducibility	Target	CV results	Testing results
Lu et al. (2018b)	MMDNN: 6 DNN + 1 DNN ensemble classifier	NO	10-CV (-)	Adam	No data and code available, no SST, **CI**, **not small dataset**, no validation, no SEN, no BA, no GTV	sMCI vs. pMCI	82.9-/79.7/-/-/-	-/-/-/-/-
						sCN vs. sAD	84.6/-/80.2/-/-	-/-/-/-/-
						sCN vs. (pMCI + sAD)	86.0/-/85.7/-/-	-/-/-/-/-
						sCN vs. (pCN + pMCI + sAD)	86.4/-/86.5/-/-	-/-/-/-/-
Liu et al. (2018c)	CNN: deep multi-task multichannel learning	NO	Train on ADNI1 and test on ADNI2 (E)	SGD	No data and code available, no SST, no CI, small dataset, **E validation**, no SEN, no BA, no GTV	CN vs. sMCI vs. pMCI vs. AD	-/-/-/-/-	51.8/-/-/-/-
						RMSE: CDR (1.7), ADAS11 (6.2), ADAS13 (8,5), MMSE (2.4)		
Nie et al. (2017)	MSMT linear model	NO	10-CV (-)	–	No data and code available, **paired t-test SST**, no CI, small dataset, no validation, no SEN, no BA, no GTV	*MMSE*: R-value (0.8794), nMSE (0.1233). *ADAS-Cog*: R-value (0.9094), nMSE (0.0347)		

Table 7 (continued)

CV results (Accuracy/precision/recall/f-score/AUC); *pSVMs*: Probabilistic SVM; *LOSO*: Leave-one-subject-out; *LMCI*: Late mild cognitive impairment; *SD*: Standard deviation; *MV*: Morphological variables; *CNF*: clinical neuropsychological features (MMSE + CDR); *SNP*: Single nucleotide polymorphism; *PD*: Parkinson's disease; *FAQ*: Functional activities questionnaire; *NPI*: Neuropsychiatric inventory; *GDS*: Geriatric depression scale; *GLCM*: Including gray-level co-occurrence matrix; *CSs*: Cognitive scores; *NB*: Neuropsychological battery, *NP*: Neuropathology; *BN*: Demographics, lab tests, vital signs, CSF data, genetic data, physical examination; *GM*: General model; *AEMVC*: Auto-Encoder based Multi-View missing data Completion framework; *GDCA*: Supervised Gaussian discriminative component analysis algorithm; *SUV*: Standardized uptake values; *3MS*: MMSEORDA (orientation subscale score—time), MMSEORLO (orientation subscale score—place), and NACCMMSE (total MMSE score); *4LM*: LOGIPREV (total score from the previous test administration), LOGIMEM (total number of story units recalled from this current test administration), MEMUNITS (total number of story units recalled), and MEMTIME (time elapsed since first recall to delayed recall); *LM*: Logical memory; *SGD*: stochastic gradient descent algorithm; *MCSVM*: Multiclass SVM; DARTEL: Diffeomorphic anatomical registration through exponentiated lie algebra; C: Classification; *FRT*: Friedman ranking test; *HPT*: Holm post-hoc test; *JMLRC*: Joint Multi-Modal Longitudinal Regression and Classification; *SNPs*: Single nucleotide polymorphisms; *R*: Regression; RMSE: Root mean squared error; *RAVLT*: Rey Auditory Verbal Learning Test; sCN: Stable Normal controls; pCN: Progressive CN; *MMDNN*: Multimodal and Multiscale Deep Neural Network; *MTRL*: Multi-task Relationship Learning; MF: Matrix factorization; *PROT*: Proteomics; META: 3 features demographics, 7 features genetics, 23 features cognitive scores, 18 features lab tests; MSMT: Multisource multitask; *nMSE*: Normalized mean-squared error; *R-value*: correlation coefficient; VHWEF: Ventricular volume, Hippocampus volume, Whole Brain volume, Entorhinal Cortical thickness, Fusiform, Middle temporal gyrus,and intracranial volume (ICV); *MKL*: Group structured sparsity penalized multiple kernel learning; *SMC*: Significant memory concern

et al. 2021b; Shen et al. 2020; Brand et al. 2020; Lee et al. 2019; Wang et al. 2019e; Zhang et al. 2020; Lei et al. 2020; Peng et al. 2019)), holdout (El-Sappagh et al. 2020; Fang, et al. 2020; Forouzannezhad, et al. 2019; Bi et al. 2020; Zhang et al. 2019b)) and using two independent datasets for training and test (Liu et al. 2018c). These studies split the data after performing preprocessing on the whole data. Forouzannezhad et al. (2019) performed a train/test split, performed FS on the training set alone, and then applied the selected features on the test set.

The multitask studies in the AD domain are mainly categorized into three groups. The first group predicts multiple cognitive scores that are related to AD progression as a regression task (Tabarestani, et al. 2019; Wang et al. 2019e; Liu et al. 2016). The second group predicts multiple cognitive scores with a regression task but also one or more classification tasks as CN vs. MCI vs. AD (El-Sappagh et al. 2020; Liu et al. 2018c). The third group uses multitask learning in data preparation steps as FS (Lei et al. 2020). Building multitask models based on multimodal longitudinal data for predicting AD and related sensitive features like cognitive scores is expected to improve the model performance because it deals with features from many sources, and multiple tasks are related in chronological order (Nie et al. 2017). Zhang et al. (2019b) proposed an AD scoring system by combining MMSE, CDR, MRI data processed with CNN (*MRIcnn*), and PET data processed with CNN (*PETcnn*) : $CDscore = \lambda \times avg(PETcnn + MRIcnn) + (1 - \lambda) \times avg(MMSEscore + CDRscore)$, where $\lambda = (1 + \gamma)/2$ and γ is the Pearson correlation coefficient between the predictions of MRI and PET models. Some studies used time-series data to predict the future state of the patient at a particular time point (El-Sappagh et al. 2020). Other studies predicted the future status of the patient at multiple time points regarding the baseline data of multiple modalities (Tabarestani, et al. 2019; Nie et al. 2017). However, there are no studies that learn multimodal longitudinal data to predict multiple scores at multiple time points.

Regarding reproducibility issues: (1) none of the studies shared the datasets. The reason could be the public availability of ADNI, which is the most popular dataset. However, some studies shared the used patient IDs from the ADNI dataset (El-Sappagh et al. 2020). (2) Only (Tabarestani, et al. 2019; Brand et al. 2020; Liu et al. 2018b) shared the code used to run experiments. (2) The statistical significance of the results is studied by Zhang et al. (2020) using Friedman ranking test and Holm post-hoc test, by El-Sappagh (2021) using Friedman and post-hoc Nemenyi, by (El-Sappagh et al. 2020; Nie et al. 2017; Lee et al. 2019; Forouzannezhad, et al. 2019; Peng et al. 2019) using Student t-test, and by Wang et al. 2019e using Mc-Nemar test. However, no study discussed the reason for selecting these specific tests. (3) A small number of studies reported the confidence intervals which describe the model stability (El-Sappagh 2021; Liu et al. 2021b; Zhang et al. 2020) (4) Most studies were based on small datasets except (El-Sappagh et al. 2020; Tabarestani, et al. 2019; Lee et al. 2019) (5) Many studies reported CV results only without any internal or external validation (Liu et al. 2021b). External validation results have been reported only by ((Bi et al. 2020; Liu et al. 2018c; Peng et al. 2019)). (6) None of the studies verified either the ground truth of the used dataset or the sensitivity of the model to small changes in input data. Moreover, no study performed bias and imbalance data analysis.

5.2 Studies that explicitly considered Trustworthy AI

Although there are several studies on the diagnosis and progression detection models for AD in the literature, there are no notable applications of these models in actual medical settings. We claim this is mainly because only a few studies take care of TAI requirements

in the pipeline, from data collection to model deployment and monitoring (see Table 8). Bruun et al. (Bruun et al. 2019) implemented the PredictND CDSS tool for diagnosing dementia diseases, including AD, based on 779 patients collected from four European memory clinics. The system is implemented under the guidance of domain experts to calculate a disease state index. Baseline data only of demographics, cognitive tests, cerebrospinal fluid biomarkers, and MRI are used. Disease state fingerprint is a tree-like visualization tool used to show the contribution of each biomarker in the final decisions. This tool achieved a statistically significant enhancement in the accuracy of AD diagnosis. Then, a CDSS increased the physician's confidence. Even though this is considered a complete and practical system in AD literature, the study ignores most TAI requirements, especially explainability, robustness, and fairness. As a result, no tools are used in other settings. In addition, responsible deployment of AI systems must consider the previously discussed TAI requirements. Unfortunately, in the AD domain, AI has no real application in clinical settings. "Human agency and oversight" and "privacy and data governance" are challenging TAI requirements that remain underexplored and require further research. Regarding "accountability", it is believed that physicians who make final decisions (no matter if with or without the support of AI) are responsible for all medical decisions. Regarding the rest of the TAI requirements, we look for answers to the set of questions asked by the HLEG. More precisely, due to space restrictions, the study selects the most critical questions from the list given in the supplementary material, which is a comprehensive questionnaire for measuring the degree of compliance with TAI requirements. The study proposes a questionnaire with 123 questions to assess the level of trustworthiness achieved by an AI-based model. These questions are highlighted in bold in Table 8 (see Q1, Q2, etc.). The full details about these questions can be found in the supplementary material. The rest of this section concentrates on three complementary TAI requirements, i.e., transparency, robustness, and fairness. It is worth noting that robustness is related to accuracy and validation. That is the reason why all studies in Table 8 have some contribution to this TAI requirement. In summary, no single study considers all these three requirements. No single real system in the AD domain deeply measures and monitors the TAI requirements.

5.2.1 Transparency

Transparency is mainly related to interpretable models (e.g., DT, LR, and BN), but it is also required by explainable models (e.g., explainable CNN). Most of the existing DL models for AD lack transparency, so it is difficult to explain why and how a model's decision is reached (Zhang et al. 2021).

Regarding interpretable models, Xiao et al. (2021) optimized an interpretable logistic regression model to detect AD progression. The model achieved good validation results and calculated weights for different MRI regions, which could help domain experts understand the logic behind the model. The essential region was the left hippocampus, with a normalized weight of 0.3478. Ding et al. (2018) proposed a hybrid regression model based on the BN technique to study the dynamic and longitudinal relationship among AD predictors over time. The BN model predicted the CDR value as an index for AD severity by identifying and visualizing the time-varying relationships between combinations of significant indicators such as MMSE, logical memory recall, MRI's GM and cerebrospinal volumes, PiB-PET's active voxels, ApoE, and age. The entropy-based FS method was used to select 8 out of 32 features (i.e., 2 demographics, 10 medical history data, 13 blood tests, 4 psychological, 4 MRI and PET). *Post-hoc interpretable models* have been used to endow

Table 8 Review of TAI-based studies in the AD literature

Ref. (Year)	Data set	Subjects	Modality (Expert?)	Task (Purpose)	Data preparation	Feature engineer	Time series	Model (optimize)
Abuhmed et al. (2021)	ADNI	CN (419), sMCI (473), pMCI (140), AD (339)	T1W MRI, PET, NB, CSs, NP (NO)	AD progression (C)	MRI and PET by standard ADNI pipeline; missing values by LSTM masking, normalization, SMOTE balancing	-	YES	Hybrid BiLSTM-RF (Adam)
Sappagh et al. (2021)	ADNI	CN(294), sMCI(254), pMCI(232), AD(268)	11 modalities (YES)	AD progression (C)	*Images*: standard ADNI pipeline; *Other data*: missing values, SMOTE balancing, outliers, data normalization	FS using RF-RFE	NO	Two-layer RF classifiers (Grid search)
Xiao et al. (2021)	ADNI	AD (51), CN (50), pMCI (51), sMCI (45)	T1W MRI (NO)	AD progression (C)	Extract gray matter volume by IPPS steps	FS by generalized elastic net regularization	NO	$L_{1/2}$ Sparse logistic regression (coordinate descent algorithm)
Zhang et al. (2021)	ADNI	ADNI1: 835, ADNI2: 258, ADNI3: 314	T1W sMRI (NO)	AD detection (C)	MRI: 1.5 T of ADNI1 and 3 T of ADNI2&3 preprocessed by ADNI standard pipeline	–	NO	3D CNN: 3D ResAttNet (Adam)
Bass, et al. (2021)	ADNI	*Train*: AD (257), MCI (674), *Test*: pMCI (61), *validate*: pMCI (61)	T1W MRI (NO)	AD progression (C)	N4 bias correction registration, normalization, resizing	MRI Volumes	NO	ICAM: VAE-GAN translation network

Table 8 (continued)

Ref. (Year)	Data set	Subjects	Modality (Expert?)	Task (Purpose)	Data preparation	Feature engineer	Time series	Model (optimize)
Wen et al. (2021)	ADNI	CN (46), MCI (97), AD (46)	DWI + T1W MRI (YES)	AD progression (C)	*T1W MRI*: WM, GM, and CSF segmentation; registration; binarization. DWI MRI: correction, registration. normalization	*T1W MRI*: GM density features from voxels and ROIs; DWI MRI: FA and MD maps	NO	Linear SVM (-)
Bron, et al. (2021)	ADNI HPNDB	*ADNI*: CN (520), sMCI (628), pMCI (231), AD (334) *HPNDB*: CN (139), sMCI (91), pMCI (48), AD (199)	T1W MRI (YES)	AD progression (C)	MRI image preprocessing using Iris pipeline + normalization to zero mean and unit variance	Modulated GM maps	NO	SVM (-)
Zhao et al. (2021)	ADNI	*ADNI set A* (210), *ADNI set B* (603), *ASIS* (48)	T1W 3D sMRI (NO)	AD progression (C)	Segmentation of GM, WM, & CSF; create brain mask, multiply mask with MRI to get initial brain image; registration & cropping	Extract 80 patches of volumetric features	YES	3D mi-GAN-3D DenseNet (Adam)
Bin Bae et al. (2020)	ADNI SNUBH	*ADNI*: CN (195), AD (195) *SNUBH*: CN (195), AD (195)	T1W 3D sMRI (YES)	AD detection (C)	Resampling, extraction of coronal slices around MTL, registration, skull-stripping	30 coronal slices extract for MTL then min–max normalization [0, 1]	NO	CNN: Inception-v4

Table 8 (continued)

Ref. (Year)	Data set	Subjects	Modality (Expert?)	Task (Purpose)	Data preparation	Feature engineer	Time series	Model (optimize)
Georges et al. (2020)	ADNI	LMCI (36), AD (41)	T1W sMRI (NO)	AD detection (C)	Reconstruct left and right cortical hemispheres, segmentation, cortical attributes calculations	Feature selection by FS-Select method	NO	SVM
Zhao et al. (2020)	ADNI	In house (715), ADNI (1228)	T1W MRI (YES)	AD progression (C)	Registration, hippocampus segmentation, normalization, and integration	SVM-RFE based FS from hippocampal radiomic features (IST)	NO	RBF SVM
Jo et al. (2020)	ADNI	CN (66), AD (66), early MCI (97), late MCI (71)	Tau PET (YES)	AD detection (C)	Normalization	–	NO	3D CNN
Lian et al. (2020a)	ADNI	*ADNI-1*: NC (229), sMCI (226), pMCI (167), AD (199), *ADNI-2:* NC (200), sMCI (239), pMCI (38), AD (159)	T1W 3D sMRI (YES)	AD progression (C)	AC-PC correction, intensity correction, skull stripping, cerebellum removing, registration by Colin27	–	NO	CNN (*with-prior* H-FCN) (Adam)
Dyrba, et al. (2020)	ADNI	ADNI-GO/2 (663), ADNI-3 (575), AIBL (606), DELCODE (474)	T1W 3D MRI (YES)	AD detection (C)	Segmentation, normalization, augmentation	Gray matter's 3D volumes	NO	3D CNN (Adam)

Table 8 (continued)

Ref. (Year)	Data set	Subjects	Modality (Expert?)	Task (Purpose)	Data preparation	Feature engineer	Time series	Model (optimize)
Fukunishi et al. (2020)	DMAJ	AD (3,095), Not AD (45,028)	Demog., comorbidity and drugs (YES)	AD progression (C)	Data coding of diseases by ICD-10, drugs by ATC, Data balancing	FS by L0 regularization (select 13 from 611 features)	NO	100 Sparse logistic regression models (-)
Qiu et al. (2020)	ADNI, AIBL, FHS, NACC	ADNI (417), AIBL (382), FHS (102), NACC (582)	T1W MRI, age, sex, MMSE (YES)	AD detection (C)	MRI: image registration, intensity normalization, and volumetric segmentation	–	NO	CNN-MLP hybrid model (Adam)
Biffi et al. (2020)	ADNI	CN (404), AD (322)	T1W MRI (YES)	AD detection (C)	3D left and right hippocampus segmentation	–	NO	LVAE-MLP (Adam)
Kim et al. (2020)	ADNI, SHD	*ADNI*: CN (347), AD (139), *SHD*: CN (68), AD (73)	FDG-PET (YES)	AD detection (C)	Registration, Intensity and spatial normalizations, slice selection (double slices)	Feature extraction using BEGAN	NO	SVM
Adeli et al. (2019)	ADNI	CN (101), MCI (202), AD (93)	MRI, FDG-PET (YES)	AD detection (C)	Standard preprocessing pipeline	FS by $\updownarrow_1$ norm	NO	RFS-LDA (ADMM + grid search)
Oh et al. (2019)	ADNI	CN (230), sMCI (101), pMCI (166), AD (198)	T1W MRI (YES)	AD progression (C)	Realignment, normalization, smooth, segmentation (GM, WM, CSF)	Convolutional AE-based feature extraction,	NO	CAE-ICAE (grid search)

Table 8 (continued)

Ref. (Year)	Data set	Subjects	Modality (Expert?)	Task (Purpose)	Data preparation	Feature engineer	Time series	Model (optimize)
Seo et al. (2019)	ADNI	CN (177), MCI (277), AD (108)	T1W MRI (NO)	AD progression (C)	MRI registration, ICV-standardize, normalization; then feature prescreening by correlation	Non-linear Manifold-based DR (LLE) (2 features)	YES	SVM (-)
Das et al. (2019)	ADNI	Plasma protein (151), CSF and plasma (141)	Plasma protein data + CSF (YES)	AD detection (C)	–	14 proteins + 2 CSF features	NO	2-stages: Rule-based SHIMR then SVM (-)
Babapour Mofrad, et al. (2019)	ADNI	CN (442), MCI (363), AD (1004)	CSF (YES)	AD detection (C)	–	Ab42, tau, and Ptau-181, APOE, sex, age	NO	CART decision tree (-)
Ding et al. (2018)	AIBL	589 subjects	Medical history, T1W MRI, PET, blood test, demog., psychological (YES)	AD detection (C)	Discretization, SMOTE balancing	Entropy-based feature selection	YES	Bayesian Network (-)

Ref. (Year)	CV	Robustness requirement	Fairness requirement	Transparency requirement	Target	CV results	Testing results
Abuhmed et al. (2021)	90:10 train: test then 10-times repeated stratified 10-CV	**Q20** (5,YES,4,NO,YES,LSTM masking), **Q35** (I), **Q32** (LSTM), **Q43** (Single), **Q38** (90:10, 10-CV), **Q31**(data preprocessing), **Q40** (YES), Q56 (YES,), **Q38** (I, YES), **Q56** (Student t-test)	**Q111** (data imbalance, SMOTE)	**Q89** (X feature importance + fuzzy-rules + DT rules: post-hoc model-agnostic), **Q90** (DE, NO), **Q96** (YES), **Q101** (S)	CN vs. MCI vs. AD	-/-/-/-/-	82.6/84.7/ 84.8/84.7/-

Table 8 (continued)

Ref. (Year)	CV	Robustness requirement	Fairness requirement	Transparency requirement	Target	CV results	Testing results
Sappagh et al. (2021)	90:10 train:test, then 5-time repeated 10-CV	**Q20** (11,NO,-,-,-,-), **Q32** (RF), **Q29** (YES), **Q35** (E), **Q43** (Single), **Q38** (90:10, 10-CV), **Q31**(data preprocessing), **Q44** (YES), **Q45** (YES)	**Q111** (data imbalance, SMOTE)	**Q89** (SHAP feature importance + fuzzy-rules + DT rules: post-hoc model-agnostic, local and global), **Q90** (DE, YES), **Q91** (YES, accuracy + fidelity), **Q93** (YES), **Q96** (YES), **Q98** (YES), **Q101** (M), **Q103** (YES)	CN vs. MCI vs. AD sMCI vs. pMCI	93.9/-/-/-/93.9 87.1/-/-/-/87.1	93.3/-/-/-/93.8 91.8/91.7/91.7/91.7/91.8
Xiao et al. (2021)	50 times 10-CV	**Q20** (1,NO,-,-,-,-), **Q32** (LR), **Q35** (I), **Q43** (Single), **Q38** (10-CV), **Q25**& **Q31**(image preprocessing),	–	**Q89** (LR: interpretable and model specific, local and global), **Q85** (YES), **Q86** (YES), **Q90** (DE, NO), **Q92** (I), **Q101** (S), **Q91** (Yes, accuracy)	CN vs. MCI sMCI vs pMCI CN vs. AD	82.8/-/86.7/-/89 72.9/-/81.3/-/75 93.3/-/92.3/-/96	-/-/-/-/- -/-/-/-/- -/-/-/-/-
Zhang et al. (2021)	5-CV on ADNI1 + test by ADNI2&3	**Q20** (1,NO,-,-,-,-), **Q31**(image preprocessing), **Q27**(LOW), **Q32**(CNN), **Q35**(E), **Q38** (train:test, Yes), **Q60** (YES, ADNI)	–	**Q89** (Grad-CAM-based V: explainable, post-hoc, and CNN model specific, local), **Q90** (DE, NO), **Q101** (S)	CN vs. AD sMCI vs. pMCI	91.3/-/91/-/98.4 82.1/-/81.2/-/92	*CN/AD*: ADNI2 (90.9/-/78.8/-/88.9), ADNI3 (89.2/-/78.8/-/88.9)
Bass, et al. (2021)	Train: test	**Q20** (1,NO,-,-,-,-), **Q31** (MRI preprocessing), **Q38** (train:test, NO), **Q63** (C), **Q43** (Single), **Q32** (VAE)	–	**Q89** (MRI visualization: explainable by VAE-GAN feature attribution maps, model specific, and local), **Q90** (DE, NO), **Q102** (YES)	MCI vs. AD	NCC (-): 0.683	NCC (+): 0.652

Table 8 (continued)

Ref. (Year)	CV	Robustness requirement	Fairness requirement	Transparency requirement	Target	CV results	Testing results
Wen et al. (2021)	250 times repeated inner stratified 10-CV	**Q20** (2,NO,-,-,-,-), **Q31** (MRI preprocessing), **Q38** (inner CV, YES), **Q63** (C), **Q43** (multiple), **Q32** (SVM), **Q29** (YES), **Q45** (YES)	**Q111** (data imbalance, undersampling balance)	–	CN vs. AD (BA) CN vs. MCI (BA) sMCI vs. pMCI (BA)	94/-/-/-/- 74/-/-/-/- 80/-/-/-/-	-/-/-/-/- -/-/-/-/- -/-/-/-/-
Bron, et al. (2021)	Stratified 5-CV	**Q20** (1,NO,-,-,-,-), **Q35** (E), **Q32** (SVM), **Q43** (Single), **Q38** (5-CV), **Q31**(MRI preprocessing), **Q63** (C/D), **Q38** (E, YES), **Q56**(McNemar's test B corrected)	–	**Q89** (MRI visualization: explainable by saliency map using CNN and guided backprop, model specific, and local), **Q90** (DE, YES)	CN vs. AD sMCI vs. pMCI	-/-/-/-/94 -/-/-/-/75.6	-/-/-/-/89.6 -/-/-/-/66.5
Zhao et al. (2021)	Train:test	**Q20** (1,YES,3,-,-,-), **Q35** (E), **Q32** (CNN), **Q43** (Single), **Q38** (train:test), **Q31**(MRI preprocessing), **Q38** (E, YES), **Q56**(two-tailed paired t test), Q7 (YES, GAN)	–	–	CN vs. MCI vs. AD sMCI vs. pMCI	-/-/-/-/- -/-/-/-/-	76.7/78.1/ 70.8/74.3/- 78.5/86.8/ 52.4/54.1/-
Bin Bae, et al. (2020)	5-time repeated 80:20 train:test	**Q20** (1,NO,-,-,-,-), **Q35** (I/E), **Q32** (CNN), **Q29** (YES), **Q43** (Single), **Q38** (train:test), **Q31**(MRI preprocessing), **Q38** (I/E, YES), **Q56**(Student's t-test)	–	–	CN vs AD (ADNI results, SNUBH results)	89/-/88/-/94 88/-/85/-/91	83/-/76/-/88 82/-/79/-/89
Georges et al. 2020)	LOOCV	**Q20** (1,NO,-,-,-,-), **Q35** (E), **Q32** (SVM), **Q43** (Single), **Q38** (LOOCV), **Q31**(sMRI preprocessing), **Q38** (E, NO)	–	–	LMCI vs. AD	70.3/-/-/-/-	-/-/-/-/-

Table 8 (continued)

Ref. (Year)	CV	Robustness requirement	Fairness requirement	Transparency requirement	Target	CV results	Testing results
Zhao et al. (2020)	Train:test	**Q20** (2,NO,-,-,-,-), **Q31** (MRI preprocessing), **Q38** (Train:test, YES), **Q43** (multiple), **Q32** (SVM), **Q29** (YES), **Q45** (YES), **Q63** (C)	**Q111** (data imbalance, undersampling balance), **Q108** (pre-processing: LR to clean the effects of the age and sex)	–	CN vs. AD	88/-/88/-/95	79/-/87/-/89
Jo et al. (2020)	4 times repeated stratified 5-CV	**Q20** (1,NO,-,-,-,-), **Q32**(CNN), **Q38** (5-CV, NO), **Q29** (YES), **Q43** (Single)	–	**Q89** (MRI visualization: explainable by CNN's LRP relevance heatmaps, model specific, and local), **Q90** (DE, YES), **Q102** (YES)	CN vs. AD	90.8/-/-/-/-	-/-/-/-/-
Lian et al. (2020a)	Train with ADNI 1 (80:20 split) and test with ADNI 2	**Q20** (1,NO,-,-,-,-), **Q31**(image preprocessing), **Q56** (t-test), **Q32**(CNN), **Q35**(E), **Q38** (train:test, NO), **Q29** (YES), **Q43** (Single)	–	**Q89** (MRI visualization: explainable by CNN class activation map, model specific, and local), **Q92** (E), **Q90** (DE, YES),	AD vs. NC pMCI vs. sMCI	-/-/-/-/- -/-/-/-/-	90.3/-/82.4/-/95.1 80.9/-/52.6/-/78.1
Dyrba, et al. (2020)	Stratified 10-CV	**Q20** (1,NO,-,-,-,-), **Q31** (MRI preprocessing), **Q38** (10-CV, YES), **Q63** (C), **Q43** (Single), **Q35** (E), **Q32** (CNN), **Q29** (YES)	**Q108** (pre-processing: LR to clean the effects of the covariates age, sex, total intracranial volume, scanner magnetic field strength)	**Q89** (MRI visualization: explainable by 3D CNN relevance heatmaps by LRP, model specific, and global and local), **Q92** (E), **Q97** (YES,-,YES), **Q90** (DE, YES), **Q102** (YES)	CN vs. AD (test results for: ADNI3, AIBL, DELCODE) CN vs. MCI (test results for: ADNI3, AIBL, DELCODE)	88/-/-/-/94.3 72.6/-/-/-/76.7	90.7/-/-/-/93.9 83.9/-/-/-/91.8 92.5/-/-/-/94 67.6/-/-/-/68.3 76.1/-/-/-/74 74.3/-/-/-/76.9

Table 8 (continued)

Ref. (Year)	CV	Robustness requirement	Fairness requirement	Transparency requirement	Target	CV results	Testing results
Fukunishi et al. (2020)	70:30 train:test, 100 bootstraps	**Q20** (3,NO,-,-,-,-), **Q31** (data encoding and balancing), **Q29** (YES), **Q32** (LR), **Q35** (I), **Q43** (Single), **Q45** (YES)	**Q111** (data imbalance, undersampling balancing and bootstrapping), **Q112** (FPR and FNR)	**Q89** (LR: interpretable and model specific, local and global), **Q85** (YES), **Q86** (YES), **Q90** (DE, YES), **Q92** (E), **Q101** (S), **Q91** (Yes, accuracy)	AD vs. Not AD	-/-/-/-/-	63.9/10.5/ 61.7/18.0/66.3
Qiu et al. (2020)	5-time repeated train on ADNI and test on other three sets	**Q20** (3,NO,-,-,-,-), **Q31** (MRI preprocessing), **Q32** (CNN-MLP), **Q35** (I/E), **Q29** (YES), **Q43** (multiple), **Q45** (YES), **Q44** (YES), **Q63** (YES), **Q38** (train:test, Yes), **Q56** (t-test + χ^2 test)	–	**Q89** (MRI visualization: explainable, model specific, and local), **Q85** (YES), **Q86** (YES), **Q90** (DE, YES), **Q92** (I), **Q101** (S)	AD vs. CN for ADNI, AIBL, FHS, and NACC	-/-/-/-/- -/-/-/-/- -/-/-/-/- -/-/-/-/-	96.8/-/95.7/ 96.5/- 93.2/-/87.7/ 81.4/- 79.2/-/74.2/ 63.3/- 85.2/-/92.4/ 82.4/-
Biffi et al. (2020)	562:64:100 train: validate: test	**Q20** (1,NO,-,-,-,-), **Q31** (MRI preprocessing), **Q29** (YES), **Q32** (LVAE-MLP), **Q35** (E), **Q38** (train:validate:test, Yes)	–	**Q89** (MRI visualization by VAE highest laten space: explainable, model specific, and global), **Q101** (S), **Q90** (DE, YES), **Q92** (E)	CN vs. AD	-/-/-/-/-	84/-/78/-/-
Kim et al. (2020)	Train:test	**Q20** (1,NO,-,-,-,-), **Q32** (SVM), **Q35** (E), **Q29** (YES), **Q38** (train:test, Yes), **Q43** (Single), **Q56** (Pearson's chi-square)	–	–	CN vs. AD	94.8/-/92.1 /-/98	94.3/-/91.8/-/98
Adeli et al. (2019)	10 times repeated 10-CV	**Q20** (1,NO,-,-,-,-), **Q31** (MRI preprocessing), **Q29** (YES), **Q32** (LDA), **Q38** (10-CV, Yes), **Q56** (FET), **Q35** (E)	–	**Q89** (MRI visualization using LDA: explainable, model specific, and global), **Q101** (S), **Q90** (DE, YES)	CN vs. AD CN vs. MCI	92.1/-/-/-/- 91.9/-/-/-/-	-/-/-/-/- -/-/-/-/-

Table 8 (continued)

Ref. (Year)	CV	Robustness requirement	Fairness requirement	Transparency requirement	Target	CV results	Testing results
Oh et al. (2019)	20-time repeated nested 5-CV	**Q20** (1,NO,-,-,-,-), **Q31** (MRI preprocessing), **Q32** (CAE-ICAE), **Q38** (nested 5-CV, Yes), **Q56** (t-test)	–	**Q89** (MRI visualization using SMV: explainable, model specific, and global), **Q101** (S), **Q90** (DE, YES)	CN vs. AD CN vs. pMCI CN vs. sMCI sMCI vs. pMCI	86.6/-/88.6/-/- 77.4/-/81/-/- 63/-/59/-/- 74/-/77.5/-/-	-/-/-/-/- -/-/-/-/- -/-/-/-/- -/-/-/-/-
Seo et al. (2019)	10-CV	**Q20** (1,YES,7,-,-,-), **Q31** (MRI preprocessing), **Q32** (SVM), **Q38** (10-CV, NO),	–	**Q89** (MRI visualization: explainable, model specific, and local), **Q101** (S), **Q90** (DE, NO), **Q89**(YES)	CN vs. AD	92/-/-/-/-	-/-/-/-/-
Das et al. (2019)	Stratified 70: 30 train:test, 10-time repeated 5-CV	**Q20** (2,NO,-,-,-,-), **Q63** (C), **Q35** (I), **Q29** (YES), **Q32** (Rule + SVM), **Q43** (Single), **Q21** (YES), **Q38** (train:test with 5-CV, Yes), **Q65** (YES),	–	**Q85** (YES), **Q89** (weighted sum of short interval conjunction rules + rule importance + V: interpretable, model specific, local and global), **Q91** (Yes, accuracy) **Q90** (DE, YES), **Q102** (YES), **Q92** (I), **Q91** (YES, rule length, # of rules, and time validations), **Q101** (S)	CN vs. AD	-/-/84/-/86	-/-/84/-/81
Babapour Mofrad et al. (2019)	50:50 train:test,	**Q20** (1,NO,-,-,-,-), **Q32** (DT), **Q35** (I), **Q29** (YES), **Q38** (train:test, Yes), **Q43** (Single), **Q56** (t-test, Kruskal–Wallis, chi-square tests)	–	**Q89** (DT: interpretable, model specific, local and global), **Q85** (YES), **Q86** (YES), **Q90** (DE, YES), **Q92** (I), **Q101** (S), **Q91** (Yes, accuracy)	CN vs. AD sMCI vs. pMCI	90/-/93/-/- -/-/-/-/-	86/-/86/-/- 76/-/84/-/-

Table 8 (continued)

Ref. (Year)	CV	Robustness requirement	Fairness requirement	Transparency requirement	Target	CV results	Testing results
Ding et al. (2018)	90:10 train:test, then 5-time repeated 10-CV	**Q20** (6,YES,4,NO,YES,-), **Q29** (YES), **Q38** (I, YES), **Q35** (I), **Q43** (Single)	**Q111** (data imbalance, SMOTE)	**Q89** (BN: interpretable and model specific), **Q85** (YES), **Q86** (YES), **Q90** (DE, YES), **Q92** (I), **Q101** (S), **Q91** (Yes, accuracy)	CN vs. MCI vs. AD	-/-/-/-	80/-/-/91

Table 8 (continued)

DMAJ: Dataset from a metropolitan area in Japan; *ATC*: Chemical Classification System; *FPR*: False positive rate; *FNR*: False negative rate; *3D ResAttNet*: Explainable 3D residual attention deep neural network; *Grad-CAM*: Gradient-based localization class activation mapping; I: Internal; *E*: External; *S*: Single explanation technique, *M*: Mixture of explanation techniques, *SHIMR*: Sparse high-order interaction model with rejection option; *V*: Visualization; C: Code, D: Data; *IPPS*: Skull stripping, spatial standardization and segmentation, modulation, smoothing, registration, selecting the gray matter volume of the 90 auto-labeled regions of interest (ROI) as features; *CART*: Classification And Regression Tree; RF-RFE: Random forest Recursive feature elimination; *X*: Using information gain, Gini index, gain ratio, correlation, feature importance of the XGBoost classifier, and permutation importance using RF; BN: Bayesian network; *AIBL*: Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing, *FHS*: Framingham Heart Study, *NACC*: National Alzheimer's Coordinating Center; *LLE*: Locally linear embedding; *LVAE*: Ladder variational autoencoder; *SMV*: Saliency map visualization technique; *CAE-ICAE*: Convolutional autoencoder -inception module-based CAE; *H-FCN*: Hierarchical fully convolutional network; *LRP*: Layer-wise relevance propagation; *LR*: Linear regression; *ICAM*: Interpretable Classification via disentangled representations and feature Attribution Mapping; *NCC*: Normalized cross correlation; *NCC(+)*: Positive NCC; *NCC(-)*: Negative NCC; *DWI*: Diffusion-weighted image; *T1W*: T1 weighted; *FA*: Fractional anisotropy; *MD*: Mean diffusivity; *BA*: Balanced accuracy; *DR*: Dimension reduction; *DE*: Domain expert; *IST*: Intensity, shape, and texture features; *ADMM*: Alternating Direction Method of Multipliers; *FET*: Fisher exact test; *SHD*: Severance Hospital dataset; *BEGAN*: Boundary Equilibrium Generative Adversarial Network; *HPNDB*: Health-RI Parelsnoer Neurodegenerative Diseases Biobank data set; B: Bonferroni; *MCIPD*: MCI prediction dataset; *DIAN*: Dominantly Inherited Alzheimer Network; RUB: Resubstitution with upper bound correction; *SNUBH*: Seoul National University Bundang Hospital; *MTL*: Medial temporal lobe;

black-box models with some level of transparency. Abuhmed et al. (2021) designed three post-hoc model-agnostic explainers associated with an LSTM model for AD progression detection. Post-hoc techniques dealt with feature importance, DTs, fuzzy rules, and classical rules. In addition, the study used SHAP feature importance and 22 explainers to interpret the RF classifier decisions for AD prediction and progression detection (Sappagh et al. 2021).

Explaining black-box models mainly depends on the visualization of the MRI modality. For example, Qui et al. (2020) proposed a CNN-MLP hybrid model for AD diagnosis based on baseline 1.5T T1-weighted MRI data integrated with the patient's age, gender, and MMSE. The model was built and validated using 417 cases from ADNI. It was also validated with external data from AIBL (382), FHS (102), and NACC (565). Explainability was supported by MRI visualization and an extracted probability map from the CNN model (i.e., AD probability for every location of the MRI image). The authors excluded cases with other dementias and patients with a history of traumatic brain injury, depression, stroke, and brain tumors to remove possible data biases. More importantly, 9 neurologists participated in the model validation. Oh et al. (Oh et al. 2019) proposed an end-to-end DL model for AD progression detection based on MRI data. The model used volumetric convolutional autoencoder-based unsupervised learning for visualization map extraction for the CN vs. AD task. Then supervised fine-tuning was applied to the classifier of CN vs. AD. Using transfer learning, the visualization map was also used to predict sMCI vs. pMCI. The authors used the class saliency visualization (CSV) technique to facilitate the understanding of spatial influence in the decisions made by the DL model. CSV calculates how much each input voxel contributes to the final activation of the target class. The study concluded that the temporal and parietal lobes were identified as key MRI regions. Dyrba et al. (2020) highlighted the significant role of hippocampus atrophy in detecting AD based on the 3D CNN's relevance maps. Lian et al. (2020a) proposed a deep CNN model based on a hierarchical, fully convolutional network to detect AD progression and automatically identify hierarchical discriminative locations of brain atrophy at both the patch-level and region-level by using the class activation map technique. Jo et al. (2020) provided explainability based on Tau-PET images and a 3D-CNN with layer-wise relevance propagation-based model. The model asserted the informative role of the hippocampus, parahippocampal gyrus, thalamus, and fusiform gyrus. The same visualization technique was used by Dyrba et al. (2020). All in all, DL models achieved good diagnostic performance for detecting AD, but they are not yet applied in clinical routine (Dyrba et al. 2020), mainly due to their lack of explainability. Such explainability usually relies only on neuroimage visualization, which is not enough for physicians. Bass et al. (2021) asserted that the saliency-based methods which analyze network gradients such as gradient-weighted class activation mapping, perturbation methods, LRP, and guided backpropagation are not efficient: (1) the visualization is noisy and has low resolution; (2) they detect similar features in both healthy and disease groups, which makes it hard to interpret reported results; and (3) using a post-hoc technique is insufficient as DL models focus on most discriminative features to predict each class accurately. Bass et al. suggested that a plausible solution to these limitations is the use of GAN models, which can translate images from one class as AD to another class as MCI by means of capturing all relevant features of a population. The interested reader is referred to Singh et al. (2020), who published a survey of XAI methods based on images.

Because there is no real system in the AD domain applied in the medical environment, the current literature provides shallow XAI techniques. Thus, it can be concluded that transparency is not adequately addressed in the AD literature. Most questions related to

this TAI requirement have not been answered. The continuous (re)assessment of the quality of the input data to the XAI module has not been discussed (**Q79**&**Q80**). Tracking back which AI rules and data led to the decision has been handled only by interpretable models (Ding et al. 2018; Xiao et al. 2021; Qiu et al. 2020; Fukunishi et al. 2020; Das et al. 2019). Most longitudinal studies did not confirm the number of time steps with domain experts and did not provide time series explainability except (Seo et al. 2019) (**Q89**). Seo et al. (2019) tracked AD progression over time using SVM and MRI data. Longitudinal MRI data (seven-time steps) are transformed into 2D space using the local linear embedding (LLE) technique, and the resulting decision probabilities of SVM were used to color the LLE map over time. The resulting visualization provided the progression path of the patient over time. However, no study clearly determined the target stakeholders of the explainability (**Q90**) or the context in which the explanations must be considered (**Q99**). The generalizability (**Q92**) and combination of XAI with different formats (**Q93**&**Q101**) have not been discussed. Our previous study (Varghese et al. 2013) provided XAI with three methodologies, but we did not solve the conflict between explainers. There is a tight relation between transparency, fairness, and trustworthiness, where the first could be used to measure and enhance the last two. No study had used XAI to investigate and enhance fairness (**Q94**) and robustness (**Q95**). Some techniques like fuzzy logic, case-based reasoning, and ontology can be integrated with the ML or DL models to enhance their performance and interpretability, but as far as we know there is not any CDSS enhanced with semantic knowledge and case-based reasoning yet (**Q96**). Only (Dyrba et al. 2020) provided users with interactive explainability (**Q97**), and only (Sappagh et al. 2021) provided users with explainability based on different data types (**Q98**). None of the studies under consideration measured the level of understanding of the system (**Q100**). In addition, no study integrated knowledge-based models (e.g., rule-based, case-based, or ontology-based models) with data-driven models (**Q103**). Notice that building robust domain knowledge based on physicians' experience or standard clinical practice guidelines and using this knowledge to build CDSSs is very critical in the medical domain. These systems are transparent, and they are expected to be trusted by domain experts. In this regard, we have already proposed a comprehensive fuzzy ontology for AD which collects the disease knowledge required to build AD diagnosis and progression detection systems (Shoaip et al. 2020). This ontology can be populated with AD diagnosis rules and used as the backend to build an interpretable CDSS. Unfortunately, no study measured the complexity of the provided XAI features and the required time for their computation and interpretation (**Q88**).

5.2.2 Robustness

Robustness is related to model accuracy, validation, and reproducibility. Accordingly, this TAI requirement is considered in all studies in Table 8. Let us give further details on the most outstanding studies. Wen et al. (2021) combined diffusion MRI and T1 weighted MRI to predict AD progression based on a linear SVM classifier. The study performed an extensive analysis to select the best combination of features from the voxel-wise and ROI-wise using GM and WM from T1 weighted MRIs. The authors used stratified nested CV to validate the model properly. The outer loop was used to measure model performance, and the inner loop was used for hyperparameter optimization based on a balanced accuracy metric. The study emphasized the criticality of incorporating FS and feature scaling steps in the CV procedure to prevent optimistic performance and to prevent data leakage. As a result, it was achieved from 5% up to a 40% relative increase in balanced accuracy. The study

explored the impact of different FS techniques, feature scaling, data imbalance, smoothing strength, and imaging modalities methods on the classifier performance. Although the majority class had twice as many examples as the minority class in several experiments, the study employed a downsampling strategy to balance the data rather than a fairness analysis technique. Even if the study asserted that their results were reproducible, they did not properly handle all the robustness requirements. The same authors have performed a similar analysis based on T1 MRI and FDG-PET collected from ADNI, OASIS, and AIBL datasets (Samper-González et al. 2018). They confirmed the critical role of proper and relevant preparation of ML pipeline to calculate correct results and to build reproducible models. For example, Bron et al. (2021) divided the dataset from the beginning into training and testing sets. They performed all model and data preparations on the training data only, while test data was kept untouched for the testing stage. Zhao et al. (Qiu et al. 2020) built an SVM model for AD classification based on 715 MRI subjects collected from 6 different scanners from ADNI to improve model stability and result reproducibility. An independent ADNI dataset of 1228 subjects was included to assess reproducibility further. The authors concentrated on hippocampus features as suitable AD biomarkers. Then, the correlation between the hippocampus and other critical neurological biomarkers (e.g., APOE, CSF Aβ, CSF Tau) to detect MCI progression has been investigated. The study confirmed that hippocampus radiomic yields robust biomarkers for both AD diagnosis and progression detection. Adeli et al. (2019) suggested that model generalizability and robustness are affected by sample outliers, and feature noise inherently exists in neuroimaging data. Authors proposed a semi-supervised classifier, i.e., robust feature sample linear discriminant analysis (RFS-LDA), to handle these two challenges based on the LDA technique. Outliers are penalized by $\updownarrow_1$ fitting function. Using labeled and unlabeled data leads to better data denoising, and FS is done by $\updownarrow_1$ norm. The proposed model was tested on AD and Parkinson's datasets. Kim et al. (2020) used slice selective learning to select specific slices to improve computations and extract unbiased features. The selected two slices helped build a model less sensitive to the PET imaging acquisition environment. GAN was used for feature extraction, and an SVM model was tuned with two independent datasets. Jiménez-Mesa et al. (2021) tested the significance level of the correlation between the class label feature and other independent covariates to measure the level of accuracy of data labeling. This step is very critical because labeling errors are a well-known problem in medical datasets. This problem is especially important in AD datasets because patient labeling is mainly based on CSs values (e.g., MMSE, CDR, or FAQ). Zhao et al. (2021) proposed the 3D batch-based mi-GAN (multi-information GAN) DL model to predict the individual's whole brain 3D sMRI image at future time points (after 1 or 4 years) conditioning on multi-information (i.e., age, sex, education level, and APOE e4) at baseline. The model achieved a structural similarity index of 0.943 between the real images in the fourth year and the generated ones. The predicted images were used by another DL model (3D DenseNet) to predict the disease's multi-class clinical stage. The hybrid model achieved an enhanced accuracy of 6.04% compared to conditional GAN and cross-entropy loss. The proposal included a module determining whether an MRI image is real or automatically generated by GAN. This module can detect adversarial attacks. To improve the robustness of the model, the training images were linearly registered, and the testing images were nonlinearly registered. Georges et al. (2020) focused on model reproducibility and robustness to choose the best FS mechanism which boosts reproducibility capabilities. However, good performance of a particular FS method does not necessarily imply that the experiment is reproducible and that the selected features are the best for generalizability. This is because the FS algorithm could select a different feature set after applying a small perturbation in

the training data. Thus, we cannot trust the FS methods to generate an interpretable feature set. This study proposed an *FS-Select model* to select the most trustworthy and reliable features from a set of FS methods by exploring the relationships among FS methods. The study built a multi-graph architecture using the reproducibility power of each feature, average accuracy, and feature stability for each FS method.

Despite the recent effort to deal with robustness, this TAI requirement is not yet fully handled by all AD literature studies. None of the studies in Table 8 paid attention to either security (**Q5**–**Q13**) or safety (**Q14**–**Q19**) requirements. Indeed, no security standards or measures have been used to detect, evaluate, and prevent adversarial attacks. No study provided online learning (**Q66**–**Q69**), and so no study provided a mechanism to filter the fake and engineered cases (**Q9**). The situations where the system is not reliable have not been identified by any study, and measures to prevent the occurrence of these situations have not been implemented (**Q13**). Some studies used multimodal data (Sappagh et al. 2021; Qiu et al. 2020; Fukunishi et al. 2020; Das et al. 2019; Wen et al. 2021; Zhao et al. 2020), while others used time-series data (Abuhmed et al. 2021; Ding et al. 2018; Seo et al. 2019; Zhao et al. 2021). However, the sufficiency of the number of utilized time steps has not been medically confirmed by a domain expert, and the effect of sparse time-series data is not discussed (**Q20**). The level of accuracy required by domain experts has not been defined before building the models (**Q21**). Many critical issues are ignored in the AD literature: System monitoring (**Q23**), medical completeness of the used dataset (**Q24**), causality between features and between features and dependent features (**Q26**), the accuracy of the ground truth (**Q27**), and label biases (**Q28**), type of missingness (**Q30**), usage of AutoML (**Q47**), and interoperability with EHR ecosystem (**Q50**). Importantly, no study measured the stability of the models in terms of the sensitivity of the models to small changes in the data and borderline cases (**Q53**). No study critically analyzed data biases and imbalance (**Q55**). No study kept an audit trail (**Q58**) and specified a clear context for model reproducibility (**Q57**). All in all, most robustness measures have not been handled. This could be the main reason for not trusting AI-based CDSSs and not deploying them in real clinical environments. Compared to the medical literature, we find studies in other domains apart from AD where authors improved the robustness of models against adversarial attacks either by detecting adversarial examples or by adding mechanisms in the DL model to deal with fake examples (Xu et al. 2021). Ma et al., (2021) discovered adversarial attacks with over 98% detection AUC. Unfortunately, these types of issues are underexplored in the AD literature.

5.2.3 Fairness

Most studies neglected the fairness requirement, and thus they probably reported unfair results because the populations used in these studies were demographically biased (Bin Bae, et al. 2020). None of the studies in Table 8 has evaluated for potential optimistic bias in the classifier performance. Mendelson et al. (2017) provided an empirical study to evaluate potential biases in resampling-based experiments. The study concluded that selection bias accounts for more than half of the apparent performance improvements resulting from pipeline optimizations, mainly in the case of small datasets. The used dataset for most studies is ADNI which is inherently biased, as discussed before. For example, most ADNI subjects have high education levels, which affect brain structures, so the question of "*how models will perform for lower education levels?*" is unreliable to answer. The same problem is seen with other protected attributes like ethnicity, age, gender, etc. Please note that domain experts must define the features that must be treated as sensitive attributes,

and datasets and models must be prepared accordingly. For example, comorbidities such as hypertension and diabetes, adverse events such as Diarrhea and Pneumonia, and medications like Aricept and Cognex may be deemed sensitive features and should be balanced among different classes. For example, ADNI is biased regarding comorbidities, medications, and adverse events because a minority of the patients have values for these features. The interested readers are invited to take a look at the *MEDHIST.csv*, *RECBLLOG.csv*, *RECCMEDS.csv*, *BACKMEDS.csv*, and *BLSCHECK.csv* files to investigate the level of bias in ADNI. It is extremely critical to report suitable performance metrics which account for individual and group fairness. Most of the studies in the literature have not considered these requirements, which means that the real generalizability of these models is unknown. For example, Bae et al. (2020) built a CNN model of AD classification based on MRI scans collected from AD patients, and age/gender-matched CN controls from two populations that differ in ethnicity/education level. The study concentrated on improving the model's generalizability by doing extensive internal and external validations with two independent datasets. It is worth noting that the used ADNI dataset has ethnicity distribution as follows: Caucasians (83.59%), African Americans (4.87%), Hispanics (5.64%), Asians (2.05%), and others (3.85%). The other dataset contained only the Koreans population. In addition, the study used two MRI datasets collected by different scanners with different MRI protocols. However, no fairness questions were answered in the study.

Gamberger et al. (Wen et al. 2021) proposed a clustering mechanism that identifies homogeneous patient subpopulations regarding clinical and biological descriptors to measure the difference between males and females regarding AD in the ADNI dataset. The authors asserted a clear difference between male and female AD patient groups. Therefore, neglecting the gender-specific properties of AD data could result in the biased performance of generated models. Liu and Caselli (2018) discovered a high correlation between APOE e4 (a major genetic risk factor for AD) and age and gender. They discovered that APOE e4 is a more significant risk factor for young patients than for the old group. Accordingly, the authors asserted that neglecting this relation results in biased classifiers.

Unfortunately, no study in the AD literature researched the potential discrimination against a specific group of patients based on protected attributes (**Q107**–**Q122**). Indeed, no fairness questions have been discussed in the studies in Table 8 except for **Q111,** which is related to the source of bias (Sappagh et al. 2021; Abuhmed et al. 2021; Ding et al. 2018; Fukunishi et al. 2020; Wen et al. 2021; Zhao et al. 2020), and **Q108** which refers to the mechanisms implemented to ensure fairness (Dyrba, et al. 2020).

6 Trustworthy AI in the industry and other domains

Many industrial sectors are rapidly applying AI to critical tasks, yet these industrial entities prefer the highest-performing AI models that are often complex and work as black boxes. Firms use data and AI to create scalable solutions, but they are concerned about AI's ethical, legal, and economic ramifications (Blackman 2020). Microsoft, Facebook, Twitter, Google, and other giant technology companies are rapidly forming teams to address the requirements of TAI arising from the ubiquitous gathering, analysis, and use of vast amounts of data (Cheng et al. Sep. 2021), especially when that data is used to train practical machine learning models (Blackman 2020). Although these companies announce these efforts, we have no way to access details on such efforts because of their confidentiality. Few recent AI-specialized startups aim to develop AI solutions that are

transparent, accountable, responsible, fair, and ethical (Fiddler 2022). These startups construct an explainable AI engine and propose several XAI-based solutions that account for bias, fairness, and transparency in AI (Gade et al. 2019). A few key challenges that face the industrial sector while applying TAI can be summarized as follows (Blackman 2020). (1) The optimal approach for integrating the recently proposed TAI solutions into the existing conventional systems. (2) The distinction between the TAI requirements that the customer is interested in and those fundamental to the underlying business model. (3) Lake of a general approach to applying TAI requirements on different ML models due to the diversity of the models, data format, applications, etc. (4) The tradeoffs and contradictions among the TAI requirements significantly affect the model performance and customer expectations. Despite usually achieving a superior performance compared to domain experts, there is a critical gap between developing and integrating AI systems in real environments. TAI is a crucial challenge in safety–critical domains such as medicine, justice, and security, and limitations in implementing TAI requirements in AI-based applications are the main reason for that gap (González-Gonzalo, et al. 2021; Li, et al. 2021). Without TAI, the use of AI is expected to lead to significant harm, discrimination, and injustice (Crockett et al. 2021). Gonzalez-Gonzalo et al. (2021) built a TAI system for the ophthalmic domain. The study asserted that building successful TAI systems depends on stakeholders (e.g., AI developers, healthcare providers, domain experts, patients, regulatory agents, etc.) with different roles and responsibilities. In addition, most current AI studies considered one TAI requirement and neglected others. Most studies in the AI literature concentrated on model performance, which is insufficient to get user trust. Zicari et al. (2021a) assessed the application of TAI and ML as supportive tools to recognize cardiac arrest in emergency calls. This tool has been used in Copenhagen in, Denmark, since Fall 2020. Zicari et al. used Z-Inspection (Zicari et al. 2021b), which is a holistic process to assess TAI in practice. Z-Inspection is based on the TAI framework defined by the AI HLEG. The study determined specific challenges and ethical trade-offs to be considered when considering AI in practice. It recommended that an independent committee of experts evaluate the implementation of TAI requirements in the search for sociotechnical validation of AI systems before deployment. Lekadir et al. (2021) proposed FUTURE-AI. It includes guidelines and best practices for guiding future AI developments in medical imaging to improve models' trust by considering fairness, universality, traceability, usability, robustness, and explainability. Ma et al. (2022) have explored TAI in the dentistry domain.

Trustworthiness is also an urgent requirement in other domains such as business, education, government, administration, home automation, etc. (Kaur et al. 2023). Although a plethora of guidelines, principles, and toolkits have been published globally for TAI, Crockett et al. (2021) discovered that limited grassroots-level implementations can be found in the literature, especially among small- and medium-sized enterprises (SMEs). Crockett et al. identified the key barriers that SMEs faced in their adoption of TAI, and then implemented 77 published toolkits based on 33 evaluation criteria. The authors tried to understand the AI ethics landscape from the SME perspective and concluded that there is no one-size-fits-all tool to handle all TAI principles. Emaminejad and Akhavian (2022) studied the applications of TAI in robotics, especially in the architecture, engineering, and construction (AEC) industry. Authors discovered that although AEC researchers and industry professionals studied and deployed AI and robotics, there is a lack of studies that explore the key TAI dimensions in the AEC context. Besides explainability, robustness, and other TAI requirements, the study discussed the "trust calibration" concept to eliminate the overtrust and undertrust of humans in robotics. Loureiro et al. (2020) surveyed the literature on applications of AI and robotics in the business context and identified TAI

as a key challenge and development trend. De Bruyn et al. (2020) surveyed the pitfalls and opportunities of AI in marketing. The authors predicted that AI would fall short of its promises in many marketing domains if its technological pitfalls were not handled, including badly defined objective functions, biased AI, explainable AI, and controllable AI. Zhang and Zhou (1912) discussed fairness in the business domain and evaluated fairness metrics in a credit card default payment example. Borg et al. (2021) provided a case study for applying ALTAI principles in an advanced driver assistance system. The study asserted that some TAI requirements, such as human agency, transparency, and societal and environmental impact, were difficult to apply in the project. Hutiri and Ding (2022) explored the concept of "trustworthy edge intelligence (TEI)." The authors determined the requirements for TEI, provided a unified framework for TEI, and provided an application scenario of voice-activated services. In the education domain, Vincent-Lancrin (2021) explored the role of AI in improving educational processes and to prepare students with new skill sets for increasingly automated economies and societies. AI can be used in personalized learning, supporting students with special needs, analyzing classroom dynamics and student engagement, online and blended learning, and foreign language learning. The authors highlighted the importance of some TAI requirements (e.g., fairness, privacy, and security) to gain the trust of stakeholders, but they asserted that TAI implementation is rare in AI-based education systems. Corbeil and Corbeil (2021) explored the benefits, challenges, and applications of AI in education and mentioned that the lack of user trust is the main reason for its limited application. Holmes et al. (2021) surveyed the opinion of 60 researchers from AI in education to answer a questionnaire about ethics and the application of AI in educational contexts. The study discovered that most researchers are not trained to tackle ethical questions. Baker and Hawn (2021) provided a comprehensive survey of the algorithmic bias in education and pointed out the potential causes of that bias. The study discussed the literature on biases from race/ethnicity, gender, nationality, socioeconomic status, disability, and military-connected status. Fenu et al. (2207) surveyed the fairness challenges of AI systems in the education through anonymous surveys and interviews with education experts. Marcus and Davis (2019) stated that the implementation of TAI depends on the domain and the context in which the AI system is used. For example, the facial recognition software that is used for auto-tagging people in social media pictures could be less reliable. However, the same tool is unacceptable if the police want to use it to find suspects in surveillance photos. The current state-of-the-art of TAI in all these domains is insufficient to get users' trust. Shneiderman (2020) proposed 15 recommendations in team, organization, and industry levels. These recommendations are intended to enhance the trustworthiness of AI systems. Writer et al. (2019) asserted that developing standards and policies to govern AI systems' development and deployment processes leads to faster technology transfer, interoperability, security, and reliability.

7 Conclusion and future research directions

TAI is essential for critical medical problems such as Alzheimer's disease to provide a more applicable and trustworthy application to medical experts. This study reviewed in depth TAI guidelines and requirements. We have evaluated the recent literature on AD diagnosis and progression detection from the perspective of TAI. Keeping in mind the inclusion/exclusion criteria, we have carefully revised more than 400 papers and provided readers with a complete picture of the state of art in this field of research. We found a

shortage in the revised AD literature regarding TAI requirements, which could demonstrate why we did not see the real effect of AI in the AD domain or why these systems have not been generally deployed in clinical practice. Since all TAI guidelines are general and abstract, and cannot be applied directly in the ML application development process, we proposed a detailed TAI-based machine learning pipeline which considered the majority of TAI requirements in all steps of the ML pipeline, including data preparation, training, validation, and deployment. This proposed pipeline will help ML researchers and developers embed TAI requirements into their proposed pipelines. Furthermore, we proposed a comprehensive questionnaire for TAI requirements that help researchers to assess, enhance, and evaluate the AI application in hand for TAI requirements. The questionnaire also includes a list of questions for every TAI requirement that could help to measure how well the TAI requirement is applied in the given application.

7.1 Learned lessons

This work provides a unique contribution to the AI community as a whole, and it is especially relevant for research on the AD domain. There are many lessons that could be learned from the current study.

1. To build real CDSSs, researchers should keep in mind all the TAI requirements collected in the supplementary material. In a separate research study, we will implement a TAI-based CDSS system that fulfills the TAI requirements for Alzheimer's disease.
2. The inclusion of patients and domain expert in all steps of the AI pipeline is critical. They must evaluate the model's robustness, fairness, and explainability from the medical perspective.
3. The standard ML pipeline should be extended to include TAI requirements. All existing pipelines mostly concentrate on the model performance, which is not enough for life-threatening applications of AI.
4. AI research has no real effect in real medical environments because patients and medical experts do not trust AI-based decisions. Model stability, certainty, generalizability, security, privacy, etc. must be considered in AI research to provide medically applicable solutions.
5. Current TAI measures are represented as high-level and abstract guidelines. Only a few software packages and studies have mapped these guidelines into measurable metrics.
6. Most studies in the literature mainly concentrated on a single TAI dimension, such as model robustness or transparency, and totally neglected other dimensions.
7. TAI dimensions sometimes contradict each other. For example, increasing model transparency could result in lower performance and lower security and privacy. A tradeoff always exists among TAI measures.
8. Most research studies did not consider the external validation of the model using datasets from other sources different from those used in model training and internal testing.
9. Many studies divided datasets into training and validation sets only after performing the data preprocessing steps, and this action resulted in information leakage, which may cause the model to provide optimistic and not real results.

7.2 Challenges

We have shed light on the current limitations and possible directions for enhancements of AI-based CDSSs for AD. The following is a list of open challenges and future research directions in the medical domain within this context.

1. Collected data for AD management are time series in nature. Most of the healthcare data is not prepared for ML, and all existing datasets have short sequences of data that are not suitable for DL models, such as LSTM and CNN models. A significant number of patients could be dropped out while the study is going on, which worsens the incompleteness problem. In addition, the sample size becomes very small, and it is hindered by the curse of dimensionality. GAN can be used to generate samples. Because AD is a slow-progressing disease and studies are short due to funding, many forms of data censoring could occur, including (1) the patient could remain in the CN stage while the study and convert to MCI just after its finishing, and (2) the patient could be in AD stage before the beginning of the study. Data censoring results in over- or under- sampling of specific AD stages may result in biased models. Efforts should be done to collect complete and uncensored data.
2. Model stability is a crucial requirement to be used in clinical environments. The reproducibility of model results using different data could be used to measure the generalizability of models (Yamanakkanavar et al. 2020). However, due to different inclusion and exclusion criteria or different demographic characteristics in clinical studies, significant under-representation of some populations could occur. As a result, ML models are usually created with non-representative cohorts of the general AD population, so when revalidated with new data, they result in poor generalization performance (Routier et al.). Reproducibility can be enhanced if the study (1) depends on standard datasets, (2) applies standardized data management, especially for neuroimages, (3) uses cross-validation to reduce overfitting, (4) tests models on completely different cohorts, and (5) makes the code publicly available.
3. Although several large-scale datasets have been collected, such as ADNI and AIBL, these studies are subject to bias because of the inclusion and exclusion criteria used to select subjects. A potential solution to this issue is to rely on the EHR data, which provides a more representative view of patient measurements—using the whole patient profile with a powerful ML model results in an individualized and patient-tailored decision. Very few studies have explored the role of EHR data in improving AD prediction, progression detection, or monitoring.
4. It is challenging to compare the model performance across different studies due to variations in the implemented steps, such as sample selection, data preprocessing, validation procedures, and reported evaluation metrics. Some studies may report a biased performance because of insufficient or unclear validation and model selection procedures. For example, using the whole dataset in the feature selection, data augmentation, or autoencoder steps results in data leakage (i.e., use of test data in any part of the training process (Rathore et al. 2017)) and thus biased performance because the same data could appear in several sets. The data should be divided into training/validation/testing sets at the very beginning, and afterwards only the train/validation sets should be used in data preparation and model selection. Especially for DL models, the test sets should be left untouched until the very end of model testing.

Whole pipeline optimization using manual or autoML techniques is better for building unbiased models (He et al. 2021).

5. Data interoperability is the ability to map data from one study to another study. This helps to build a model using a dataset from one study and validate this model with data from another study. This way of doing is likely to improve the generalizability of models and solve the problem of small sample sizes. However, studies are built independently, and finding a mapping rule between different datasets is complex. This process requires assessing the comparability of features and subjects in different datasets. First, feature mapping requires specifying the relationships between data elements and standardizing the used vocabulary in both sources. For example, ADNI specified the hippocampus volume biomarker as "Hippocampus", but EPAD specified it as "lhvr" (right hemisphere) and "lhvl" (left hemisphere). Second, subjects in each study should be comparable. For example, if the distribution of gender is not comparable in the datasets, then cognitive impairment scores cannot be compared directly. Data standardization and normalization is a suitable strategies for enhancing data integration. Also, a standardized set of markers and biomarkers could be collected from the integrated sets. Interoperability plays a major role in integrating CDSSs with EHR systems. Standard ontologies such as SNOMED CT and unified data formats such as FHIR can help in facing this challenge.
6. The explainability of models is a crucial requirement for AD models to have a clinical impact (Graham et al. 2019). In the AD domain, there are many markers and biomarkers to consider, and DL models can achieve higher accuracy. However, these models are not causal and capture the non-linear correlation between features. Moreover, DL models are considered black boxes, which do not explain their decisions. On the other hand, simpler models like linear regression and DTs are more interpretable but less accurate than DL models. XAI is a hot topic today, but further efforts are required to explore how to apply XAI techniques to understand AD models. For example, combining multiple models such as DTs, fuzzy rule-based systems, Bayesian networks, case-based reasoning, and data visualization with the black-box deep learning models, so-called hybrid models, could bridge the gap between explainability and accuracy.
7. Data fusion of related neuroimaging (e.g., sMRI, fMRI, PET), clinical, cognitive, and fluid biomarker data is crucial to understand the disease and to improve its early diagnosis, accurate prediction, and patient monitoring (Bayram et al. 2018). Fusion of multimodal data results in big complex data, including multimedia data (image, sound, etc.), time-series data, structured data (lab tests, symptoms, demographics, etc.), and semantic data (medications, symptoms, and comorbidities), towards semi-structured and unstructured text documents. This fusion calls for complex algorithms to extract meaningful features and to learn from these features. Although many techniques exist for analyzing big data, they have rarely been rigorously tested and put into practical use in medical environments (Brati and Zoran 2018). In addition, the resulting models are very complex and hard to interpret by domain experts.
8. Knowledge-based models are based on domain experts, CPG knowledge, and literature. The extraction of this knowledge is a challenge. In addition, the number of publications in the AD domain is exponentially increasing. For example, nearly 15,000 publications appeared in 2017 (Khatami et al. 2020). Collecting the most recent findings in a representative high-quality literature corpus and using text mining and NLP techniques to extract meaningful and computable knowledge representations is a big challenge (Gyori et al. 2017). Moreover, the automatic checking of the medical applicability and compatibility of the extracted knowledge is time-consuming and difficult. Repre-

senting this knowledge in the form of IF–THEN rules, knowledge graphs, Bayesian networks, or semantic ontologies is another challenge. More research is needed in this direction to improve the knowledge extraction and representation process. In addition, research work is needed to shed light on the possible ways for improving the integration between this general knowledge and patient-specific knowledge extracted from data-driven models.

9. A way to address the challenge of data labeling is to use clustering techniques as a step before classification to categorize data into groups. However, different clusters could be found among MCI subjects, so further research is also required here.
10. TAI techniques are still in an immature stage. For example, the tools, techniques, and metrics for dealing with AI fairness face many dilemmas, including the lack of a unified definition of fairness, the need to measure equity versus equality as well as the tradeoff between fairness and model performance, facing the disagreement and incompatibility of fairness metrics, and tensions with context and policy (Caton and Haas 2020; Mehrabi et al. 2019).
11. Longitudinal data analysis faces two main challenges to be used in a clinically meaningful way in AD progression detection. *The first challenge is temporal alignment*. To consistently compare patients, the acquired markers should be temporally aligned, but different patients can be (1) at different stages of the disease at the same acquisition time and (2) at different biological ages. Moreover, the collected data show only a short period of the full AD onset, more than 20 years. There are many approaches to handle this issue like using reference variables, including (1) time from baseline acquisition (Franke 2012), (2) patient age (R.J. Bateman et al. 2012), (3) normalized age (Lorenzi et al. 2014), (4) cognitive scores (Guerrero et al. 2016), and (5) data-driven progression scores (Casanova et al. 2018). However, this issue is still a major challenge that needs to be considered either as a preprocessing step or as a standalone task. The issue becomes more sophisticated in multimodal longitudinal data because (1) each patient could have a different number of acquisitions, what causes imbalanced data, (2) missing data results from missing acquisitions, (3) data are not collected at the same time point for all patients, and (4) the time window between follow-ups may not to be uniform. *The second issue is the missing values*. This is a very common issue in longitudinal data. Handling missing values is required to avoid biased models, leveraging available data, etc. Missing longitudinal data have many types depending on the pattern of missing values, but most studies assumed that data are missing at random. The issue becomes more complex when handling missing data in multimodal longitudinal data (Martí-Juan et al. 2020).
12. Further investigations of neuroimaging modalities are required to define the most accurate modality and the best pipeline to prepare data. Some studies have reported that sMRI is more discriminative than PET (Suk and Lee 2014) or DTI (Ebrahimighahnavieh et al. 2020); others reported MRI to be as discriminative as PET (Suk and Lee 2015) or less discriminative (Lu et al. 2018c), and some studies reported fMRI (Sarraf 2016) as more helpful.
13. Preparing neuroimaging data carefully for DL models is crucial. For example, using 2D slices as input instead of the whole 3D prevents the generation of millions of training parameters and results in simpler neural network models. Voxel-based methods can obtain all 3D information in a single brain scan. However, these methods have the limitations of (1) all brain regions are treated uniformly without caring about the special anatomical structures, (2) they ignore local information because they treat each voxel independently, and (3) they carry high feature dimensionality and high computational

load. These raw brain scans suffer from noise at different levels and from different sources such as equipment, operator issues, patient's random neural activity, and environment. The generated complex patterns of voxels from single scans create difficulties in classifying and interpreting features. ROI-based methods select a limited number of discrete predefined and more representative regions, which are easier to implement and interpret in a real environment. These methods are based on the predefined domain expert's knowledge of the most important brain regions, and their performance is based on the number of defined regions. As a result, the abnormalities of the neglected regions might be ignored, what leads to the loss of some discriminative information and limit the discriminative power of the extracted features. Hippocampus is the most significant region where its volume, shape, and texture are affected and have been used for early AD detection (Ebrahimighahnavieh et al. 2020). However, a few studies have combined the volume, shape, thickness, intensity, and texture features to perform an integrative analysis for AD detection and prediction (Rémi Cuingnet et al. 2011). In addition, the role of longitudinal change for these features has not been studied in the literature. Notice that the effect of other brain regions has not been deeply studied as for the hippocampus.

14. To provide an efficient CDSS for real clinical settings, developers should study how to enrich CDSSs with multiple comorbidities simultaneously, how to provide and to estimate the effect of CDSSs on the clinical and organizational outcomes, and how CDSSs can be more effectively integrated into the workflow and deployed across diverse settings (Lebedev et al. 2014). Validation of CDSSs is a factor that can make their implementation successful but providing evidence of clinical efficiency is a resource-consuming task that requires support from a sustainable validation methodology (Ito et al. 2011; Zhou et al. 2013).
15. The role of non-functional methods for TAI like regulation, standardization, certification, regulation, education and awareness, and accountability by the government, should be further deeply studied.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10462-023-10415-5.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2021R1A2C1011198), (Institute for Information & communications Technology Planning & Evaluation) (IITP) grant funded by the Korea government (MSIT) under the ICT Creative Consilience Program (IITP-2021-2020-0-01821), and AI Platform to Fully Adapt and Reflect Privacy-Policy Changes (No. 2022-0-00688). This research was funded by the Spanish Ministry for Science and Innovation (grants PID2020-112623GB-I00, PDC2021-121072-C21, PID2021-123152OB-C21 and TED2021-130295B-C33) and the Galician Ministry of Education, University and Professional Training (grants ED431C2022/19, and ED431G2019/04). All these grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

Abadi M et al (2016) Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp 308–318

Abrol A, Bhattarai M, Fedorov A, Du Y, Plis S (2020) Calhoun V (2020) Deep residual learning for neuroimaging: an application to predict progression to Alzheimer's disease. J Neurosci Methods 339:108701

Abuhmed T, El-sappagh S, Alonso JM (2021) Robust hybrid deep learning models for Alzheimer's progression detection. Knowledge-Based Syst 213:106688
Adeli E et al (2019) Semi-supervised discriminative classification robust to sample-outliers and feature-noises. IEEE Trans Pattern Anal Mach Intell 41(2):515–522
Adler P et al (2018) Auditing black-box models for indirect influence. Knowl Inf Syst 54(1):95–122
Ahmed R et al (2019a) Neuroimaging and machine learning for dementia diagnosis : recent advancements and future prospects. IEEE Rev Biomed Eng 12:19–33
Ahmed S et al (2019b) Ensembles of patch-based classifiers for diagnosis of alzheimer diseases. IEEE Access 7:73373–73383
AI HLEG (2019) Policy and investment recommendations for trustworthy AI. Brussels Indep High-Level Expert Gr Artif Intell (AI HLEG), Rep Publ by Eur. Commun, p 52
Alberdi A, Aztiria A, Basarab A (2016) On the early diagnosis of Alzheimer's disease from multimodal signals: a survey. Artif Intell Med 71:1–29
Ali S, El-Sappagh S, Ali F, Imran M, Abuhmed T (2022) Multitask deep learning for cost-effective prediction of patient's length of stay and readmission state using multimodal physical activity sensory data. IEEE J Biomed Heal Inform, 1–14
Alonso J, Castiello C, Magdalena L, Mencar C (2021) Explainable fuzzy systems: paving the way from interpretable fuzzy systems to explainable AI systems, 1st edn. Springer, Berlin Heidelberg
Alonso JM, Bugar A (2019) ExpliClas : automatic generation of explanations in natural language for weka classifiers. In: IEEE international conference on fuzzy systems, pp 1–6
Al-Rubaie M, Chang JM (2019) Privacy-preserving machine learning: threats and solutions. IEEE Secur Priv 17(2):49–58
Altaf T, Anwar SM, Gul N, Majeed MN, Majid M (2018) Multi-class Alzheimer's disease classification using image and clinical features. Biomed Signal Process Control 43:64–74
An N, Ding H, Yang J, Au R, Ang TFA (2020) Deep ensemble learning for Alzheimer's disease classification. J Biomed Inform 105:103411
Arachchige PCM, Bertok P, Khalil I, Liu D, Camtepe S, Atiquzzaman M (2020) A trustworthy privacy preserving framework for machine learning in industrial IoT systems. IEEE Trans Ind Inform 16(9):6092–6102
Armañanzas R, Iglesias M, Morales DA, Alonso-Nanclares L (2017) Voxel-based diagnosis of alzheimer's disease using classifier ensembles. IEEE J Biomed Heal Informatics 21(3):778–784
Arnold M et al (2018) FactSheets: Increasing trust in ai services through supplier's declarations of conformity. IBM J Res Dev 63(4):1–13
Arras L, Montavon G, Müller K-R, Samek W (2017) Explaining recurrent neural network predictions in sentiment analysis. arXiv1706.07206
Arrieta AB et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI". Inf Fusion 58:82–115
Arya V et al (2020) Ai explainability 360: an extensible toolkit for understanding data and machine learning models. J Mach Learn Res 21:1–6
Ashmore R, Calinescu R, Paterson C (2019) Assuring the machine learning lifecycle: desiderata, methods, and challenges. arXiv
Association A (2019) Alzheimer's disease factsand figures. Alzheimer's Dement 15(3):321–387
Auret L, Aldrich C (2012) Interpretation of nonlinear relationships between process variables by use of random forests. Miner Eng 35:27–42
Babapour-Mofrad R et al (2019) Decision tree supports the interpretation of CSF biomarkers in Alzheimer's disease. Alzheimer's Dement Diagn Assess Dis Monit 11:1–9
Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7):e0130140
Bai Z, Watson F, Yu DR, Shi H, Yuan Y, Zhang Y (2009) Abnormal resting-state functional connectivity of posterior cingulate cortex in amnestic type mild cognitive impairment. Brain Res 1302:167–174
Baker RS, Hawn A (2021) Algorithmic bias in education. Springer, New York
Baniecki H, Kretowicz W, Piatyszek P, Wisniewski J, Biecek P (2020) Dalex: responsible machine learning with interactive explainability and fairness in python. arXiv, pp 1–7
Barakat NH, Bradley AP (2007) Rule extraction from support vector machines: a sequential covering approach. IEEE Trans Knowl Data Eng 19(6):729–741
Barreno M, Nelson B, Joseph AD, Tygar JD (2010) The security of machine learning. Mach Learn 81(2):121–148
Basaia S et al (2019) Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. NeuroImage Clin. 21:101645

Basheera S, Sai-Ram MS (2019) Convolution neural network–based Alzheimer's disease classification using hybrid enhanced independent component analysis based segmented gray matter of T2 weighted magnetic resonance imaging with clinical valuation. Alzheimer's Dement Transl Res Clin Interv 5:974–986

Basheera S, Satya-Sai-Ram M (2020) A novel CNN based Alzheimer's disease classification using hybrid enhanced ICA segmented gray matter of MRI. Comput Med Imaging Graph 81:101713

Baskar D, Jayanthi VS, Jayanthi AN (2019) An efficient classification approach for detection of Alzheimer's disease from biomedical imaging modalities. Multimed Tools Appl 78(10):12883–12915

Bass C et al (2021) ICAM-reg: interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans. IEEE Trans Med Imaging 20:1–13

Bastani O, Ioannou Y, Lampropoulos L, Vytiniotis D, Nori A, Criminisi A (2016) Measuring neural net robustness with constraints. arXiv1605.07262

Bateman RJ et al (2012) Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N Engl J Med 367(9):795–804

Bayram E, Caldwell JZK, Banks SJ (2018) Current understanding of magnetic resonance imaging biomarkers and memory in Alzheimer ' s disease. Alzheimer's Dement Transl Res Clin Interv 4:395–413

Beheshti I, Demirel H, Matsuda H (2017) Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. Comput Biol Med 83(February):109–119

Bellamy RKE et al (2019) AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM J Res Dev 63(4/5):1–1

Ben-Braiek H, Khomh F (2020) On testing machine learning programs. J Syst Softw 164:110542

Bender EM, Friedman B (2018) Data statements for natural language processing: toward mitigating system bias and enabling better science. Trans Assoc Comput Linguist 6(May):587–604

Benton A, Mitchell M, Hovy D (2017) Multi-task learning for mental health using social media text. arXiv

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2018) Fairness in criminal justice risk assessments: The state of the art. Sociol Methods Res, 0049124118782533

Beutel A, Chen J, Zhao Z, Chi EH (2017) Data decisions and theoretical implications when adversarially learning fair representations. arXiv

Beutel A, et al (2019) Putting fairness principles into practice: challenges, metrics, and improvements. In: Proceedings of the 2019 AAAI/ACM conference on AI, Ethics, and Society, pp 453–459

Bi XA, Hu X, Wu H, Wang Y (2020) Multimodal data analysis of alzheimer's disease based on clustering evolutionary random forest. IEEE J Biomed Heal Informatics 24(10):2973–2983

Biffi C et al (2020) Explainable anatomical shape analysis through deep hierarchical generative models. IEEE Trans Med Imaging 39(6):2088–2099

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

Biggio B, Fumera G, Roli F (2013) Security evaluation of pattern classifiers under attack. IEEE Trans Knowl Data Eng 26(4):984–996

Bin Bae J, et al (2020) Identification of Alzheimer's disease using a convolutional neural network model based on T1—weighted magnetic resonance imaging. *Sci Rep*, pp 1–10

Binth M, Noor T, Zenia NZ, Kaiser MS, Al Mamun S, Mahmud M (2020) Application of deep learning in detecting neurological disorders from magnetic resonance images : a survey on the detection of Alzheimer ' s disease, Parkinson ' s disease and schizophrenia. Brain Inform

Birkenbihl C et al (2020) Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia—lessons for translation into clinical practice. EPMA J 11(3):367–376

Blackman R (2020) A practical guide to building ethical AI. Harvard Bus Rev, vol 15

Blanco-Justicia A, Domingo-Ferrer J, Martínez S, Sánchez D (2020) Machine learning explainability via microaggregation and shallow decision trees. Knowledge-Based Syst 194:105532

Blennow H, Zetterberg K (2018) Biomarkers for Alzheimer's disease: current status and prospects for the future. J Intern Med 284:643–663

Borg M, et al (2021) Exploring the assessment list for trustworthy ai in the context of advanced driver-assistance systems. In: Proc.—2021 IEEE/ACM 2nd int. work. ethics softw. eng. res. pract. ethics 2021, pp 5–12

Brand L, Nichols K, Wang H, Shen L, Huang H (2020) Joint multi-modal longitudinal regression and classification for alzheimer's disease prediction. IEEE Trans Med Imaging 39(6):1845–1855

Brati B, Zoran O (2018) Machine learning for predicting cognitive diseases : methods , data sources and risk factors

Bron EE et al (2021) Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. NeuroImage Clin 31:102712
Bruun M et al (2019) Impact of a clinical decision support tool on prediction of progression in early-stage dementia: a prospective validation study. Alzheimer's Res Ther 11(1):1–11
Bucholc M et al (2019) A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. Expert Syst Appl 130:157–171
Budd S, Robinson EC, Kainz B (2019) Survey on active learning and human-in-the-loop deep learning for medical image analysis. arXiv 71, 102062
Buruk B, Ekmekci PE, Arda B (2020) A critical perspective on guidelines for responsible and trustworthy artificial intelligence. Med Heal Care Philos 23(3):387–399
Buvaneswari PR, Gayathri R (2021) Deep learning-based segmentation in classification of alzheimer's disease. Arab J Sci Eng 46(6):5373–5383
Calders T, Verwer S (2010) Three naive Bayes approaches for discrimination-free classification. Data Min Knowl Discov 21(2):277–292
Calmon FP, Wei D, Vinzamuri B, Ramamurthy KN, Varshney KR (2017) Optimized pre-processing for discrimination prevention. In Proceedings of the 31st international conference on neural information processing systems, pp 3995–4004
Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp), pp 39–57
Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1721–1730
Carvalho CM, Seixas FL, Conci A, Muchaluat-Saade DC, Laks J, Boechat Y, A dynamic decision model for diagnosis of dementia, Alzheimer's disease and Mild Cognitive Impairment. Comput Biol Med 126
Caton S, Haas C (2020) Fairness in machine learning: a survey. arXiv, pp 1–33
Cearns M, Hahn T, Baune BT (2019) Recommendations and future directions for supervised machine learning in psychiatry. Transl Psychiatry. https://doi.org/10.1038/s41398-019-0607-2
Celik B, Vanschoren J (2021) Adaptation strategies for automated machine learning on evolving data. IEEE Trans Pattern Anal Mach Intell
Celis LE, Huang L, Keswani V, Vishnoi NK (2019) Classification with fairness constraints: a meta-algorithm with provable guarantees. In: Proceedings of the conference on fairness, accountability, and transparency, pp 319–328
Chang H, Shokri R (2020) On the privacy risks of algorithmic fairness. arXiv
Chen Y, Xia Y (2021) Iterative sparse and deep learning for accurate diagnosis of Alzheimer's disease. Pattern Recognit 116:107944
Cheng L, Varshney KR, Liu H (2021) Socially responsible AI algorithms: issues, purposes, and challenges. J Artif Int Res 71:1137–1181
Cherepanova V, Nanda V, Goldblum M, Dickerson JP, Goldstein T (2021) Technical challenges for training fair neural networks. arXiv2102.06764
Chetelat G (2018) Multimodal neuroimaging in alzheimer's disease : early diagnosis, physiopathological mechanisms, and impact of lifestyle. J Alzheimer's Dis 64:S199–S211
Chiappa S (2019) Path-specific counterfactual fairness. Proc AAAI Conf Artif Intell 33(01):7801–7808
Chitradevi D, Prabha S (2020) Analysis of brain sub regions using optimization techniques and deep learning method in Alzheimer disease. Appl Soft Comput J 86:105857
Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J (2016) RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. Adv Neural Inf Process Syst 3512–3520
Choi JY, Lee B (2020) Combining of multiple deep networks via ensemble generalization loss, based on MRI images, for alzheimer's disease classification. IEEE Signal Process Lett 27:206–210
Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163
Chouldechova A, G'Sell M (2017) Fairer and more accurate, but for whom?.arXiv1707.00046
Corbeil ME, Corbeil JR (2021) Establishing trust in artificial intelligence in education. In: Trust, organizations and the digital economy. Routledge, pp 49–60
Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, pp 797–806.
Cortez P, Embrechts MJ (2013) Using sensitivity analysis and visualization techniques to open black box data mining models. Inf Sci (ny) 225:1–17
Cotter A et al (2019) Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. J Mach Learn Res 20(172):1–59

Cowgill B, Tucker C (2017) Algorithmic bias: a counterfactual perspective. NSF Trust. Algorithms
Crockett KA, Gerber L, Latham A, Colyer E (2021) Building trustworthy AI solutions: a case for practical solutions for small businesses. IEEE Trans Artif Intell
Crone SF, Lessmann S, Stahlbock R (2006) The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing. Eur J Oper Res 173(3):781–800
Cruz RMO, Sabourin R, Cavalcanti GDC (2018) Dynamic classifier selection: recent advances and perspectives. Inf Fusion 41:195–216
Cruz AF, Saleiro P, Belém C, Soares C, Bizarro P Promoting fairness through hyperparameter optimization, vol 1, no 1. Association for Computing Machinery.
Cui R, Liu M (2019) RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. Comput Med Imaging Graph 73:1–10
Cutillo CM et al (2020) "Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. Npj Digit Med 3(1):1–5
Das D, Ito J, Kadowaki T, Tsuda K (2019) An interpretable machine learning model for diagnosis of Alzheimer's disease. PeerJ 7:e6543
Dasgupta P, Collins J (2019) A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks. AI Mag 40(2):31–43
De Bruyn A, Viswanathan V, Beh YS, Brock JKU, von Wangenheim F (2020) Artificial intelligence and marketing: pitfalls and opportunities. J Interact Mark 51:91–105
Deng H (2019) Interpreting tree ensembles with inTrees. Int J Data Sci Anal 7(4):277–287
Di Stefano PG, Hickey JM, Vasileiou V (2020) Counterfactual fairness: removing direct effects through regularization. arXiv2002.10774
Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 10(7):1895–1923
Dignum V (2019) Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer Nature, NewYork
Dimitriadis SI, Liparas D (2018) Random forest feature selection, fusion and ensemble strategy: combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: from the alzheimer's disease neuroimaging initiative (ADNI) data. J Neurosci Methods 302:14–23
Ding X et al (2018) A hybrid computational approach for efficient Alzheimer's disease classification based on heterogeneous data. Sci Rep 8(1):1–10
Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R (2020) Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. J Am Med Inform Assoc 27(4):592–600
Divya R, Shantha Selva Kumari R (2021) Genetic algorithm with logistic regression feature selection for Alzheimer's disease classification. Neural Comput Appl, vol 0123456789
Dua M, Makhija D, Manasa PYL, Mishra P (2020) A CNN–RNN–LSTM based amalgamation for Alzheimer's disease detection. J Med Biol Eng 40(5):688–706
Duc NT, Ryu S, Qureshi MNI, Choi M, Lee KH, Lee B (2020) 3D-deep learning based automatic diagnosis of alzheimer's disease with joint MMSE prediction using resting-state fMRI. Neuroinformatics 18(1):71–86
Duchesne S, Caroli A, Geroldi C, Collins DL, Frisoni GB (2009) Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. Neuroimage 47(4):1363–1370
Dunn C, Moustafa N, Turnbull B (2020) Robustness evaluations of sustainable machine learning models against data poisoning attacks in the internet of things. Sustainability 12(16):6434
Durán JM, Jongsma KR (2021) Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics. https://doi.org/10.1136/medethics-2020-106820
Durongbhan P et al (2019) A dementia classification framework using frequency and time-frequency features based on EEG signals. IEEE Trans Neural Syst Rehabil Eng 27(5):826–835
Dwork C, Immorlica N, Kalai AT, Leiserson M (2018) Decoupled classifiers for group-fair and efficient machine learning. In: Conference on fairness, accountability and transparency, pp 119–133
Dyrba M et al (2020) Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease. arXiv
Ebrahimighahnavieh A, Luo S, Chiong R (2020) Deep learning to detect Alzheimer ' s disease from neuroimaging: a systematic literature review. Comput Methods Programs Biomed 187:105242
Eke CS, Jammeh E, Li X, Carroll C, Pearson S, Ifeachor E (2021) early detection of Alzheimer's disease with blood plasma proteins using support vector machines. IEEE J Biomed Heal Inform 25(1):218–226

El Sappagh S, Alonso JM, Islam SMR, Sultan AM (2021) A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci Rep. https://doi.org/10.1038/s41598-021-82098-3
El-Gamal FEZA et al (2020) Personalized computer-aided diagnosis for mild cognitive impairment in alzheimer's disease based on sMRI and C PiB-PET analysis. IEEE Access 8:218982–218996
El-Rashidy N et al (2022) Sepsis prediction in intensive care unit based on genetic feature optimization and stacked deep ensemble learning. Neural Comput Appl 34(5):3603–3632
El-Sappagh S, Elmogy M, Riad AM (2015) A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis. Artif Intell Med 65(3):179–208
El-Sappagh S, Alonso JM, Ali F, Ali A, Jang JH, Kwak KS (2018) An ontology-based interpretable fuzzy decision support system for diabetes diagnosis. IEEE Access 6:37371–37394
El-Sappagh T, Abuhmed SM, Riazul-Islam KS (2020) Kwak, "Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data." Neurocomputing 412:197–215
El-Sappagh S, Ali F, Abuhmed T, Singh J, Alonso JM (2022a) Automatic detection of Alzheimer's disease progression: an efficient information fusion approach with heterogeneous ensemble classifiers. Neurocomputing 512:203–224
El-Sappagh S, Saleh H, Ali F, Amer E, Abuhmed T (2022b) Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time. Neural Comput Appl 34:14487–14509
El-Sappagh S et al (2021) Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data. Futur Gener Comput Syst 115
Emaminejad N, Akhavian R (2022) Trustworthy AI and robotics: Implications for the AEC industry. Autom Constr 139:104298
Er F, Goularas D (2021) Predicting the prognosis of MCI patients using longitudinal MRI data. IEEE/ACM Trans Comput Biol Bioinform 18(3):1164–1173
Casanova R, Barnard RT, Gaussoin SA, Saldana S, Hayden KM, Manson JE, Wallace RB, Rapp SR, Resnick SM, Espeland MA, Chen J-C, et al (2018) Using high-dimensional machine learning methods to estimate an anatomical risk factor for Alzheimer's disease across imaging databases. Neuroimage 183:401–411
Hochrangige Expertengruppe für Künstliche Intelligenz (HEG-KI), High-Level Expert Group on Artificial Intelligence (2020) Assessment list for trustworthy AI (ALTAI) for self assessment
Falahati F, Institutet K, Westman E, Institutet K, Simmons A (2014) Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. J Alzheimer's Dis 41:685–708
Fang C et al (2020) Gaussian discriminative component analysis for early detection of Alzheimer's disease: A supervised dimensionality reduction algorithm. J Neurosci Methods, 344(July):108856
Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 259–268
Feng R, Yang Y, Lyu Y, Tan C, Sun Y, Wang C (2019) Learning fair representations via an adversarial framework. arXiv1904.13341
Fenu G, Galici R, Marras M (2022) Experts' view on challenges and needs for fairness in arti cial intelligence for education. arXiv:2207.01490v1
Fiddler (2022) Build ethical AI using explainable AI
Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam L, Kohane IS (2019) Adversarial attacks on medical machine learning. Science 363(6433):1287–1290
Finlayson SG, Kohane IS, Beam AL (2018) Beam adversarial attacks against medical deep learning systems. arXiv1804.05296
Fisher CK, Smith AM, Walsh JR (2018) Deep learning for comprehensive forecasting of Alzheimer's disease progression. arXiv1807.03876
Forouzannezhad P et al (2019) A survey on applications and analysis methods of functional magnetic resonance imaging for Alzheimer's disease. J Neurosci Methods 317:121–140
Forouzannezhad P et al (2020) A Gaussian-based model for early detection of mild cognitive impairment using multimodal neuroimaging. J Neurosci Methods 333:108544
Franke CGK (2012) Longitudinal changes in individual Brain AGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. GeroPsych (bern) 25(4):235–245
Fukunishi H, Nishiyama M, Luo Y, Kubo M, Kobayashi Y (2020) Alzheimer-type dementia prediction by sparse logistic regression using claim data. Comput Methods Programs Biomed 196:105582

Fung BCM, Wang K, Chen R, Yu PS (2010) Privacy-preserving data publishing: a survey of recent developments. ACM Comput Surv 42(4):1–53
Gacto MJ, Alcalá R, Herrera F (2011) Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. Inf Sci (ny) 181(20):4340–4360
Gade K, Geyik SC, Kenthapadi K, Mithal V, Taly A (2019) Explainable AI in industry. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp 3203–3204
Gao J, Wang B, Lin Z, Xu W, Qi Y (2017) Deepcloak: masking deep neural network models for robustness against adversarial samples. arXiv1702.06763
Gardiner J, Nagaraja S (2016) On the security of machine learning in malware c&c detection: a survey. ACM Comput Surv 49(3):1–39
Gardner J, Brooks C (2017) Statistical approaches to the model comparison task in learning analytics.MLA/BLAC@ LAK
Gebru T et al (2021) Datasheets for datasets. Commun ACM 64(12):86–92
Georges N, Mhiri I, Rekik I (2020) Identifying the best data-driven feature selection method for boosting reproducibility in classification tasks. Pattern Recognit 101:107183
Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G (2018) Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 178(11):1544–1547
Gillen S, Jung C, Kearns M, Roth A Online learning with an unknown fairness metric. arXiv Prepr. arXiv1802.06936 (2018)
Gillespie N, Curtis C, Bianchi R, Akbari A, van Vlissingen R (2020) Achieving trustworthy AI: a model for trustworthy artificial intelligence
Goel N, Amayuelas A, Deshpande A, Sharma A (2020) The importance of modeling data missingness in algorithmic fairness: a causal perspective. arXiv2012.11448
Goh G, Cotter A, Gupta M, Friedlander M (2016) Satisfying real-world goals with dataset constraints. Adv neural inf. process. syst., Nips 2016, pp 2423–2431
Gómez-Sancho M, Tohka J, Gómez-Verdejo V (2018) Comparison of feature representations in MRI-based MCI-to-AD conversion prediction. Magn Reson Imaging 50(March):84–95
Gonzalez Zelaya CV (2019) Towards explaining the effects of data preprocessing on machine learning. In: Proc.—int. conf. data eng., vol. 2019-April, pp 2086–2090
González-Gonzalo C et al. (2021) Trustworthy AI: closing the gap between development and integration of AI systems in ophthalmic practice. Prog Retin Eye Res
Goodfellow I, et al (2014) Generative adversarial nets. Adv Neural Inf Process Syst
Gopinath D, Pasareanu CS, Wang K, Zhang M, Khurshid S (2019) Symbolic execution for attribution and attack synthesis in neural networks. In: Proc—2019 IEEE/ACM 41st int. conf. softw. eng. companion, ICSE-companion 2019, pp 282–283
Graham SA et al (2020) Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: a conceptual review. Psychiatry Res 284:112732
Grgic-Hlaca N, Zafar MB, Gummadi KP, Weller A (2016) The case for process fairness in learning: feature selection for fair decision making. NIPS Symp Mach Learn Law 1:1
Gu S, Rigazio L (2015) Towards deep neural network architectures robust to adversarial examples. In: Workshop paper at international conference on learning representative (ICLR)
Guerrero R, Schmidt-Richberg A, Ledig C, Tong T, Wolz R, Rueckert D et al (2016) Instantiated mixed effects modeling of Alzheimer's disease markers. Neuroimage 142:113–125
Guidotti R, Monreale A, Ruggieri S, Turini F, Pedreschi D, Giannotti F (2018) A survey of methods for explaining black box models. ACM Comput Surv 51(5):931–9342
Gyori G, Bachman BM, Subramanian JA, Muhlich K, et al (2017) "From word models to executable models of signaling networks using automated assembly. Mol Syst Biol, 13:954
Haas C (2019) The price of fairness-a framework to explore trade-offs in algorithmic fairness. In: 40th international conference on information systems, ICIS
Hajian S, Bonchi F, Castillo C (2016) Algorithmic bias: From discrimination discovery to fairness-aware data mining. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 2125–2126
Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. arXiv1610.02413
Hatherley JJ (2020) Limits of trust in medical AI. J Med Ethics 46(7):478–481
He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K (2019) The practical implementation of artificial intelligence technologies in medicine. Nat Med 25(1):30–36
He X, Zhao K, Chu X (2021) AutoML: a survey of the state-of-the-art. Knowledge-Based Syst 212:106622

He Y, Meng G, Chen K, Hu X, He J (2020) Towards security threats of deep learning systems: a survey. IEEE Trans Softw Eng
Hedayati R, Khedmati M, Taghipour-Gorjikolaie M (2021) Deep feature extraction method based on ensemble of convolutional auto encoders: application to Alzheimer's disease diagnosis. Biomed Signal Process Control 66:102397
Holmes W, et al. (2021) Ethics of AI in education: towards a community-wide framework. Int J Artif Intell Educ, 504–526
Hong X et al (2019) Predicting alzheimer's disease using LSTM. IEEE Access 7:80893–80901
Hossain MA, Ferdousi R, Alhamid MF (2020) Knowledge-driven machine learning based framework for early-stage disease risk prediction in edge environment. J Parallel Distrib Comput 146:25–34
Huang M et al (2017) Longitudinal measurement and hierarchical classification framework for the prediction of Alzheimer's disease. Sci Rep 7:1–13
Huang X et al (2020) A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. Comput Sci Rev 37:100270
Hutiri WT, Ding AY (2022) Towards trustworthy edge intelligence : insights from voice-activated services. In: IEEE serv. comput. conf., 2022.
Ignatiev A, Cooper MC, Siala M, Hebrard E, Marques-Silva J (2020) Towards formal fairness in machine learning. In: Lect. notes comput. sci. (including subser. lect. notes artif. intell. lect. notes bioinformatics), vol 12333 LNCS, pp 846–867
Iosifidis V, Fetahu B, Ntoutsi E (2019) Fae: a fairness-aware ensemble framework. In 2019 IEEE international conference on big data (big data), pp 1375–1380
Ito K et al (2011) Disease progression model for cognitive deterioration from Alzheimer's disease neuroimaging initiative database. Alzheimer's Dement 7(2):151–160
Jacobs M et al. (2021) Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In: Proc. 2021 CHI conf. hum. factors comput. syst., pp 1–14
Jacovi A, Marasović A, Miller T, Goldberg Y (2021) Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: FAccT 2021—proc 2021 ACM conf. fairness, accountability, transpar, Section 2, pp 624–635
Jain A, Ravula M, Ghosh J (2020) Biased models have biased explanations. arXiv2012.10986
Japkowicz N (2006) Why question machine learning evaluation methods. In: AAAI workshop on evaluation methods for machine learning, pp 6–11
Ji M, Yu G, Xi H, Xu T, Qin Y (2021) Measures of success of computerized clinical decision support systems: An overview of systematic reviews. Heal Policy Technol 10(1):196–208
Jie B, Liu M, Liu J, Zhang D, Shen D (2017) Temporally constrained group sparse learning for longitudinal data analysis in alzheimer's disease. IEEE Trans Biomed Eng 64(1):238–249
Jiménez-Mesa C et al. (2021) Deep learning in current neuroimaging: a multivariate approach with power and type I error control but arguable generalization ability
Jin M, Deng W (2018) Predication of different stages of Alzheimer's disease using neighborhood component analysis and ensemble decision tree. J Neurosci Methods 302:35–41
Jo T, Nho K, Saykin AJ (2019) Deep learning in alzheimer ' s disease : diagnostic classification and prognostic prediction using neuroimaging data". Front Aging Neurosci. https://doi.org/10.3389/fnagi.2019.00220
Jo T, Nho K, Risacher SL, Saykin AJ (2020) Deep learning detection of informative features in tau PET for Alzheimer's disease classification. BMC Bioinform 21(21):1–14
Joshi G, Walambe R, Kotecha K (2021) A review on explainability in multimodal deep neural nets. IEEE Access 9:59800–59821
Ju R, Hu C, Li Q (2017) Early diagnosis of Alzheimer's disease based on resting-state brain networks and deep learning. IEEE/ACM Trans Comput Biol Bioinforma 16(1):244–257
Juraev F, El-Sappagh S, Abdukhamidov E, Ali F, Abuhmed T (2022) Multilayer dynamic ensemble model for intensive care unit mortality prediction of neonate patients. J Biomed Inform 135:104216
Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33(1):1–33
Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Fairness-aware classifier with prejudice remover regularizer. In Joint European conference on machine learning and knowledge discovery in databases, pp 35–50
Katell M et al (2020) Toward situated interventions for algorithmic equity: lessons from the field. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 45–55
Katzir Z, Elovici Y (2018) Quantifying the resilience of machine learning classifiers used for cyber security. Expert Syst Appl 92:419–429

Kaur D, Uslu S, Rittichier KJ, Durresi A (2023) Trustworthy artificial intelligence: a review. ACM Comput Surv 55(2):105–115
Keane MT, Kenny EM (2019) The twin-system approach as one generic solution for XAI: an overview of ANN-CBR twins for explaining deep learning. arXiv1905.08069
Khatami SG et al (2020) Challenges of integrative disease modeling in Alzheimer ' s disease. Front Mol Biosci 6:1–13
Khedher L, Ramírez J, Górriz J, Brahim A (2015) Early diagnosis of Alzheimer's disease based on partial least, squares principal componentanalysis and support vector machine using segmented MRI images. Neurocomputing 151:139–150
Khvostikov A, Aderghal K, Krylov A, Catheline G, Benois-Pineau J (2018) 3D inception-based CNN with sMRI and MD-DTI data fusion for Alzheimer's disease diagnostics. arXiv:1809.03972
Kim HW, Lee HE, Lee S, Oh KT, Yun M, Yoo SK (2020) Slice-selective learning for Alzheimer's disease classification using a generative adversarial network: a feasibility study of external validation. Eur J Nucl Med Mol Imaging 47(9):2197–2206
Kim MP, Reingold O, Rothblum GN (2018) Fairness through computationally-bounded awareness. arXiv1803.03239
Kindermans P-J, et al (2017) Learning how to explain neural networks: patternnet and patternattribution. arXiv1705.05598
Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. arXiv1609.05807
Kovalchuk SV, Kopanitsa GD, Derevitskii IV, Savitskaya DA (2020) Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. J Biomed Inform 127:1–23
Krasanakis E, Spyromitros-Xioufis E, Papadopoulos S, Kompatsiaris Y (2018) Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In: Proceedings of the 2018 World Wide Web conference, pp 853–862
Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 5(4):221–232
Kuo KL, Fuh CS (2011) A rule-based clinical decision model to support interpretation of multiple data in health examinations. J Med Syst 35(6):1359–1373
Kurakin A, Goodfellow I, Bengio S (2018) Adversarial examples in the physical world. In: Secur AIS (ed) RV yampolskiy. Chapman and Hall/CRC, Boca Raton, pp 99–112
Kuznetsov SO (2001) Machine learning on the basis of formal concept analysis. Autom Remote Control 62(10):1543–1564
Kwon BC et al (2019) RetainVis: visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. IEEE Trans vis Comput Graph 25(1):299–309
Lamy JB, Sekar B, Guezennec G, Bouaud J, Séroussi B (2019) Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. Artif Intell Med 94:42–53
Lazli L, Boukadoum M (1894) A survey on computer-aided diagnosis of brain disorders through MRI based on machine learning and data mining methodologies with an emphasis on alzheimer disease diagnosis and the contribution of the multimodal fusion. Appl Sci 10:2020
Lebedev AV et al (2014) Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. NeuroImage Clin 6:115–125
Ledley RS, Lusted LB (1959) Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Nucl Phys 13(1):104–116
Lee G et al (2019) Predicting Alzheimer's disease progression using multi-modal deep learning approach. Sci Rep 9(1):1–12
Lei B, Hou W, Zou W, Li X, Zhang C, Wang T (2019) Longitudinal score prediction for Alzheimer's disease based on ensemble correntropy and spatial–temporal constraint. Brain Imaging Behav 13(1):126–137
Lei B et al (2020) Self-calibrated brain network estimation and joint non-convex multi-task learning for identification of early Alzheimer's disease. Med Image Anal 61:101652
Lei B et al (2020) Deep and joint learning of longitudinal data for Alzheimer's disease prediction. Pattern Recognit, vol 102
Lekadir K, et al (2021) FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging
Li H, Wang X, Ding S (2018) Research and development of neural network ensembles: a survey. Artif Intell Rev 49(4):455–479
Li K, Wu Z, Peng K-C, Ernst J, Fu Y (2019a) Guided attention inference network. IEEE Trans Pattern Anal Mach Intell 42(12):2996–3010
Li W, Zhao Y, Chen X, Xiao Y, Qin Y (2019b) Detecting Alzheimer's disease on small dataset: a knowledge transfer perspective. IEEE J Biomed Heal Informatics 23(3):1234–1242

Li W, Lin X, Chen X (2020) Detecting Alzheimer's disease based on 4D fMRI: an exploration under deep learning framework. Neurocomputing 388:280–287
Li H, Habes M, Wolk DA, Fan Y (2019c) A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal MRI. Alzheimer's Dement, pp 1–12
Li A, Li F, Elahifasaee F, Liu M, Zhang L (2021) Hippocampal shape and asymmetry analysis by cascaded convolutional neural networks for Alzheimer's disease diagnosis. Brain Imaging Behav
Li B, et al (2021) Trustworthy AI: from principles to practices
Lian C, Liu M, Zhang J, Shen D (2020a) Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE Trans Pattern Anal Mach Intell 42(4):880–893
Lian C, Liu M, Pan Y, Shen D (2020b) Attention-guided hybrid network for dementia diagnosis with structural MR images. IEEE Trans Cybern, pp 1–12
Liang Y, Li S, Yan C, Li M, Jiang C (2021) Explaining the black-box model: a survey of local interpretation methods for deep neural networks. Neurocomputing 419:168–182
Liu L, Caselli RJ (2018) Age stratification corrects bias in estimated hazard of APOE genotype for Alzheimer's disease. Alzheimer's Dement Transl Res Clin Interv 4:602–608
Liu F, Zhou L, Shen C, Yin J (2014) Multiple kernel learning in the primal for multimodal alzheimer's disease classification. IEEE J Biomed Heal Informatics 18(3):984–990
Liu X, Goncalves AR, Cao P, Zhao D, Banerjee A (2018) Modeling Alzheimer's disease cognitive scores using multi-task sparse group lasso. Comput Med Imaging Graph 66:100–114
Liu Q, Li P, Zhao W, Cai W, Yu S, Leung VCM (2018a) A survey on security threats and defensive techniques of machine learning: a data driven view. IEEE Access 6:12103–12117
Liu M, Cheng D, Wang K, Wang Y (2018b) Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis. Neuroinformatics 16(3–4):295–308
Liu M, Zhang J, Adeli E, Shen D (2018c) Joint classification and regression via deep multi-task multichannel learning for alzheimer's disease diagnosis. IEEE Trans Biomed Eng 66:1195–1206
Liu R, Rong Y, Peng Z (2020) A review of medical artificial intelligence. Glob Heal J 4(2):42–45
Liu W, Qiu JL, Zheng WL, Lu BL (2021a) Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Trans Cogn Dev Syst 8920:1–15
Liu Y et al (2021b) Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. Med Image Anal 69:101953
Liu Y, Radanovic G, Dimitrakakis C, Mandal D, Parkes DC (2017) Calibrated fairness in bandits. arXiv1707.01875
Ljubic B et al (2020) Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction. Comput Methods Programs Biomed 197:105765
Lorenzi M, Pennec X, Frisoni GB, Ayache N et al (2014) Disentangling normal aging from Alzheimer's disease in structural MR images. Neurobiol Aging 16(9):801
Lorenzi M, Filippone M, Frisoni GB, Alexander DC (2017) Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer ' s disease. Neuroimage 190:56–68
Lou Y, Caruana R, Gehrke J, Hooker G (2013) Accurate intelligible models with pairwise interactions. In: Proc. ACM SIGKDD int. conf. knowl. discov. data min, vol Part F1288, pp 623–631
Loureiro SMC, Guerreiro J, Tussyadiah I (2021) Artificial intelligence in business: state of the art and future research agenda. J Bus Res 129:911–926
Lu S, Xia Y, Cai W, Fulham M, Feng DD (2017a) Early identification of mild cognitive impairment using incomplete random forest-robust support vector machine and FDG-PET imaging. Comput Med Imaging Graph 60:35–41
Lu D, Popuri K, Ding GW, Balachandar R, Beg MF (2018a) Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease. Med Image Anal 46:26–34
Lu D, Popuri K, Ding W, Balachandar R, Beg MF (2018b) Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural MR and FDG-PET images. Sci Rep 8(1):5697
Lu MFBD, Popuri K, Ding GW, Balachandar R (2018c) Multimodal and mul- tiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. Sci Rep 8(1):5697
Lu J, Issaranon T, Forsyth D (2017b) Safetynet: detecting and rejecting adversarial examples robustly. In: Proceedings of the IEEE international conference on computer vision, pp 446–454

Ma X et al (2021) Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognit 110:107332
Ma J et al (2022) (2022) Towards trustworthy AI in dentistry. J Dent Res 10:002203452211060
Madaio MA, Stark L, Wortman Vaughan J, Wallach H (2020) Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In: Conf hum factors comput syst - proc, pp 1–14
Maksymiuk S, Gosiewska A, Biecek P (2020) Landscape of R packages for eXplainable artificial intelligence, vol XX, pp 1–26
Malakar S, Ghosh M, Bhowmik S, Sarkar R, Nasipuri M (2020) A GA based hierarchical feature selection approach for handwritten word recognition. Neural Comput Appl 32(7):2533–2552
Maleki F, Muthukrishnan N, Ovens K, Reinhold C, Forghani R (2020) Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment. Neuroimaging Clin N Am 30(4):433–445
Marcus G, Davis E (2019) Rebooting AI: Building artificial intelligence we can trust. Vintage, New York
Mariotti E, Alonso-Moral JM, Gatt A (2021) Prometheus: harnessing fuzzy logic and natural language for human-centric explainable artificial intelligence. In: XX Spanish congress on fuzzy logic and technologies, pp 274–279
Markus AF, Kors JA, Rijnbeek PR (2021) The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 113:103655
Martí-Juan G, Sanroma-Guell G, Piella G (2020) A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease. Comput Methods Programs Biomed 189:105348
McDaniel P, Papernot N, Celik ZB (2016) Machine learning in adversarial settings. IEEE Secur Priv 14(3):68–72
McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M (2021) Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med. https://doi.org/10.1126/scitranslmed.abb1655
McGill School of Computer Science (2020) The machine learning paper paper reproducibility checklist, pp 0–1
McGraw G, Bonett R, Figueroa H, Shepardson V (2019) Security engineering for machine learning. Computer 52(8):54–57
McLachlan S, Dube K, Hitman GA, Fenton NE, Kyrimi E (2020) Bayesian networks in healthcare: Distribution by medical condition. Artif Intell Med 107:101912
Mehdipour-Ghazi M et al (2019) Training recurrent neural networks robust to incomplete data: application to Alzheimer's disease progression modeling. Med Image Anal 53:39–46
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2019) A survey on bias and fairness in machine learning. arXiv
Mendelson AF, Zuluaga MA, Lorenzi M, Hutton BF (2017) Selection bias in the reported performances of AD classification pipelines. NeuroImage Clin 14:400–416
Middleton B, Sittig DF, Wright A (2016) Clinical decision support: a 25 year retrospective and a 25 year vision. Yearb Med Inform 1:S103–S116
Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38
Minhas S, Khanum A, Riaz F, Khan SA, Alvi A (2017) Predicting progression from mild cognitive impairment to Alzheimer's disease using autoregressive modelling of longitudinal and multimodal biomarkers. IEEE J Biomed Heal Informatics 22(3):818–825
Mitchell M, et al (2019) Model cards for model reporting. In: FAT* 2019: proc. 2019 conf fairness, accountability, transpar., no Figure 2, pp 220–229
Moher D, Liberati A, Tetzlaff J, Altman DG, Group P (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 6(7):e1000097
Mohseni S, Zarei N, Ragan ED (2018) A multidisciplinary survey and framework for design and evaluation of explainable AI systems. arXiv1811.11839
Molnar C (2018) Interpretable machine learning: a guide for making black box models explainable. Leanpub
Moosavi-Dezfooli S-M, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2574–2582
Moosavi-Dezfooli S-M, Fawzi A, Fawzi O, Frossard P (2017) Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1765–1773

Moustakidis S, Papandrianos NI, Christodolou E, Papageorgiou E, Tsaopoulos D (2020) Dense neural networks in knee osteoarthritis classification: a study on accuracy and fairness. Neural Comput Appl 5
Muhammed-Niyas KP, Thiyagarajan P (2021) Alzheimer's classification using dynamic ensemble of classifiers selection algorithms: a performance analysis. Biomed Signal Process Control 68:102729
Bacanin Ruxandra N-S, Zivkovic M, Petrovic, A, Rashid, TA, Bezdan T (2021) Performance of a novel chaotic firefly algorithm with enhanced exploration for tackling global optimization problems: application for dropout regula, "performance of a novel chaotic firefly algorithm with enhanced exploration for tackling global optimization problems: application for dropout regularization. Mathematics 9(21): 2705
Nanni L, Lumini A, Zaffonato N (2018) Ensemble based on static classifier selection for automated diagnosis of mild cognitive impairment. J Neurosci Methods 302:42–46
Narayanan A (2018) Translation tutorial: 21 fairness definitions and their politics. In: Proc. conf. fairness accountability transp, New York, USA, 2018, vol 2, no 3, pp 2–6
Narayanan A, Shmatikov V (2008) Robust de-anonymization of large datasets (how to break anonymity of the netflix prize dataset). In: University of Texas at Austin
Nasiri S, Zahedi G, Kuntz S, Fathi M (2019) Knowledge representation and management based on an ontological CBR system for dementia caregiving. Neurocomputing 350:181–194
Neri E, Coppola F, Miele V, Bibbolino C, Grassi R (2020) Artificial intelligence: who is responsible for the diagnosis? Radiol Medica 125(6):517–521
Nguyen L, Wang S, Sinha A (2018) A learning and masking approach to secure learning. In International conference on decision and game theory for security, pp 453–464
Nicolae MI et al (2018) Adversarial robustness toolbox v0.4.0. arXiv, 1–34
Nie L, Zhang L, Meng L, Song X, Chang X, Li X (2017) Modeling disease progression via multisource multitask learners: a case study with alzheimer's disease. IEEE Trans Neural Networks Learn Syst 28(7):1508–1519
Nordberg A, Rinne JO, Kadir A, Långström B (2010) The use of PET in Alzheimer disease. Nat Rev Neurol 6(2):78–87
Oh K, Chung YC, Kim KW, Kim WS, Oh IS (2019) Classification and visualization of alzheimer's disease using volumetric convolutional neural network and transfer learning. Sci Rep 9(1):1–16
Olah C et al (2018) The building blocks of interpretability. Distill 3(3):e10
On R (2019) High-level expert group on artificial intelligence ethics guidelines for trustworthy AI. Eur Commun, 09(04)
P 1607-U GOES Procedures (1978) Code of federal regulations
Page MJ et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. J Clin Epidemiol 134:178–189
Pan X, Adel M, Fossati C, Gaidon T, Guedj E (2019) Multilevel feature representation of FDG-PET brain images for diagnosing alzheimer's disease. IEEE J Biomed Heal Informatics 23(4):1499–1506
Pan X et al (2021) Multi-view separable pyramid network for AD prediction at MCI stage by 18F-FDG brain PET imaging. IEEE Trans Med Imaging 40(1):81–92
Pang T, Xu K, Du C, Chen N, Zhu J (2019) Improving adversarial robustness via promoting ensemble diversity. In: 36th Int. conf. mach. learn. ICML 2019, vol 2019-June, pp 8759–8771
Papangelou K, Sechidis K, Weatherall J, Brown G (2018) Toward an understanding of adversarial examples in clinical trials. In Joint European conference on machine learning and knowledge discovery in databases, pp 35–51
Papernot N, McDaniel P (2018) Deep k-nearest neighbors: towards confident, interpretable and robust deep learning. arXiv1803.04765
Papernot N, Song S, Mironov I, Raghunathan A, Talwar K, Erlingsson Ú (2018) Scalable private learning with pate. arXiv1802.08908
Park JH, et al (2020) Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. npj Digit Med
Pellegrini E et al (2018) Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. Alzheimer's Dement 10:519–535
Peng J, Zhu X, Wang Y, An L, Shen D (2019) Structured sparsity regularized multiple kernel learning for Alzheimer's disease diagnosis. Pattern Recognit 88:370–382
Pesapane F et al (2020) Myths and facts about artificial intelligence: why machine- and deep-learning will not replace interventional radiologists. Med Oncol 37(5):1–9
Pineau J, et al (2020) Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program)
Pitropakis N, Panaousis E, Giannetsos T, Anastasiadis E, Loukas G (2019) A taxonomy and survey of attacks against machine learning. Comput. Sci. Rev. 34:100199

Platero C, Lin L, Tobar MC (2019) Longitudinal neuroimaging hippocampal markers for diagnosing alzheimer's disease. Neuroinformatics 17(1):43–61
Poloni KM, Duarte de Oliveira IA, Tam R, Ferrari RJ (2021) Brain MR image classification for Alzheimer's disease diagnosis using structural hippocampal asymmetrical attributes from directional 3-D log-Gabor filter responses. Neurocomputing 419:126–135
Puente-Castro A, Fernandez-Blanco E, Pazos A, Munteanu CR (2020) Automatic assessment of Alzheimer's disease diagnosis based on deep learning techniques". Comput Biol Med 120:103764
Qayyum A, Qadir J, Bilal M, Al-Fuqaha A (2021) Secure and Robust machine learning for healthcare: a survey. IEEE Rev Biomed Eng 14:156–180
Qiu S, Chang GH, Panagia M, Gopal DM, Au R, Kolachalama VB (2018) Fusion of deep learning models of MRI scans, Mini-Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment. Alzheimer's Dement. Diagnosis. Assess. Dis. Monit. 10:737–749
Qiu S et al (2020) Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain 143(6):1920–1933
Rajani NF, Mooney R (2018) Stacking with auxiliary features for visual question answering. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol. 1 (Long Papers), pp 2217–2226
Ramírez J et al (2018) Ensemble of random forests one vs. rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares. J Neurosci Methods 302:47–57
Raschka S (2018) Model evaluation, model selection, and algorithm selection in machine learning
Rashed-Al-Mahfuz M, Haque A, Azad A, Alyami SA, Quinn JMW, Moni MA (2021) Clinically applicable machine learning approaches to identify attributes of chronic kidney disease (CKD) for use in low-cost diagnostic screening. IEEE J Transl Eng Heal Med 9:1–11
Rathore C, Habes S, Iftikhar M, Shacklett MA, Davatzikos A (2017) A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. Neuroimage 155:530–548
Regulation (EU) and 2016/679, European Union (2016) general data protection regulation
Rémi-Cuingnet O, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert M-O, Chupin M, Benali H et al (2011) Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56(2):766–781
Ribeiro MT, Singh MT, Guestrin C (2016) Why should I trust you?': explaining the predictions of any classifier
Richards et al (2018) Bidirectional RNN for medical event detection in electronic health records. Physiol Behav 176(5):139–148
Richhariya B, Tanveer M, Rashid AH, Neuroimaging D (2020) Biomedical signal processing and control diagnosis of alzheimer ' s disease using universum support vector machine based recursive feature elimination (USVM-RFE), vol 59
Riley KP, Snowdon DA, Desrosiers MF (2005) Early life linguisticability, late life cognitive function, and neuropathology: findings from the Nun study. Neurobiol Aging 26(3):341–347
Rojat T, Puget R, Filliat D, Del Ser J, Gelin R, Díaz-Rodríguez N (2021) Explainable artificial intelligence (XAI) on TimeSeries data: a survey
Routier A, Bottani S, Dormont D, Burgos N, Colliot O Convolutional neural networks for classification of alzheimer's disease: overview and reproducible evaluation."
Ruan W, Wu M, Sun Y, Huang X, Kroening D, Kwiatkowska M (2019) Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, vol 1, no 5
S V-L, Van DerVlies R (2021) Trustworthy AI in education: promises and challenges
Sagi O, Rokach L (2018) Ensemble learning: a survey. Wiley Interdiscip Rev Data Min Knowl Discov 8(4):1–18
Saleiro P et al (2018) Aequitas: a bias and fairness audit toolkit. arXiv
Salimi B, Howe B, Suciu D (2019) Data management for causal algorithmic fairness. arXiv1908.07924
Saluja R, Malhi A, Knapič S, Främling K, Cavdar C (2021) Towards a rigorous evaluation of explainability for multivariate time series
Samangouei P, Kabkab M, Chellappa R (2018) Defense-GAN: Protecting classifiers against adversarial attacks using generative models. arXiv1805.06605
Samper-González J et al (2018) Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. Neuroimage 183(July):504–521

Sanchez-Martinez S et al (2019) Machine learning for clinical decision-making: challenges and opportunities. Preprints 2019110278:1–38
Sarraf GTS (2016) DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. J BioRxiv 070441:070441
Schneeberger D, Stöger K, Holzinger A (2020) The European legal framework for medical AI. In: Lect notes comput sci (including subser. lect. notes artif. intell. lect. notes bioinformatics), vol 12279 LNCS, pp 209–226
Seo K, Pan R, Lee D, Thiyyagura P, Chen K (2019) Visualizing Alzheimer's disease progression in low dimensional manifolds. Heliyon 5(8):e02216
Shen HT et al (2021) Heterogeneous data fusion for predicting mild cognitive impairment conversion. Inf. Fusion 66:54–63
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(4):1–31
Shoaip N, Rezk A, El-Sappagh S, Alarabi L, Barakat S, Elmogy MM (2020) A comprehensive fuzzy ontology-based decision support system for alzheimer's disease diagnosis. IEEE Access 9:31350–31372
Shrikumar A, Greenside P, Shcherbina A, Kundaje A (2016) Not just a black box: Learning important features through propagating activation differences. arXiv1605.01713
Singh N, Singh P (2020) Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. Biocybern Biomed Eng 40(1):1–22
Singh A, Sengupta S, Lakshminarayanan V (2020) Explainable deep learning models in medical image analysis. J Imaging 6(6):1–18
Skillen KL, Chen L, Nugent CD, Donnelly MP, Burns W, Solheim I (2014) Ontological user modelling and semantic rule-based reasoning for personalisation of Help-On-Demand services in pervasive environments. Futur Gener Comput Syst 34:97–109
Song Y, Kim T, Nowozin S, Ermon S, Kushman N (2017) Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. arXiv1710.10766
Sørensen L et al (2017) Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry. NeuroImage Clin 13:470–482
Štrumbelj E, Kononenko I (2010) An efficient explanation of individual classifications using game theory. J Mach Learn Res 11:1–18
Su G, Wei D, Varshney KR, Malioutov DM (2016) Interpretable two-level Boolean rule learning for classification
Su D, Zhang H, Chen H, Yi J, Chen PY, Gao Y (2018) Is robustness the cost of accuracy?—a comprehensive study on the robustness of 18 deep image classification models. In: Lect notes comput sci (including subser lect notes artif intell lect notes bioinformatics), LNCS 11216:644–661
Su KSJ, Vargas DV (2019) One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput 23(5):828–841
Suk DSH-I, Lee S-W (2014) Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. Neuroimage 101:569–582
Suk DSH-I, Lee S-W (2015) Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct Funct 220(2):841–859
Suryanarayanan P et al (2020) A canonical architecture for predictive analytics on longitudinal patient records. arXiv
Syed AH, Khan T, Hassan A, Alromema NA, Binsawad M, Alsayed AO (2020) An ensemble-learning based application to predict the earlier stages of alzheimer's disease (AD). IEEE Access 8:222126–222143
Tabarestani S et al (2020) A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. Neuroimage 206:116317
Tanveer M, Richhariya B, Khan RU, Rashid AH (2020) Machine learning techniques for the diagnosis of alzheimer's disease: a review. ACM Trans Multimed Comput Commun Appl 16:1–35
Tatman R, Vanderplas J, Dane S (2018) A practical taxonomy of reproducibility for machine learning research. Rml@Icml 2018
Thiebes S, Lins S, Sunyaev A (2021) Trustworthy artificial intelligence. Electron Mark 31(2):447–464
Tong T, Gao Q, Guerrero R, Ledig C, Chen L, Rueckert D (2017) A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer's disease. IEEE Trans Biomed Eng 64(1):155–165
Toreini E, Aitken M, Coopamootoo K, Elliott K, Zelaya CG, van Moorsel A (2020) The relationship between trust in AI and trustworthy machine learning technologies. In: FAT* 2020—proc. 2020 conf. fairness, accountability, transpar, 272–283

Toreini E et al (2020) Technologies for trustworthy machine learning: a survey in a socio-technical context. arXiv
Tramer F, et al (2017) Fairtest: discovering unwarranted associations in data-driven applications. In IEEE European symposium on security and privacy (EuroS&P), 401 416
Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P (2017) Ensemble adversarial training: Attacks and defences. arXiv1705.07204
Üstün B, Melssen WJ, Buydens LMC (2007) Visualisation and interpretation of Support Vector Regression models. Anal Chim Acta 595(1–2):299–309
Vaithinathan K, Parthiban L (2019) A novel texture extraction technique with T1 weighted MRI for the classification of alzheimer's disease. J Neurosci Methods 318(January):84–99
Varghese T, Sheelakumari R, James JS (2013) A review of neuroimaging biomarkers of Alzheimer's disease. Neurol Asia 18(3):239–248
Varona D, Lizama-Mue Y, Suárez JL (2021) Machine learning's limitations in avoiding automation of bias. AI Soc 36(1):197–203
Vemuri P, Jack Jr R (2010) Role of structural MRI in Alzheimer's disease. Alzheimer's Res Ther 2:23
Verma S, Rubin J (2018) Fairness definitions explained. In: Proc—int conf softw eng, pp 1–7
Vesnic-Alujevic L, Nascimento S, Pólvora A (2020) Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. Telecomm. Policy 44(6):101961
Wang X, Li J, Kuang X, Tan Y, Li J (2019a) The security of machine learning in an adversarial setting: a survey. J Parallel Distrib Comput 130:12–23
Wang H et al (2019c) Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease. Neurocomputing 333:145–156
Wang M, Lian C, Yao D, Zhang D, Liu M, Member S (2019d) Spatial-temporal dependency modeling and network hub detection for functional MRI analysis via convolutional-recurrent network. IEEE Trans Biomed Eng. https://doi.org/10.1109/TBME.2019.2957921
Wang M, Zhang D, Shen D, Liu M (2019e) Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data. Med Image Anal 53:111–122
Wang S, Wang Y, Wang D, Yin Y, Wang Y, Jin Y (2020a) An improved random forest-based rule extraction method for breast cancer diagnosis. Appl Soft Comput J 86:105941
Wang L, Liu Y, Zeng X, Cheng H, Wang Z, Wang Q (2020b) Region-of-Interest based sparse feature learning method for Alzheimer's disease identification. Comput Methods Programs Biomed 187:105290
Wang J, Sun J, Zhang P, Wang X (2018a) Detecting adversarial samples for deep neural networks through mutation testing. arXiv1805.05010
Wang T, Qiu RG, Yu M (2018b) Predictive modeling of the progression of alzheimer ' s disease with recurrent neural networks. Sci Rep, pp 1–12
Wang Y-X, Balle B, Kasiviswanathan SP (2019b) Subsampled Rényi differential privacy and analytical moments accountant. In: The 22nd international conference on artificial intelligence and statistics, pp 1226–1235
Wang D et al. (2021) 'Brilliant AI doctor' in rural China: tensions and challenges in AI-powered CDSS deployment. arXiv:2101.01524v2
Weiner MW et al (2017) Recent publications from the Alzheimer's disease neuroimaging initiative: reviewing progress toward improved AD clinical trials. Alzheimer's Dement 13(4):e1–e85
Wen J et al (2020) Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med Image Anal 63:101694
Wen J et al (2021) Reproducible evaluation of diffusion MRI features for automatic classification of patients with alzheimer's disease. Neuroinformatics 19(1):57–78
Wexler J, Pushkarna M, Bolukbasi T, Wattenberg M, Viégas F, Wilson J (2019) The what-if tool: Interactive probing of machine learning models. IEEE Trans vis Comput Graph 26(1):56–65
Wiens J et al (2019) Do no harm: a roadmap for responsible machine learning for health care. Nat Med 25:1337–1340
Williams JP, Storlie CB, Therneau TM, Jr CRJ, Hannig J (2020) A Bayesian approach to multistate hidden Markov models: application to dementia progression. J Am Stat Assoc 115(529):16–31
Writer ND, Ahmed S, Bajema NE, Bendett S, Chang BA, et al (2019) Artificial intelligence, China, Russia, and the global order technological, political, global, and creative perspectives
Wiśniewski J, Biecek P (2021a) Fairmodels: a flexible tool for bias detection, visualization, and mitigation. 1–15
Wiśniewski J, Biecek P (2021b) Fairmodels: a flexible tool for bias detection, visualization, and mitigation. arXiv2104.00507

Xiao R, Cui X, Qiao H, Zheng X, Zhang Y (2021) Early diagnosis model of Alzheimer's disease based on sparse logistic regression. Multimed Tools Appl 80(3):3969–3980
Xiong P, Buffett S, Iqbal S, Lamontagne P, Mamun M, Molyneaux H (2021) Towards a Robust and trustworthy machine learning system development. J ACM
Xu M, Zhang T, Li Z, Liu M, Zhang D (2021) Towards evaluating the robustness of deep diagnostic models by adversarial attack. Med Image Anal 69:101977
Xu W, Evans D, Qi Y (2017) Feature squeezing: detecting adversarial examples in deep neural networks. arXiv1704.01155
Xu K, et al (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, 2015, pp 2048–2057
Yamanakkanavar N, Choi JY, Lee B (2020) MRI segmentation and classification of human brain using deep learning for diagnosis of alzheimer's disease : a survey. Sensors 20:3243
Yan JN, Gu Z, Lin H, Rzeszotarski JM (2002) Silva: interactively assessing machine learning fairness using causality. In: Conf. Hum. Factors comput. Syst. - proc.
Yang L, Shami A (2020) On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 415:295–316
Yao D, Calhoun VD, Fu Z, Du Y, Sui J (2018) An ensemble learning system for a 4-way classification of Alzheimer's disease and mild cognitive impairment. J Neurosci Methods 302:75–81
Yu T, Zhu H (2020) Hyper-parameter optimization: a review of algorithms and applications. arXiv, pp 1–56
Yuan X, He P, Zhu Q, Li X (2019) Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Networks Learn Syst 30(9):2805–2824
Yue L, Gong X, Li J, Ji H, Li M, Nandi AK (2019) Hierarchical feature extraction for early Alzheimer's disease diagnosis. IEEE Access 7:93752–93760
Zeiler MD, Taylor GW, Fergus R (2011) Adaptive deconvolutional networks for mid and high level feature learning. In: International conference on computer vision, 2011, pp 2018–2025
Zeng N, Qiu H, Wang Z, Liu W, Zhang H, Li Y (2018) A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease. Neurocomputing 320:195–202
Zhang R, Simon G, Yu F (2017) Advancing Alzheimer's research: a review of big data promises. Int J Med Inform 106(July):48–56
Zhang F, Li Z, Zhang B, Du H, Wang B, Zhang X (2019b) Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease. Neurocomputing 361:185–195
Zhang Y, Wang S, Xia K, Jiang Y, Qian P (2021) Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. Inf. Fusion 66:170–183
Zhang X, Han L, Zhu W, Sun L, Zhang D (2021) An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE J Biomed Heal Inform 14(8):5289–5297
Zhang Y, Zhou L (2019) Fairness Assessment for Artificial Intelligence in Financial Industry. arXiv:1912.07211v1
Zhang X, Wei Y, Feng J, Yang Y, Huang TS (2018) Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1325–1334
Zhang JM, Harman M, Ma L, Liu Y (2019a) Machine learning testing: Survey, landscapes and horizons. arXiv
Zhao K et al (2020) Independent and reproducible hippocampal radiomic biomarkers for multisite Alzheimer's disease: diagnosis, longitudinal progress and biological basis. Sci Bull 65(13):1103–1113
Zhao Y, Jiang P, Zeng D, Wang X, Li S (2021) Prediction of Alzheimer's disease progression with multi-information generative. IEEE J Biomed Heal INFORMATICS 25(3):711–719
Zhou J, Liu J, Narayan VA, Ye J (2013) Modeling disease progression via multi-task learning. Neuroimage 78:233–248
Zhu X, Il-Suk H, Lee SW, Shen D (2019) Discriminative self-representation sparse regression for neuroimaging-based alzheimer's disease diagnosis. Brain Imaging Behav 13(1):27–40
Zhu Y, Kim M, Zhu X, Kaufer D, Wu G (2021) Long range early diagnosis of Alzheimer's disease using longitudinal MR imaging data. Med Image Anal 67:101825
Zicari RV, et al. (2021a) On assessing trustworthy AI in healthcare. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Front Hum Dyn, p 30.
Zicari RV et al (2021b) Z—inspection ®: a process to assess trustworthy AI
Zilke JR, Mencía EL, Janssen F (2016) Deepred–rule extraction from deep neural networks. In: International conference on discovery science, pp 457–473

Authors and Affiliations

Shaker El-Sappagh[1,2,3] · Jose M. Alonso-Moral[4] · Tamer Abuhmed[3] · Farman Ali[5] · Alberto Bugarín-Diz[4]

[1] 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] College of Computing and Informatics, Sungkyunkwan University, Suwon, South Korea

[4] Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain

[5] Department of Software, Sejong University, Seoul, South Korea