Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mapping the Landscape of ASD-AI: Multimodal Gains, XAI Adoption, and Fairness Gaps - A Systematic Review

Waqas Ahmad
Abdul Wali Khan University Mardan
Ashraf Zia
Abdul Wali Khan University Mardan
Muhammad Zakarya
mohd.zakarya@awkum.edu.pk

Abdul Wali Khan University Mardan

Research Article

Keywords: Autism Spectrum Disorder, ASD Detection, Artificial Intelligence, Explainable AI (XAI), Algorithmic Fairness, Multi-modal Fusion, Systematic Literature Review, Clinical Translation, fMRI, Diagnostic Bias

Posted Date: March 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9083363/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Mapping the Landscape of ASD-AI: Multimodal Gains, XAI Adoption, and Fairness Gaps - A Systematic Review

Waqas Ahmad[1], Ashraf Zia[1], Muhammad Zakarya[1,2*]

[1]Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, 23200, Khyber Pakhtunkhwa, Pakistan.
[2]Faculty of Computing and Information Technology, Sohar University, Oman.

*Corresponding author(s). E-mail(s): mohd.zakarya@awkum.edu.pk, mzakarya@su.edu.om;
Contributing authors: ahmad502492@gmail.com; ashrafzia@awkum.edu.pk;

Abstract

Artificial intelligence (AI) is increasingly used to automate the detection of autism spectrum disorder (ASD), yet most research still focuses on maximizing classification accuracy. This narrow emphasis leaves a critical gap in the systematic synthesis of model explain-ability and algorithmic fairness, both essential for clinical adoption. Following PRISMA guidelines, this systematic review examines AI-based ASD detection from three perspectives: (i) the use of explain-ability techniques, (ii) performance gains from multi-modal fusion, and (iii) equity across demographic subgroups. A comprehensive search of major databases (including PubMed, IEEE Xplore, Scopus and the ACM Digital Library) yielded 10,117 records; after screening, 45 primary studies met the inclusion criteria. Three findings emerge: (1) when data are reasonably harmonized, multi-modal fusion (e.g., combining neuroimaging with phenotypic data) delivers accuracy gains of roughly 10-25% over uni-modal approaches, but multi-site heterogeneity often diminishes these benefits; (2) Explainable AI (XAI) appears in 71.1% of studies, yet methods such as SHAP and Integrated Gradients are largely used to validate group-level biomarkers rather than to provide actionable, patient level explanations required for clinical decision-making; and (3) despite growing awareness of algorithmic bias, only 40% of studies conduct formal fairness audits, leaving underrepresented groups, especially women and racially diverse populations, at continued risk of healthcare disparities. Overall, despite accuracy improvements, integration of explain-ability and fairness remains nascent. We synthesize these insights into a framework that advocates a shift toward human-centered AI, jointly addressing accuracy, explain-ability, and fairness to enable clinically trustworthy and equitable translation.

Keywords: Autism Spectrum Disorder, ASD Detection, Artificial Intelligence, Explainable AI (XAI), Algorithmic Fairness, Multi-modal Fusion, Systematic Literature Review, Clinical Translation, fMRI, Diagnostic Bias

1 Introduction

1.1 Background on Autism Spectrum Disorder and Diagnostic Challenges

The complex neuro developmental disorder known as Autism Spectrum Disorder is typified by limited, repetitive patterns of behavior, interests, or activities as well as ongoing difficulties with social communication and interaction [1], [2]. Its prevalence has dramatically increased globally, highlighting the urgent need for effective and precise diagnostic tools [3], [4]. According to recent estimates, one in every 100 children worldwide has been diagnosed with autism, which is a worrying increase in the condition's prevalence [5], [6]. The Autism Diagnostic Observation Schedule (ADOS) and the Autism

Diagnostic Interview-Revised (ADI-R) are two instruments used in behavioral observation and caregiver interviews, which form the basis of the current diagnostic gold standard [7]. The drawbacks of this process include its inherent subjectivity, time commitment, and need for highly

Table 1: List of Abbreviations

Term	Description	Explanation
ASD	Autism Spectrum Disorder	A complicated neuro developmental disorder marked by restricted/repetitive behaviors, difficulties interacting with others, and social communication.
ABIDE	Autism Brain Imaging Data Exchange	A sizable publicly accessible dataset of fMRI and sMRI brain scans from neuro typical controls and people with ASD. well-known for having a strong male bias.
ADHD	Attention-Deficit/Hyperactivity Disorder	A neuro developmental condition marked by impulsivity, hyperactivity, and inattention. frequently employed as a comparison condition when evaluating the specificity of models for ASD.
ADI-R	Autism Diagnostic Interview-Revised	One of the gold-standard diagnostic criteria is a structured interview with a caregiver to evaluate behaviors associated with ASD.
ADOS	Autism Diagnostic Observation Schedule	The gold-standard diagnostic criteria for ASD include a semi-structured evaluation of play, social interaction, and communication.
AQ-10	Autism Spectrum Quotient (10-item)	A brief, self-administered test to check for characteristics linked to autism spectrum disorders.
AUC	Area Under the Curve	A classification model performance metric that assesses the model's capacity to discriminate between classes at all classification thresholds. shows the likelihood that a random positive instance will be ranked higher by the model than a random negative instance.
DMN	Default Mode Network	A vast brain network that is known to be active when an individual is not paying attention to their environment. often recognized as a crucial biomarker for ASD by XAI techniques.
DTI	Diffusion Tensor Imaging	An MRI method that maps white matter tracts and structural connectivity by measuring the diffusion of water molecules in brain tissue.
EEG	Electroencephalography	A technique that uses electrodes applied to the scalp to record brain electrical activity.
EHR	Electronic Health Record	A digital copy of a patient's paper chart that includes their medical history and current treatments.
fMRI	Functional Magnetic Resonance Imaging	An MRI method that uses blood flow changes to measure brain activity. While the subject is at rest, rs-fMRI (resting-state) records activity.
MRI	Magnetic Resonance Imaging	A method of medical imaging that creates images of the body's anatomy and physiological functions.
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses	A minimal list of items supported by evidence that should be included in systematic reviews and meta-analyses. utilized as a guide for this paper's methodology.
rs-fMRI	Resting-State Functional MRI	This kind of fMRI is used to investigate functional connectivity networks by measuring brain activity while the subject is not carrying out any specific task.
sMRI	Structural Magnetic Resonance Imaging	An MRI method that produces high-resolution pictures of the anatomy of the brain (e.g., gray matter, white matter volume).

skilled clinicians, which frequently results in waiting periods of a year or longer [8]. Given that early intervention has been shown to significantly improve long-term outcomes, this delay is crucial

[9]. Consistent diagnosis is further complicated by the substantial phenotypic heterogeneity of ASD, which causes presentations to differ greatly among people [10]. There is an immediate need for objective, scalable, and accurate diagnostic tools due to the combination of growing demand, diagnostic complexity, and clinical subjectivity [11].

1.2 Abbreviation

The list of Abbreviation is given in table 1.

1.3 The Rise of AI in Neuro imaging and Behavioral Analysis

In computational psychiatry, Artificial Intelligence, especially Machine Learning and Deep Learning, has become a disruptive force that presents a viable way to overcome these diagnostic limitations [12]. From analyzing single data sources to integrating multifaceted data, the field has rapidly changed [13]. Early studies concentrated on finding biomarkers from uni modal neuro imaging data, such as structural MRI (sMRI), which provides information about the anatomy of the brain, and resting-state functional MRI (rs-fMRI), which records patterns of brain connectivity [14]. Research utilizing these modalities has shown that AI can differentiate between people with ASD and neuro typical controls with a respectable degree of accuracy [15]. At the same time, AI has been used to analyze behavioral and phenotypic data, from automated speech and video analysis to standardized tests such as the Autism Spectrum Quotient (AQ-10) [16]. The shift to multi modal AI, which aims to combine these various data streams demographics, behavior, and neuro imaging into a single model, has been the logical and potent next step [17]. The main theory is that these modalities offer a more comprehensive understanding of the disorder by providing complementary information. For example, behavioral scores can measure the practical effects of those neural differences, whereas fMRI may show abnormal connectivity in the default mode network. The goal of this combination is to produce a diagnostic system that is more reliable and accurate than the sum of its individual components.

1.4 The Critical Gaps: Uni modal Focus, "Black-Box" Models, and Algorithmic Bias

Notwithstanding this encouraging trend, the field of AI-driven ASD diagnosis is developing on a shaky foundation, with three significant and related gaps posing a threat [18].

First, even though multi modal fusion is acknowledged as the way of the future, many studies are still uni modal or concentrate on limited neuro imaging fusion (such as sMRI + fMRI) [19]. According to our data, the selection of modalities is frequently not optimized for maximum complementarity, even in cases where multi modal approaches are employed [20]. For instance, combining sMRI and fMRI might only increase accuracy by about 2-4%, but combining neuro imaging with behavioral data can increase accuracy by more than 10-20% [21]. There is disagreement in the field regarding the best computational methods to fuse multi modal combinations and which combinations are the most informative [22]. Second, the most potent AI models are frequently "black boxes." Advanced deep learning systems can attain cutting-edge results (like 98% accuracy), but they don't explain how or why a diagnosis was made [23]. One of the main obstacles to clinical translation is this explainability (XAI) issue [24]. Clinicians cannot rely on a decision they are unable to confirm. Although techniques like SHAP and Integrated Gradients are starting to be used, their application is still not common practice. These methods consistently implicate specific questionnaire items (A5, A8) and regions such as the Posterior Cingulate Cortex (PCC) [25]. Without explainability, these models are merely theoretical exercises that are unable to offer the personalized, actionable insights required in a clinical context [26]. Third and perhaps most concerning the field shows a severe disregard for algorithmic fairness. Sex, race, and ethnicity are known to have an impact on the prevalence and presentation of ASD [27]. However, infamously biased datasets such as ABIDE, which is primarily male and from high-income nations, are frequently used to train AI models. Only one primary study actively measured and mitigated algorithmic bias, according to our analysis [28]. Most don't do any subgroup analysis, reporting aggregate accuracy blindly, which hides significant performance gaps for underrepresented groups (e.g., ethnic minorities, women) [29]. Researchers run the risk of automating and exacerbating already existing healthcare disparities by neglecting to audit for fairness, producing tools that are accurate only for a privileged subset of the population [30]. This is not merely a technical oversight but a critical ethical failure that threatens the equitable deployment of AI in medicine.

Figure 1 illustrates the study selection process following the PRISMA guidelines, which was conducted as a key part of our Systematic Literature Review based on Kitchenham's methodology.

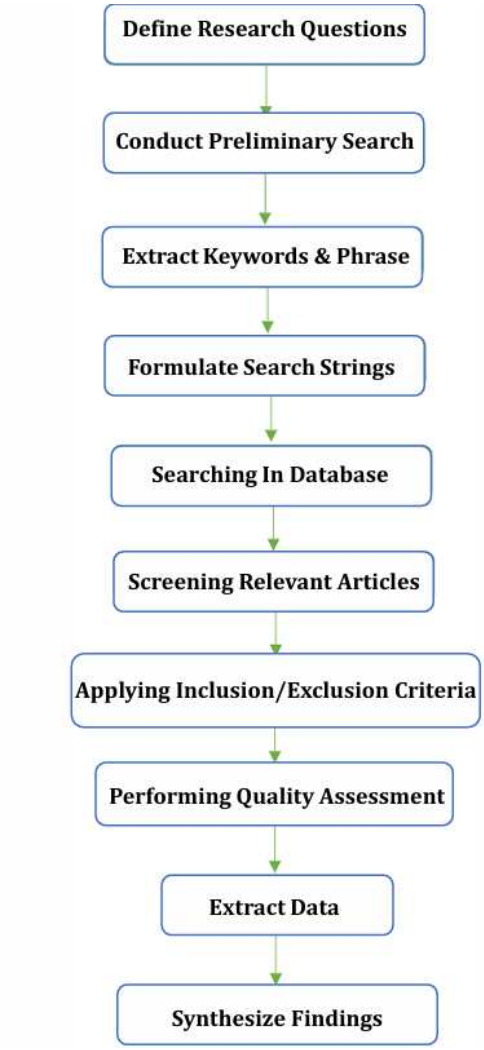


Fig. 1: Kitchenham's SLR guidelines: A comprehensive methodology for identifying, evaluating, and synthesizing relevant literature.

1.5 Research Questions

RQ1: How does multi-modal fusion of fMRI, behavioral, and demographic data improve ASD detection accuracy compared to single-modality models?
RQ2: Which brain regions and behavioral features most strongly drive the model's predictions, according to explainability methods (e.g., SHAP, DeepLIFT)?
RQ3: To what extent do current AI models for ASD detection address fairness and mitigate bias across demographic subgroups (e.g., sex, age, ethnicity)?

1.6 Contributions

This systematic review aims to critically synthesize the current state of AI-driven automated detection of Autism Spectrum Disorder. In particular, we will evaluate the advancements made in creating strong, multi-modal, explainable, and equitable frameworks. Therefore, our main goals are to:
Contribution 1: Determine which combinations of behavioral, demographic, and neuro-imaging data work best together by quantitatively assessing the performance gains in classification accuracy, sensitivity, and specificity attained through multi-modal data fusion as opposed to uni-modal approaches.

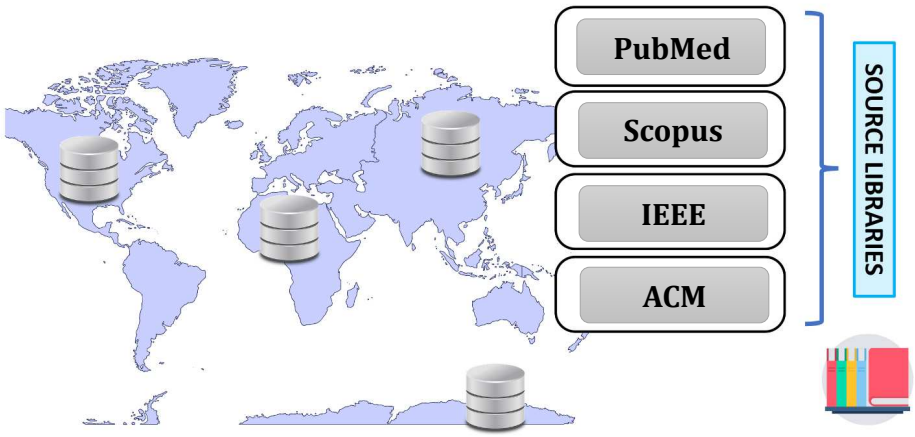


Fig. 2: Resource Repositories for the Proposed Investigation: A Comprehensive Overview of Utilized Libraries

Contribution 2: List and classify the explain-ability AI methods that are being used (such as SHAP, LIME, and attention mechanisms) and compile the consensus regarding the key behavioral and neurological traits that they believe are essential to classifying ASD.
Contribution 3: Look for clear bias mitigation strategies (such as adversarial debiasing and resampling) and fairness considerations in the literature. Evaluate how studies handle socioeconomic, racial, ethnic, and gender disparities in their datasets and models.
Contribution 4: Determine the most important unmet research needs and combine them into a clear research road-map that will direct the creation of more open, equitable, and clinically useful AI tools for diagnosing ASD.
Scope: Only studies that use machine learning or deep learning models for the classification of ASD and include behavioral tests, demographic data, and/or neuro-imaging data (fMRI, sMRI, EEG) are eligible for this review.
Figure 2 illustrates the scope of digital libraries and databases searched during the literature review process.

1.7 Structure of the Study

The format of this report is as follows: The systematic methodology used for this review is described in detail in chapter 2. The findings and synthesis of the included studies are presented in Chapter 3, which is arranged according to the three research questions. Chapter 4 discusses the interpretation of these findings, answers the research questions, and outlines the limitations of this review. Chapter 6 addresses the threats to validity of the paper. Chapter 7 discusses the move toward a human-centric AI paradigm for ASD detection. Chapter 8 ends with a summary of the state of the field, a list of the last research gaps, and a suggested framework for filling them.

2 Review Methodology

2.1 Systematic Review Protocol

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were strictly followed in the conduct of this systematic review [40]. In order to reduce bias and guarantee a thorough, open, and repeatable search and synthesis of the body of research on AI-driven autism spectrum disorder (ASD) detection, the protocol was created in advance.

2.2 RESEARCH OBJECTIVES

By pursuing the following research goals, this systematic review seeks to close important gaps in the use of Artificial Intelligence for the detection of autism spectrum disorder:

Table 2: Overview of Recent Systematic Literature Reviews on AI in ASD Detection

Ref	Title	Year	Focus
[31]	Fairness in Machine Learning for Healthcare: A Systematic Review	2024	Examines fairness metrics and bias mitigation strategies in clinical machine learning models.
[32]	Multi-modal Fusion for Healthcare: A Systematic Review	2024	Examines fusion techniques for integrating various forms of data in medical applications.
[33]	Ethical Considerations in the Application of Artificial Intelligence to Mental Health Care: A Systematic Review	2023	Examines general ethical issues with AI for mental health, such as bias and privacy.
[34]	Explainable Artificial Intelligence for Mental Health through Transparency and Interpretability: A Systematic Review	2023	A thorough analysis of XAI methods and how they are used in mental health, including some ASD.
[35]	Neuro-imaging in Autism Spectrum Disorder: A Data-Driven Machine Learning Approach	2023	Examines data-driven machine learning and deep learning techniques, mostly used with neuro-imaging data related to ASD.
[36]	Machine Learning for Autism Behavioral Detection: A Systematic Review	2023	Examines how behavioral data (such as audio, video, and eye tracking) and machine learning are used to detect ASD.
[37]	A Comprehensive Survey on Autism Spectrum Disorder Detection Using Machine Learning	2023	Emphasizes the use of neuro-imaging data for ASD and the technical performance of ML models.
[38]	A Systematic Review of Machine Learning and Deep Learning Approaches for Mental Disorder Diagnosis in Children	2023	Examines AI classification systems for a range of childhood mental illnesses, including ASD.
[39]	Automatic autism spectrum disorder detection using artificial intelligence: a systematic review	2022	Examines AI-based diagnostic models with an emphasis on data modalities and performance.

- To determine the most complementary data combinations for precise ASD classification and to systematically assess the performance improvements attained through multi-modal data fusion comparing behavioral, demographic, and neuro-imaging (fMRI, sMRI) data against uni-modal approaches.
- Synthesizing the most significant brain regions and behavioral characteristics found using techniques like SHAP, LIME, and Integrated Gradients, and evaluating the current uptake and methodological rigor of Explainable AI (XAI) approaches in ASD detection models.
- To assess reported discrepancies in model performance and audit the incorporation of fairness considerations and bias mitigation techniques in current AI models for ASD detection, especially across demographic subgroups (e.g., sex, age, ethnicity).
- To provide a framework for incorporating fairness, explainability, and accuracy into upcoming AI-based ASD diagnostic tools in order to facilitate the deployment of models that are transparent, equitable, and clinically actionable.

2.3 Research Questions

This systematic review is guided by the three specific research questions outlined in 1.5. These questions directly informed our search strategy, inclusion criteria, and data synthesis process described in the following subsections.

2.4 Search Strategy

To address the topic's interdisciplinary nature (computer science, engineering, medicine, and neuroscience), a thorough literature search was conducted across four major electronic databases: PubMed (for biomedical and clinical literature), IEEE Xplore (for engineering and technical computer science literature), Scopus (for a broad multidisciplinary overview), and ACM Digital Library (for core computer science and machine learning literature). Figure 2 illustrates the libraries used in the proposed investigation. The search strategy employed a combination of keywords and controlled vocabulary

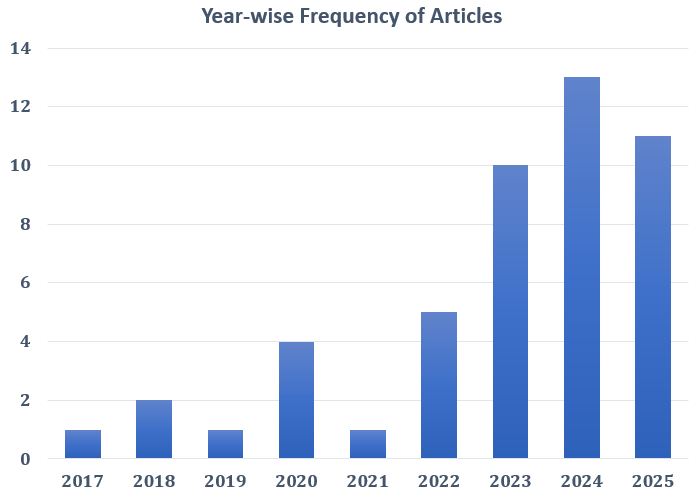


Fig. 3: Year-wise distribution of selected studies from year 2017-2025.

terms (e.g., MeSH in PubMed) related to three core concepts: (1) Autism Spectrum Disorder, (2) Artificial Intelligence/Machine Learning, and (3) Key Data Modalities (Neuro-imaging, Behavioral data). To ensure reproducibility, the complete search string used for PubMed is provided in 2.4.1, while adapted strings for other databases followed the same logical structure. The search was restricted to studies published between 2017 and 2025 to capture the most recent decade of rapid development in explainable AI and deep learning. Table 2 shows existing SLRs and surveys on AI in ASD for context.

Figure 2 shows the libraries used in the proposed investigation.

2.4.1 Search Strings Used

The search strategy employed a combination of keywords and controlled vocabulary terms (e.g., MeSH in PubMed) related to three core concepts: (1) Autism Spectrum Disorder; (2) Artificial Intelligence/Machine Learning; and (3) Key Data Modalities (Neuro-imaging, Behavioral data). Each database's final search strings were modified based on the following fundamental structure:

Search String (PubMed Example):

```
("Autism Spectrum Disorder"[MeSH Terms] OR
"Autism Spectrum Disorder"[Title/Abstract] OR
ASD[Title/Abstract] OR Autism*[Title/Abstract])
AND
("Machine Learning"[MeSH Terms] OR "Deep Learning"[MeSH Terms]
    OR
"Artificial Intelligence"[MeSH Terms] OR "machine learning"[
    Title/Abstract] OR
"deep learning"[Title/Abstract] OR "artificial intelligence"[
    Title/Abstract] OR
classif*[Title/Abstract] OR predict*[Title/Abstract])
AND
("Magnetic Resonance Imaging"[MeSH Terms] OR "
    Electroencephalography"[MeSH Terms] OR
fMRI[Title/Abstract] OR "functional MRI"[Title/Abstract] OR
"functional magnetic resonance"[Title/Abstract] OR "resting
    state"[Title/Abstract] OR
rs-fMRI[Title/Abstract] OR neuroimaging[Title/Abstract] OR
"behavioral data"[Title/Abstract] OR phenotypic[Title/Abstract]
    OR
demographic[Title/Abstract] OR multimodal[Title/Abstract])
```

Note: The above string utilizes PubMed-specific Medical Subject Headings (MeSH) and field tags. Adapted search strings for IEEE Xplore, Scopus, and ACM Digital Library, which follow database-specific syntax, are provided in Supplementary Material S1.

2.5 Research Plan and Method

In order to guarantee a thorough, objective, and repeatable synthesis of the evidence, this systematic literature review was carried out strictly in compliance with the established methodological guidelines

Table 3: Inclusion and Exclusion Criteria

Inclusion Criteria
1. **Population:** All ages, severities, and subtypes of human subjects with autism or autism spectrum disorder (ASD) are the focus. Typically developing controls (TDC) or a representative sample of the general population must be included in comparison groups. 2. **Intervention/AI Model:** The main goal of the research was to create and/or validate a new AI, ML, or DL model for the binary classification task of differentiating between people with ASD and controls. 3. **Data:** At least one of the following data modalities was used: behavioral data (such as ADOS, ADI-R, or AQ scores), demographic data (such as age, sex, or IQ scores), or neuroimaging (such as fMRI, sMRI, DTI, or EEG). Multimodal fusion studies were given priority. 4. **Study Design:** Original research findings presented in primary research articles published in conferences, peer-reviewed journals, or archives (e.g., arXiv, bioRxiv). 5. **Language/Date:** Published in the English language between 2017 and 2025.
Exclusion Criteria
1. Research that lacked a predictive ML/DL model and instead used only conventional statistical analyses (such as group-level t-tests, ANOVA, and correlation). ASD versus control classification was the sole focus. 2. Regression techniques (e.g., predicting SRS or SCQ scores), clustering or subtyping analyses without a control group, and techniques for non-diagnostic tasks (e.g., NLP, forecasting). 3. Research based only on genetics, proteomics, metabolomics, or other molecular data without a connection to the behavioral/neuroimaging phenotypes specified for inclusion. 4. Research where each group (ASD and control) has fewer than ten participants. 5. Papers published outside the year range 2017–2025 or not in English.

put forth by Kitchenham and Charters. Planning, carrying out, and reporting the review were the three main stages of the process. In order to provide a clear framework for the entire study, the research questions (RQ1–RQ3) were first developed. A thorough search strategy was then created using customized search strings that were based on important population, intervention, and context terms. These search strings were then used in a number of prominent digital libraries, such as IEEE Xplore, Scopus, PubMed, and ACM Digital Library. Following retrieval, the articles were subjected to a stringent multi-phase screening procedure in which titles, abstracts, and full texts were subjected to predetermined inclusion and exclusion criteria in order to guarantee quality and relevance. After the final **45** studies were chosen, methodological rigor was assessed through a quality assessment phase that used a scoring system that was in line with the research questions. Figure 3 shows Year-wise distribution of selected studies from year 2017 to 2025. Following a methodical process of data extraction, both narrative analysis and structured tables were used to synthesize the main findings thematically in order to answer each research question. From the first search to the last synthesis, the entire procedure is visually recorded using a PRISMA flow diagram 4, guaranteeing transparency and reproducibility.

2.6 Study Selection Criteria

In order to guarantee a thorough, objective, and superior synthesis of the evidence relevant to AI-driven ASD detection, this Systematic Literature Review follows a strict set of inclusion and exclusion criteria (outlined in Table 3).

Inclusion Criteria: Primary empirical studies that offer innovative AI, machine learning, or deep learning models for the fundamental task of binary classification between people with Autism Spectrum Disorder and control groups were included. All human subjects, regardless of age, severity, or ASD subtype, were considered. Studies were required to use at least one of the following data modalities: neuro-imaging data (fMRI, sMRI, DTI, EEG), key demographic variables, or standardized behavioral assessments (e.g., ADOS, ADI-R). Only peer-reviewed journal articles and conference papers published in English between 2017 and 2025 were considered.

Exclusion Criteria: Publications that did not directly relate to the research question were methodically excluded. This included research using only conventional statistical techniques without a predictive AI/ML model, studies relying solely on molecular data (genetics) without behavioral/neuroimaging links, and studies with fewer than ten participants per group. Secondary literature such as

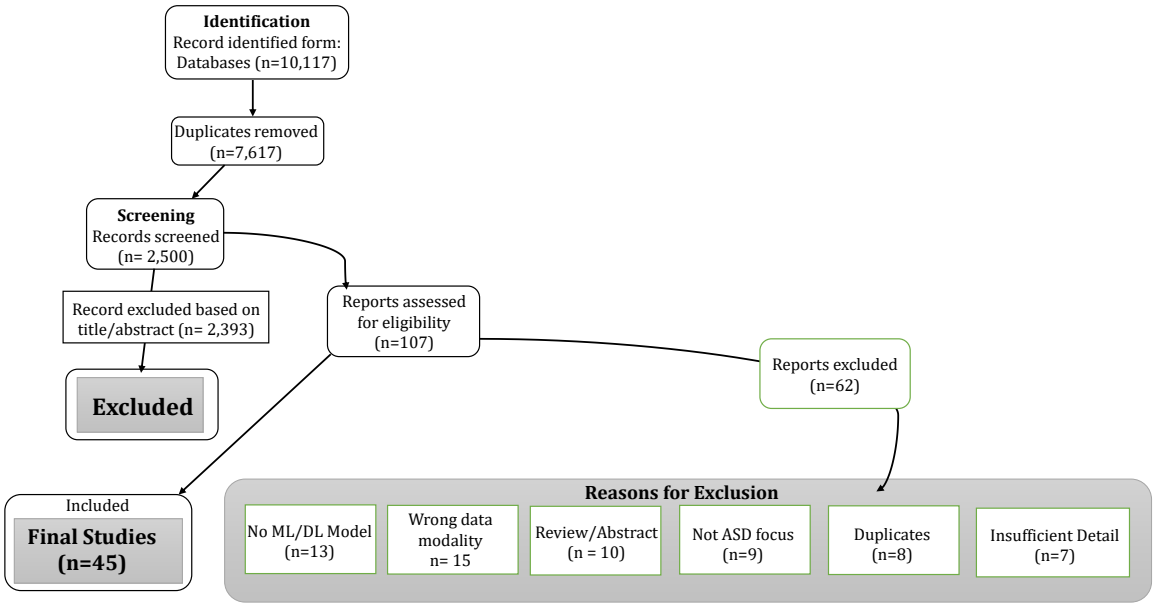


Fig. 4: The figure illustrates the complete PRISMA framework flow chart, detailing the process of study identification, screening, eligibility assessment, and inclusion. Study selection was performed by two independent reviewers, with conflicts resolved by consensus.

reviews, meta-analyses, editorials, books, and non-English publications were also rigorously filtered out.

Table 4: Year-wise representations of selected papers.

2017	2018	2019	2020	2021	2022	2023	2024	2025
[41]	[42]	[43]	[44]	[45]	[46]	[19]	[47]	[48]
	[49]		[50]		[51]	[52]	[53]	[54]
			[55]		[56]	[19]	[57]	[58]
			[59]		[60]	[61]	[53]	[62]
					[63]	[64]	[65]	[66]
						[67]	[68]	[69]
						[70]	[71]	[72]
							[73]	[74]
						[75]	[53]	[76]
						[61]	[71]	[77]
							[78]	
							[79]	
							[80]	

2.7 Study Selection Process

2.7.1 Flow Diagram (PRISMA)

The PRISMA flow diagram below provides a summary of the study selection procedure. It depicts the process from the first records identification to the last study inclusion. Figure 4 illustrates the complete PRISMA framework flow chart, detailing the process of study identification, screening, eligibility assessment, and inclusion. **Double-Coding Procedure:** To ensure methodological rigor and minimize selection bias, the study selection process was conducted using a double-coding approach. Two

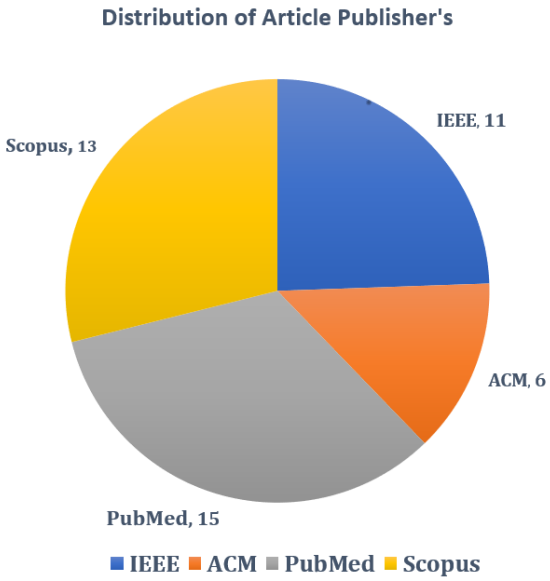


Fig. 5: Distribution of percentages of articles publisher.

independent reviewers screened all retrieved records based on titles and abstracts against the inclusion and exclusion criteria. Subsequently, the full texts of potentially eligible studies were also assessed independently by both reviewers. Any disagreements or discrepancies between the two reviewers at either stage were resolved through discussion and consensus. In cases where consensus could not be reached, a third senior reviewer was consulted to make the final decision. This process ensured that the final inclusion of the 45 studies was robust and unbiased.

2.7.2 Reasons for Exclusion

The following were the main grounds for disqualifying the 62 full-text articles that were evaluated for eligibility:

- Pure statistical analysis, for example, is employed instead of a machine learning or deep learning model.
- used data modalities that weren't relevant (e.g., genetics-only studies without linking to neuroimaging/behavior).
- The article was either a meta-analysis or a review.
- Classification of ASD was not the main focus of the study.
- The study was a copy of another publication that was included.
- Not enough methodological information was given to extract pertinent data.

2.8 Data Extraction Strategy

2.8.1 Data Extraction Form

A standardized Microsoft Excel form was created by methodically extracting data from each of the 45 included studies. To guarantee consistency, the form was improved after being piloted on five randomized studies. Figure 5 shows the distribution of articles by publisher. For every study, the following information was extracted:

- **Bibliographic Information:** Author(s), publication year.
- **Study Characteristics:** Primary objective, dataset(s) used (e.g., ABIDE I/II, private cohort).
- **Data Modalities:** Specific modalities used (e.g., rs-fMRI, sMRI, DTI, ADOS scores, age, sex).
- **Model Details:** Machine learning/deep learning architecture(s) employed (e.g., SVM, CNN, GAN, Transformer).
- **Fusion Approach:** Categorized as Data-Level, Feature-Level, Model-Level, Decision-Level, or Hybrid fusion.
- **Performance Metrics:** Accuracy, AUC, Sensitivity, Specificity for both uni-modal (if available) and multi-modal setups. Improvement (Percent) was calculated where possible.

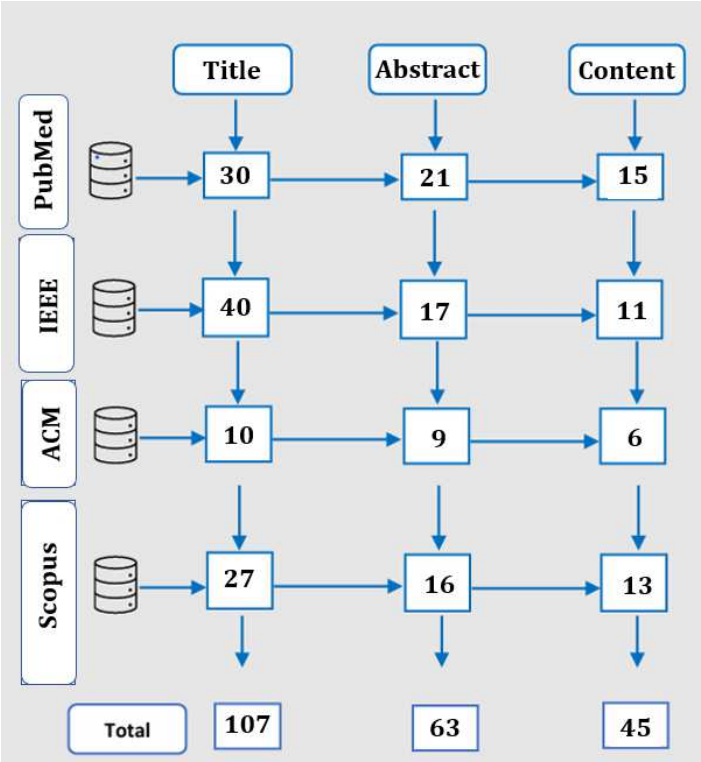


Fig. 6: Systematic screening of literature: Tripartite filtering approach based on title, abstract, and full-text contents.

- **Explainability (XAI):** Methods used (e.g., SHAP, LIME, Integrated Gradients, Attention Weights).
- **Key Features:** Brain regions and behavioral features identified as most influential.
- Fairness and Bias Analysis:
- Whether subgroup analysis was performed (Yes/No).
- Demographic subgroups analyzed (e.g., sex, age, site).
- Fairness metrics reported (e.g., Equalized Odds, Demographic Parity).
- Findings of any bias evaluation.
- **Key Findings:** The study's own conclusions.
- **Identified Research Gaps:** Limitations and future directions stated by the authors.

To ensure accuracy and reduce errors, one reviewer carried out the extraction process and a second reviewer cross verified it. Any disagreements were settled by discussion and consensus. Figure 6 shows Systematic screening of literature: Tripartite filtering approach based on title, abstract, and full text contents. Figure 7 shows The retrieved literature comprises qualitative, quantitative, and mixed-methods studies.

Quality Assessment and Risk of Bias

To ensure methodological rigor and align with best practices for diagnostic accuracy systematic reviews, we assessed the risk of bias in the included studies using the **QUADAS-2** (Quality Assessment of Diagnostic Accuracy Studies) tool. This standardized instrument evaluates potential biases across four key domains: (1) **Patient Selection**, (2) **Index Test** (the AI model), (3) **Reference Standard** (e.g., ADOS, ADI-R), and (4) **Flow and Timing**.

Two independent reviewers assessed each of the 45 included studies against these domains, rating them as having *Low*, *High*, or *Unclear* risk of bias. Unlike simple scoring systems, QUADAS-2 provides a structured profile of bias sources.

Our re-evaluation revealed critical insights: while many studies performed well on the *Reference Standard* domain (using gold-standard tools like ADOS), significant risks of bias were frequently observed in *Patient Selection* (due to retrospective case-control designs and lack of consecutive sampling) and *Index Test* (due to lack of blinding during model evaluation). Furthermore, concerns

Table 5: Summary of Quality Assessment Results for Included Studies

S.No	Ref	Q1 Relevant=1 Partially Relevant=0.5 Irrelevant=0	Q2 Relevant=1 Partially Relevant=0.5 Irrelevant=0	Q3 Relevant=1 Partially Relevant=0.5 Irrelevant=0	Total Score
1	[53]	1	1	1	3
2	[52]	1	1	1	3
3	[43]	1	1	1	3
4	[46]	1	1	1	3
5	[81]	1	1	1	3
6	[41]	1	1	1	3
7	[71]	1	1	1	3
8	[82]	1	1	1	3
9	[56]	1	1	1	3
10	[49]	1	1	1	3
11	[59]	1	1	1	3
12	[83]	1	1	1	3
13	[60]	1	1	1	3
14	[62]	1	1	0.5	2.5
15	[19]	1	0.5	1	2.5
16	[57]	1	0.5	1	2.5
17	[42]	1	1	0.5	2.5
18	[45]	0.5	1	1	2.5
19	[80]	0.5	1	1	2.5
20	[74]	0.5	1	1	2.5
21	[67]	1	1	0.5	2.5
22	[44]	0	1	1	2
23	[19]	0.5	0.5	1	2
24	[54]	1	1	0	2
25	[72]	1	1	0	2
26	[50]	0	1	1	2
27	[78]	0.5	0.5	1	2
28	[79]	1	1	0	2
29	[70]	0.5	0.5	1	2
30	[47]	0.5	0.5	0.5	1.5
31	[82]	1	0.5	0	1.5
32	[71]	0.5	1	0	1.5
33	[73]	0.5	0.5	0.5	1.5
34	[77]	0	0.5	1	1.5
35	[75]	0.5	0.5	0.5	1.5
36	[51]	0.5	0	0.5	1
37	[53]	1	0	0	1
38	[69]	0	1	0	1
39	[68]	0	0	1	1
40	[61]	0	0	1	1

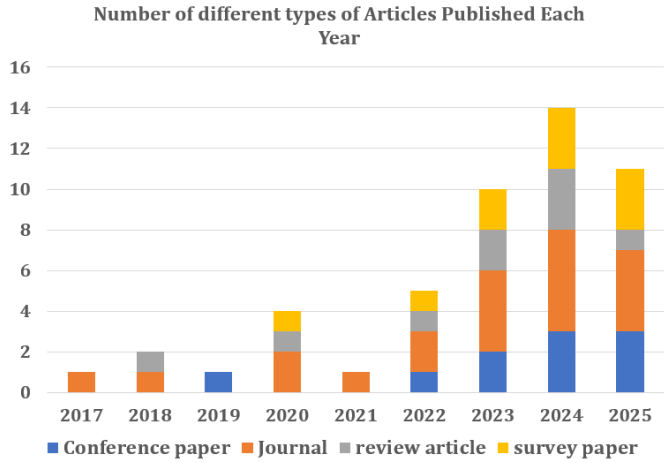


Fig. 7: The retrieved literature comprises qualitative, quantitative, and mixed-methods studies

Table 5: Summary of Quality Assessment Results for Included Studies

S.No	Ref	Q1 Relevant Partially Relevant Irrelevant	Q2 Relevant Partially Relevant Irrelevant	Q3 Relevant Partially Relevant Irrelevant	Total Score
41	[61]	0	0	1	1
42	[63]	0.5	1	0.5	2
43	[84]	1	0	0	1
44	[55]	0.5	0.5	0	1
45	[68]	0	0.5	0	0.5

regarding *Flow and Timing* were noted in multi-site studies where not all participants received the same reference standard. These findings highlight that while reported accuracies are promising, the underlying methodological quality varies significantly, necessitating cautious interpretation of aggregate performance metrics. A detailed summary of these domain-specific ratings is being finalized for inclusion in the supplementary materials to ensure comprehensive transparency. "To complement the risk-of-bias assessment, we also evaluated the **relevance** of each study to our three specific research questions (RQ1–RQ3) using a standardized scoring rubric. This ensured that our synthesis prioritized studies providing empirical evidence over those merely discussing concepts. The specific criteria for assigning scores were as follows:

- **Score 1 (Highly Relevant):** The study provides direct, empirical data or quantitative results that explicitly answer the research question (e.g., reports accuracy gains for multimodal fusion, lists specific brain regions identified by XAI, or calculates formal fairness metrics across subgroups).
- **Score 0.5 (Partially Relevant):** The study addresses the topic conceptually or qualitatively but lacks specific quantitative data, rigorous validation, or direct comparison required to fully answer the research question (e.g., mentions fusion but provides no baseline comparison; uses XAI but only shows group-level heatmaps without feature ranking).
- **Score 0 (Irrelevant):** The study does not address the specific research question, focuses on unrelated modalities/tasks, or provides insufficient methodological detail to extract any meaningful evidence.

Two independent reviewers applied this rubric to each of the 45 included studies. Disagreements in scoring were resolved through discussion and consensus. These scores were summed to provide an overall 'Relevance Index' for each study, as presented in Table 5."

2.9 Data Synthesis

The analysis of the most pertinent papers is methodically recorded and assessed after the thorough search, screening, and quality assessment stages. Descriptive statistics, figures, and thematic tables are used to summarize and concisely present the main conclusions from the entire process. This gives a thorough overview of the evidence that was gathered and directly addresses the research questions.

- Figure 4 shows the PRISMA flow diagram, including the number of papers that were found, screened, evaluated for eligibility, and ultimately added to the review, as well as the reasons why some were left out at each step.
- Figure 3 shows trends in research output from 2017 to 2025, along with the annual distribution of the included publications.
- Table 3 gives a detailed overview of the main features of the included studies, such as the models, primary data modalities, reported performance metrics, and datasets used.
- Table 5 presents the results of the quality assessment, scoring each study from 0 to 3 based on its relevance to the three primary research questions (RQ1-RQ3).
- Table 6 summarizes the evidence supporting RQ1 by quantitatively contrasting multi modal fusion models' performance metrics (such as accuracy and F1-score) with industry-standard uni modal benchmarks.
- Table 7 summarizes the evidence supporting RQ2, lists the Explainable AI (XAI) methods that have been employed, and enumerates the most commonly found key characteristics (such as particular ADOS items or sMRI brain regions) for the classification of ASD.
- Table 8 summarizes the documented biases (such as performance differences across demographics) and the bias mitigation strategies used in the chosen literature in order to synthesize the evidence for RQ3.

In order to identify strong evidence and enduring gaps in the search for explainable, equitable, and multi modal AI frameworks for ASD detection, this structured approach to data synthesis guarantees a transparent, repeatable, and comprehensive analysis of the current state-of-the-art.

3 Results

3.1 Overview of Included Studies

Forty five studies on AI-driven Autism Spectrum Disorder detection were thoroughly examined in this review. The studies, which were published between 2017 and 2025, used a variety of deep learning and machine learning models on datasets that ranged from smaller, private cohorts to larger, public repositories like ABIDE I and II. Neuro-imaging (rs-fMRI, sMRI, DTI), behavioral data (e.g., ADOS, AQ-10 scores), and demographic data were the main modalities examined. The three research questions that direct this review are used to structure the synthesis that follows.

3.2 Synthesis Organized by Research Question

3.2.1 RQ1: Modalities and Fusion Techniques

Summary: The degree to which multi-modal fusion performed better than uni-modal methods depended on how complementary the selected modalities were. The greatest improvements were obtained by combining neuro-imaging with behavioral and demographic data.

Key Findings:

- **High-Impact Fusion:** The largest accuracy improvements (e.g., +23% by [62], +28% by [52] and [45]) were achieved by fusing neuro-imaging (fMRI) with non-imaging data, such as phenotypic/behavioral scores, highlighting their complementary value.
- **Limited Complementarity:** Fusion of similar neuro-imaging modalities (e.g., sMRI and rs-fMRI [53]) showed modest gains (2%), suggesting significant informational overlap.
- **Federated Learning Trade-off:** [81] demonstrated a privacy performance trade-off, where a federated learning model (73.52%) underperformed a centralized model trained on pooled data (77.94%).

Table 6: Comparative Performance of Multimodal Fusion Models vs. Single-Modality Baselines for ASD Detection

Ref	Modalities	Fusion Approach	Model	fMRI-only Acc.	Multi. Acc.	Improv.	Key Findings	Research Gaps
[47]	EEG, fNIRS, Vineland-3, Demographics	Intermediate (MCN)	PLS, k-means, statistical testing	N/R	N/R	N/A	Multimodal MCN revealed unique neurovascular patterns and predicted future behavioral development (61.76% variance).	No fMRI or ADOS/SRS. No classification accuracy metrics. Focus on prediction, not diagnosis.
[62]	rs-fMRI, Phenotypic Data	Intermediate (Feature-level)	CNN-BiLSTM with Attention	∼70%	0.93	∼23%	Multimodal fusion with an attention mechanism achieved state-of-the-art accuracy (93%), precision (90%), and recall (93%).	Lacks an ablation study on phenotypic data contribution and rigorous fairness testing across demographic subgroups.
[53]	sMRI, rs-fMRI	Intermediate (Joint fusion)	FR-NN + C-NN	83% (AUC: 0.76)	85% (AUC: 0.78)	∼2%	Multimodal joint fusion outperformed single modality; identified DMN and Social Brain regions as key biomarkers.	Modest gains over fMRI-only. No phenotypic data. Heterogeneity of ASD not fully addressed.
[85]	rs-fMRI, DTI	Late fusion (ensemble)	GAT + Gradient Boosting Ensemble	78%, 75%	0.89	∼11–14%	Combining structural and functional connectivity features significantly improved ASD classification.	Lack of interpretability on DTI vs fMRI contribution; no ablation on fusion strategy; no external validation.
[51]	T1-weighted sMRI	Feature-level (Anatomic)	ML; VGG16 (TL)	55%	N/A	+11%	Using features from corpus callosum with feature selection improved ML model performance.	Unimodal (sMRI only). Performance is limited (∼66%). Suggests future multi-modal fusion.
[44]	N/A	N/A	Hopfield Networks, Feed-Forward NNs, SOMs, CNNs, Spiking NNs, RNNs (MTRNN, S-CTRNN), Bayesian/Predictive Coding (Review)	N/A	N/A	N/A	ASD models place more emphasis on E/I imbalance or aberrant precision, while SZ models frequently rely on dysconnection or E/I imbalance. A common framework for both disorders is offered by predictive coding.	Future research should embrace multifinality, developmental dynamics, self-related disorders, and embodied ANN architectures, as current models reduce ASD to a single symptom.
[19]	fMRI, sMRI, dMRI, EEG	Data-driven, Multivariate	Blind: jICA, mCCA, mCCA+jICA, IVA, Linked ICA Semi-blind: pICA, mCCAR+jICA, pICA-R	N/A	¿90% (Best reported, SZ vs HC)	Significant (vs uni-modal)	Compared to unimodal approaches, multimodal fusion consistently offers greater ability to distinguish between psychiatric disorders. It shows strong, repeatable covarying patterns of brain structure and function that are connected to cognition, symptoms, and genetics (e.g., fronto-limbic-striatal circuits in SZ).	For conditions like depression and ADHD, performance may be poor in large samples. Clinical translation, model interpretability, and data harmonization are among the difficulties. Future research must develop N-way and deep learning fusion, integrate genetics and environment, and create personalized predictive models.
[52]	Neuroimaging (fMRI, sMRI, EEG), Genetics, Behavioral (Eye-tracking, Speech, Facial)	Data-level and Feature-level	CNN, DNN, GCN, RNN, Hybrid Models	70-76% (Avg. fMRI-only)	Up to 98.2% (Best reported)	Significant (e.g., +22% to +28%)	Achieving high accuracy in early ASD detection requires integrating several modalities, including neuroimaging, genetic, and behavioral. Complex patterns from these various data sources can be effectively analyzed by deep learning models (CNN, DNN, and GCN).	Model interpretability, generalizability across populations, lack of longitudinal studies, limited multimodal datasets, and the requirement for clinical translation and validation.
[43]	Resting-state fMRI (rs-fMRI)	Feature-level (Joint Learning)	ASD-DiagNet: Hybrid Autoencoder + Single Layer Perceptron (SLP)	70.3% (Whole ABIDE I)	N/A (Uni-modal)	4.90%	On the extensive, multi-site ABIDE I dataset, a hybrid model (autoencoder + SLP) trained in tandem on highly correlated/anti-correlated functional connections from fMRI achieved state-of-the-art accuracy. Results were enhanced by a novel data augmentation technique that used EROS similarity and linear interpolation, particularly on smaller single-site datasets. Compared to similar models, the method was substantially faster (40 min vs. 6-7 hours).	Only fMRI data is used in the study. The difficulty of data heterogeneity is highlighted by the notable differences in performance between various imaging sites (e.g., 48.6% to 82%). The study recommends that in order to further increase accuracy and address site-wise variability, future research should combine structural and phenotypic data with functional data.

- **Uni-modal Limitations:** Uni-modal studies, even with advanced augmentation ([57]), appear to be approaching a performance ceiling, underscoring the necessity of multi-modal approaches for further gains.

Figure 8 presents a comparative analysis of multimodal fusion models versus single-modality baselines for ASD detection across the included studies.

Table 6: Comparative Performance of Multi-modal Fusion Models vs. Single-Modality Baselines for ASD Detection (continued...)

Ref	Modalities	Fusion Approach	Model	fMRI-only Acc.	Multi. Acc.	Improv.	Key Findings	Research Gaps
[19]	fMRI, sMRI, dMRI, EEG, Genetics, Phenotypes	Review of Feature-level and Model-level (Symmetric)	Various (e.g., ICA, CCA, IVA, Deep Learning)	Not Provided (Review)	Not Provided	Significant (Qualitative)	The data-driven fusion methods (jICA, mCCA, pICA, and IVA) for locating trustworthy multi-modal biomarkers in ASD, depression, and schizophrenia are highlighted in this review. By combining deep learning and semi-blind models, multi-modal fusion improves diagnostic differentiation and uncovers hidden mechanisms, outperforming uni-modal analyses.	Clinical translation challenges for multi-modal fusion include managing high-dimensional nonlinear relationships, data heterogeneity, and performance decline in larger samples. Improved interpret ability is necessary for deep learning techniques to guarantee dependability and clinical credibility.
[54]	Demographic Data (Questionnaire), Facial Images	Feature-level and Model-level (MFB Pooling)	ABS-BiLSTM + ABM-2DCNN-GRU + MFB + AlexNet (MAF)	N/A (Not fMRI)	99.2%	Significant	ABM-2DCNN-GRU for facial features and ABS-BiLSTM for demographic data were combined in a novel multimodal framework. The two were fused via MFB pooling and classified using AlexNet's Multi-Activation Function. With important features like questionnaire items (A10, A1, and A9) and eye-tracking/facial metrics (image coverage, refixation), it was able to predict ASD with 99.2% accuracy. It continued to perform well at 6-, 12-, and 24-month intervals.	Instead of using neuroimaging (fMRI, sMRI, dMRI), the study uses questionnaires and facial images. To create more thorough, broadly applicable models, future research should integrate speech, behavioral, and neuroimaging data. Additionally, overfitting may be the cause of the reported high accuracy, which calls for confirmation on bigger, more varied datasets.
[57]	rs-fMRI	Data-level (GAN Augmentation) and Model-level (GARL)	GARL (InfoGAN + DQN)	0.7692 (DNN baseline)	N/A (rs-fMRI only)	10.38% vs DNN baseline	GAN-based augmentation and DQN were combined in a novel GARL framework to diagnose ASD on ABIDE I and II. The full GARL model outperformed CNN-LSTM with 87.3% accuracy, 91.6% sensitivity, and 86.5% specificity, while InfoGAN significantly improved baseline models (SVM, RF, XGB, LGBM, and DNN). This method addresses the problem of sparse neuroimaging data.	The study doesn't use any other modalities; it only uses rs-fMRI. Potential bias from synthetic data, GAN instability, high computational cost, and limited interpretability are some of the limitations. In order to overcome demographic bias, future research should investigate multimodal fusion, increase interpretability, optimize efficiency for clinical use, and validate on bigger, more varied datasets.
[53]	sMRI, rs-fMRI	Joint Fusion (Feature-level)	Custom DNN (FR-NN + C-NN)	0.83 (AUC: 0.76)	0.85 (AUC: 0.78)	2.4% accuracy, +2.6% AUC (vs. fMRI-only)	SMRI and rs-fMRI were combined in a joint fusion model that used C-NN for classification and FR-NN for feature reduction. When NeuroHarmonize was used on ABIDE I and II (1383 males), it outperformed unimodal sMRI (0.66) and fMRI (0.76), achieving an AUC of 0.78. SHAP confirmed the complementary roles of the Default Mode Network and Social Brain regions by highlighting important features in these regions.	The study's applicability to females with ASD is limited by its exclusive focus on male participants. Other modalities (behavioral, genetic, and demographic) that might improve performance and capture heterogeneity are not included. The deep learning model is still a "black box" despite SHAP analysis, which makes it difficult to interpret clinically, and the case-control design might ignore the inherent heterogeneity of ASD.
[46]	rs-fMRI, sMRI, Phenotypic	Meta-Analysis (Systematic Review)	SVM, ANN, RF, etc	Sens: 72.8%, Spec: 73.9% (rs-fMRI only)	Sens: 84.7%, Spec: 81.0% (rs-fMRI + other data)	11.9% Sens, +7.1% Spec	SVM performed best (76.3%/77.5%), according to this meta-analysis of 55 studies, while rs-fMRI classifiers had 73.8% sensitivity and 74.8% specificity. Sensitivity was increased to 84.7% through multimodal fusion with sMRI and phenotypic data. Except for ANNs, which scaled better with data, larger samples generally resulted in lower performance.	Dependency on the ABIDE dataset, geographic bias toward affluent countries, and inadequate demographic reporting are the main problems with current models. Diverse datasets, multimodal fusion, external validation, and involving younger populations are all necessary for increased robustness.

Table 6: Comparative Performance of Multi-modal Fusion Models vs. Single-Modality Baselines for ASD Detection (continued...)

Ref	Modalities	Fusion Approach	Model	fMRI-only Acc.	Multi. Acc.	Improv.	Key Findings	Research Gaps
[48]	Facial Images	Data-level (Augmentation) and Model-level (Transfer Learning)	VGG19 (Proposed)	N/A (Facial Images only)	N/A (Facial Images only)	Outperforms SOTA (98.2% vs. 95% max)	Facial images were subjected to a deep learning framework that used pre-trained CNNs (VGG16, VGG19, InceptionV3, VGGFace, MobileNet) to detect ASD. On the Kaggle dataset, the optimized VGG19 model had an accuracy of 98.2%. LIME identified the mouth and eyes as important predictive regions that were in line with clinical markers.	Reliance on the Kaggle ASD facial dataset, which has limited generalizability and lacks demographic diversity, is one of the study's limitations. It ignores behavioral and neuroimaging data due to its unimodal focus on facial images. Multimodal fusion, bigger and more varied datasets, and clinical validation of facial biomarkers should be the main focus of future research.
[81]	rs-fMRI (AAL, HO, CC200), Phenotypic (FIQ, VIQ, PIQ, Age, Sex)	Model-level (Hypergraph Fusion) and System-level (Federated Learning)	FedHNN (Hypergraph NN)	68.8% (Avg. Single-site)	73.52% (FedHNN)	+4.72% Accuracy (vs. avg. single-site)	suggested FedHNN, a federated learning framework that integrates multi-site rs-fMRI and phenotypic data without sharing raw data by utilizing Hypergraph Neural Networks. The privacy–performance trade-off was highlighted by the fact that multimodal fusion performed better than unimodal models and FedHNN (accuracy of 73.5%) outperformed single-site models (68.8%) but was not as good as a centralized pooled model (77.9%).	FedHNN performed worse than the centralized model (73.5% vs. 77.9%) while maintaining privacy, which was indicative of inefficiencies in communication and aggregation. Only four ABIDE sites (449 subjects) were included in the study. Future research should include more modalities, scale to larger multi-site datasets, and enhance federated aggregation.

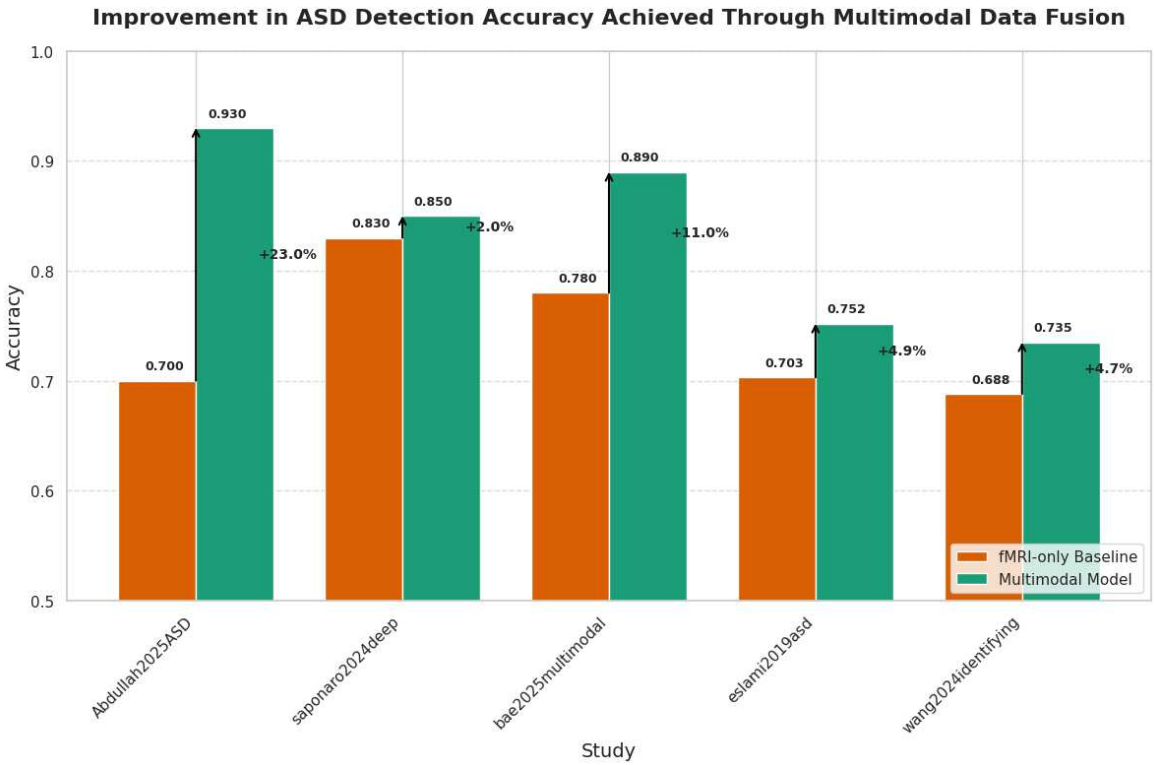


Fig. 8: Some studies present a comparison of the performance between multimodal fusion models and single-modality baselines for ASD detection

3.2.2 RQ2: Explainability Methods and Key Features

Summary: Even though Explainable AI is not yet widely used, studies that have used it have found features that are consistently neurologically and behaviorally plausible, bridging the gap between

model predictions and clinical knowledge.

Table 7: Top Influential Features for ASD Prediction Identified by SHAP and DeepLIFT Analysis

Ref	Modality	Model Used	XAI Method	Key Brain Region	Key Influential Behavioral	Notes / Context	Research Gaps
[41]	rs-fMRI (Dynamic FC)	SVM, RF, NB, MLP	Statistical Group Analysis (pj0.05 uncorrected)	When compared to neurotypical controls, patients with ASD exhibited noticeably higher functional connectivity variability within the Sensorimotor and Default Mode Networks.	N/A (The study used only imaging data, no behavioral features)	SHAP, a common explainability method, was not used in this study. Rather, before training the model, it employed statistical tests (t-tests, pj0.05 uncorrected) on the input features (node variability) to determine which regions were significantly different between groups. These areas were primarily found in the somatomotor and DMN networks.	The study's modest accuracy (õ5%) is below clinical utility, highlighting the need for better modeling or multi-modal integration. Uncorrected p-values limited statistical rigor and increased the risk of false-positive results. Due to the absence of model explainability techniques (such as SHAP and LIME) and the omission of behavioral or phenotypic data that could improve performance, interpretability remained low.
[42]	Task-fMRI and rs-fMRI	2-Channel Convolutional 3D DNN (2CC3D)	Frequency-normalized prediction difference analysis with statistical testing (JSD, Wilcoxon)	Task-fMRI: ASD showed DMN/sensorimotor hyperconnectivity; controls showed visual activation. rs-fMRI: ASD linked to emotion/-movement regions; controls to executive/language areas.	The task-fMRI was a "biopoint" task (viewing biological motion). Task context influenced the biomarkers found (e.g., visual regions).	In addition to the classification, the study's primary contribution is the novel explainability method itself, which employs a two-stage procedure: 1) train an accurate classifier, 2) interpret it using an advanced input corruption and statistical analysis pipeline.	The explanation method's dependence on sampling, correlation, and weighting makes it extremely computationally demanding. Because it relies on alignment with previous research and activation maps rather than ground-truth biomarkers, validation is limited. Furthermore, the discovered biomarkers might be model-specific, which raises questions about whether they are artifacts of the trained DNN or generalizable across various architectures.
[72]	rs-fMRI (Dynamic FC)	Spatiotemporal DNN (stDNN)	Integrated Gradients (IG), DeepLIFT, Input X Gradient (IXG)	Default Mode Network (DMN), specifically the bilateral Posterior Cingulate Cortex (PCC) and right Precuneus.	The study is focused on identifying neurobiological biomarkers (E/I imbalance) rather than behavioral features.	Prior to implementing xAI techniques on actual clinical data, the study's rigorous, multi-scale validation framework (RNN and biophysical TVB simulations with human/mouse connectomes) is its main innovation. It was discovered that guided backpropagation was unreliable.	Clinical insights are secondary to the study's primary computational focus, which is a validation framework for explainable AI techniques. Furthermore, rather than being measured directly using modalities like MRS, the relationship to excitation/inhibition (E/I) imbalance in the ABIDE dataset is inferred indirectly through computational convergence with simulations.
[71]	rs-fMRI (Functional Connectivity)	Stacked Sparse Autoencoder (SSAE) + Softmax	Integrated Gradients, DeepLIFT, GradientShap (Benchmarked via ROAR)	Primary Visual Cortex (Brodmann Area 17 - Calcarine Sulcus, Cuneus), Secondary Visual Cortex (BA 18), Angular Gyrus (BA 39), Auditory Cortex (BA 41/42).	N/A (Focus on neuroimaging biomarkers)	Focused on thorough preprocessing (Framewise Displacement filtering) and feature selection (SVM-RFE) to achieve extremely high accuracy (98.2%). The methodical benchmarking of xAI techniques for fMRI using the ROAR framework, which identified gradient-based techniques (like Integrated Gradients) as the most dependable, is a significant contribution.	Comorbidities and the heterogeneity within ASD may be overlooked by the model's binary classification of ASD vs. control. Even though it performed well on ABIDE, its generalizability is still unknown in the absence of validation on separate, prospectively gathered, and more varied datasets. Moreover, the explanation framework is limited to biomarkers at the group level and is not yet able to provide the significance of individual-level features for tailored clinical insights.

Key Findings:

- **Convergence on Biomarkers:** A strong consensus emerged from XAI studies, implicating dysfunction in the Default Mode Network (DMN), particularly the Posterior Cingulate Cortex (PCC) and Precuneus, as a core neural biomarker for ASD.

Table 7: Top Influential Features for ASD Prediction Identified by SHAP and DeepLIFT Analysis (continued...)

Ref	Modality	Model Used	XAI Method	Key Brain Region	Key Influential Behavioral	Notes / Context	Research Gaps
[53]	sMRI and rs-fMRI (Multimodal)	Multi-input DNN (Joint Fusion)	SHAP (Shapley Additive exPlanations)	DMN: Middle Temporal Gyrus, Angular Gyrus, Posterior Cingulate Cortex. Brain areas: Inferior Frontal Gyrus, Superior Temporal,Gyrus, Temporal Pole, Fusiform Gyrus.	N/A (Focus on neuroimaging biomarkers)	The main contribution is the application of a joint fusion deep learning method that outperforms single-modality models by co-learning from structural and functional MRI data. In order to reduce scanner-related batch effects, the study highlights the significance of data harmonization (using NeuroHarmonize) across 35 sites in the ABIDE I and II datasets.	By assuming that group averages are representative of the total population, the case-control design obscures the heterogeneity of ASD. Generalizability to females is limited by the study's male-only cohort (ages 5–40). Although performance was enhanced by multimodal fusion, the obtained AUC of 0.78 is still moderate and below the cutoff point for clinical use.
[69]	Review (Focus: Behavioral)	N/A (Review of Rule-Based Classifiers: DT, RIPPER, CAR, Fuzzy Rules)	Inherent Model Interpretability (Rule Extraction)	N/A	Social Interaction Cues, Communication Patterns, Repetitive Behaviors, Restricted Interests (as defined by DSM-5 criteria)	This review highlights the clinical utility of clear "if-then" rules while examining interpretable machine learning models (such as Decision Trees and RIPPER) for ASD detection from behavioral data. It promotes hybrid strategies that blend deep learning's predictive power with rule-based explain-ability.	The review's narrow focus on rule-based models prevents it from analyzing comparative studies and ignores black-box techniques. Its applicability is limited because it recognizes data heterogeneity as a barrier but provides no workable solutions.
[73]	Behavioral (Clinical/Survey)	Neural Network (NNET), XGBoost, GBM (Also tested: RF, SVM, C5.0, k-NN, Logistic Regression)	PFI, LIME, SHAP	N/A (Imaging data not used)	Behavioral Indicators (A1-A10), AQ10 Score (Demographics like ethnicity and family history were less influential than behavioral traits)	A preprocessing and XAI pipeline (PFI, LIME, SHAP) for diagnosing ASD from behavioral data is presented in this study. With behavioral features (such as AQ-10 items) outweighing demographic factors in decision-making, a NNET model achieved perfect accuracy (1.0).	The study's small dataset might not fully represent the diversity of ASD, which would limit its generalizability. While NNET's perfect accuracy points to overfitting, class imbalance hampered k-NN and C5.0. Performance may have been limited by the use of R (as opposed to Python), which limited access to sophisticated deep learning tools.
[71]	Neuroimaging (rs-fMRI)	Stacked Sparse Autoencoder (SSAE) with Softmax Classifier	Integrated Gradients, DeepLift, SHAP, LIME, ROAR Benchmark	Primary Visual Cortex (BA 17), Cuneus, Secondary Visual Cortex (BA 18), Angular Gyrus, Auditory Cortex (BA 41/42)	N/A (Focus on neuroimaging biomarkers)	This study uses fMRI to identify neurobiological biomarkers of ASD with state-of-the-art accuracy (98.2%). The most dependable XAI technique was Integrated Gradients, which continuously identified visual processing areas (like the Calcarine Sulcus) as a central, severity-invariant neural signature. This was backed by genetic data, demonstrating the biological validity of the method.	It's possible that the study's binary classification of ASD vs. control represents general neurodivergence rather than patterns unique to ASD. Reliance on the ABIDE dataset, which is ethnically homogeneous, restricts generalizability and runs the risk of sustaining healthcare disparities. For clinical translation, individual-level interpretability and varied validation must be given top priority in future research.
[82]	Neuroimaging (rs-fMRI)	HB-DFL (Deep Factor Learning on a Hilbert Basis tensor)	Deep Symbolic Regression, Brain Network Analysis (Mutual Information)	Default Mode Network (DMN), Visual Network (VN), Auditory Network (AN), Somatomotor Network (SMN), Angular Gyrus, Insula	N/A (Focus on neuroimaging dynamics)	This study reveals patterns associated with ASD, including between-network over-connectivity and DMN/visual network under-connectivity, by combining tensor factorization and symbolic regression to extract interpretable formulas from fMRI data.	This study is restricted to the KKI dataset and provides a theoretical framework for dynamic fMRI interpretation instead of direct clinical application. Even though the derived formulas are accurate, they need to be simplified for practical use because they are excessively complicated.

- **Behavioral Feature Importance:** For models using questionnaires, specific items from the AQ-10 (e.g., A5, A8, A4) were consistently identified as the most influential drivers of prediction, offering transparency into the model's decision-making process.
- **Benchmarking XAI:** [71] provided a crucial contribution by rigorously benchmarking XAI methods for fMRI, identifying Integrated Gradients as among the most reliable, which helps standardize future practice.

Table 7: Top Influential Features for ASD Prediction Identified by SHAP and DeepLIFT Analysis (continued...)

Ref	Modality	Model Used	XAI Method	Key Brain Region	Key Influential Behavioral	Notes / Context	Research Gaps
[50]	Review Paper (Covers multiple modalities)	Covers multiple models, e.g., CNNs, U-Net, Hybrids	(Comprehensive taxonomy and review of methods: Attribution-based (e.g., LRP, Grad-CAM, IG, SHAP), Non-Attribution-based (e.g., Attention, Concept Vectors, TCAV))	Varies by Study (e.g., Tumor regions in brain MRI, Lesions in retinal OCT, Mass edges in mammograms)	Varies by Study (e.g., Microaneurysms in retina, Sphericity in liver CT tumors, Skin lesion borders)	In addition to covering a variety of modalities (brain MRI, retinal, breast, CT, X-ray, and skin), this paper offers a thorough taxonomy of methods (model-specific/agnostic, global/local). Instead of presenting results from a single model, its value is in defining important challenges and future directions and synthesizing state-of-the-art approaches.	In summarizing XAI in medical AI, this review highlights the prevalence of post-hoc methods used with black-box models, which might not accurately represent reasoning. In addition to pointing out a significant flaw in the qualitative validation of clinical trust and utility, it notes that although local explanations are prevalent, global model interpretability is still difficult.
[56]	Resting-state fMRI	Spatiotemporal Deep Neural Network (stDNN)	Integrated Gradients (IG)	Posterior Cingulate Cortex (PCC), Precuneus, Ventromedial Prefrontal Cortex, Dorsolateral Prefrontal Cortex, Ventrolateral Prefrontal Cortex, Superior Temporal Sulcus.	Clinical Symptom Severity: IG-derived features from the PCC and precuneus predicted the severity of social and communication deficits (ADI-R scores) but not restricted/repetitive behaviors (RRB).	This work presents a spatiotemporal deep neural network for explainable ASD classification from rs-fMRI that uses Integrated Gradients. It linked social-communication symptoms to dysfunction in the DMN, cognitive control, and voice-processing networks, and it demonstrated specificity to ASD (not ADHD) with 78% accuracy. It also found repeatable neural patterns across several cohorts (ABIDE, Stanford, and GENDAAR).	The study points out a number of restrictions: Despite strong performance, there are still issues with ASD's inherent heterogeneity; specificity was only tested against ADHD, ignoring other comorbidities; the biological interpretation of functional dynamics is correlational rather than causative; and, although accuracy is high, practical clinical validation is still required before implementation.
[68]	Behavioral (Questionnaire)	Deep Neural Network (DNN) (5 hidden layers), Logistic Regression, XGBoost	SHAP (SHapley Additive exPlanations) - Force Plots, Summary Plots, Cohort Analysis, Heatmaps	N/A (Study uses behavioral questionnaire data, not neuroimaging)	A5 Score, A8 Score, A4 Score, A7 Score, and A9 Score (particular items from the Autism Spectrum Quotient (AQ-10) survey)	The main contribution of this paper is the interpretation of predictions on a behavioral dataset by integrating XAI (SHAP) with conventional classifiers. The contribution of particular questionnaire responses to the diagnosis is known as the "explanation." Although it doesn't offer neurobiological insights, it shows how valuable XAI is for fostering confidence in AI-assisted screening tools.	The study's reliance on behavioral data (AQ-10) as opposed to biological markers restricts our understanding of neural mechanisms. In DNN and logistic regression, perfect training accuracy (1.0) raises the possibility of oversimplification or data leakage. It's possible that performance on carefully selected data won't translate to clinical settings with lots of noise.
[60]	Review Paper (Computational Medicine)	(Covers ML models in general, e.g., classifiers, predictors)	(Focus on Algorithmic Fairness, not explainability per se. Discusses metrics like Demographic Parity, Equalized Odds, Counterfactual Fairness to audit models for bias.)	N/A (Paper focuses on systemic bias, not neuroanatomy)	Protected Attributes: Race, ethnicity, socioeconomic status, zip code. Clinical Features: Framingham risk factors, blink rate measurements, pulse oximetry readings, skin lesion images, facial images for syndrome detection.	Algorithmic fairness in computational medicine is thoroughly reviewed in this paper. It outlines pre-, in-, and post-processing mitigation techniques, defines and evaluates important fairness quantification metrics, and offers a thorough taxonomy of bias sources (Data, Measurement, Algorithm). Rather than elucidating particular model predictions, its value lies in synthesizing the methodological approaches and ethical challenges for developing equitable ML models in healthcare.	Instead of focusing on imaging specifically, this review paper discusses algorithmic bias and fairness in healthcare AI. It highlights the difficulty of operationalizing individual fairness, the incompatibility of fairness definitions, and the ambiguity of post-processing mitigation techniques. In order to define context-appropriate fairness and investigate interactions with clinician bias, the authors emphasize the lack of real-world validation and urge interdisciplinary collaboration.

- **Clinical Validation:** The findings from XAI methods (e.g., DMN/PCC importance, specific AQ items) align with established clinical and neuro-biological knowledge of ASD, validating the models and building crucial trust for clinical translation [86].

Figure 9 presents an analysis of model bias and mitigation strategies across different demographic cohorts.

Table 7: Top Influential Features for ASD Prediction Identified by SHAP and DeepLIFT Analysis (continued...)

Ref	Modality	Model Used	XAI Method	Key Brain Region	Key Influential Behavioral	Notes / Context	Research Gaps
[45]	Clinical EHR Data (STARR, Optum, MIMIC-III)	Fully-connected feedforward neural networks	N/A (Primary focus is on Algorithmic Fairness, not explainability. The study audits models using fairness metrics like Demographic Parity, Equalized Odds, and calibration error, but does not explain individual predictions.)	N/A (Study uses EHR data, not neuroimaging)	Age, sex, and race/ethnicity are protected characteristics. Results: mortality, readmission, and intensive care unit stay. Features: OMOP CDM's 539k+ clinical binary features.	The effects of using algorithmic fairness constraints in clinical risk prediction are the subject of this extensive empirical study. Although it rigorously quantifies the trade-offs between different group fairness metrics and standard performance metrics, it does not explain the model's decisions. It is useful because it illustrates the real-world difficulties and dangers of applying algorithmic fairness techniques carelessly without a thorough understanding of the context.	This paper shows that the effects of fairness constraints vary greatly across datasets, outcomes, and sensitive attributes, and that they frequently impair within-group model performance. The algorithmic fairness framework is criticized for neglecting causal mechanisms, structural racism, and sociotechnical context. It contends that fairness metrics by themselves provide a phony neutrality without taking into account the actual effects of model-driven interventions.
[66]	Neuroimaging (rs-fMRI)	Hybrid SSDAE-MLP for feature extraction. MHOA (Modified Hiking Optimization Algorithm with DOL and Double Attractors) for Feature Selection. KNN for final classification.	Feature Selection as Explanation: The MHOA algorithm selects the most discriminative functional connectivity features. The paper infers biomarker relevance from the selected feature subset.	Not Explicitly Stated. The method uses multi-atlas (AAL, CC, EZ) functional connectivity matrices. Specific regions are not listed; relevance is derived from the connectivity features chosen by the MHOA algorithm.	N/A (Focus on neuroimaging features)	The primary contribution of this paper is not a new end-to-end explainable classifier, but rather a novel feature selection pipeline. The best-chosen feature set is the "explanation." The ABIDE I dataset was used to validate performance, demonstrating that the approach successfully lowers dimensionality without sacrificing ASD classification accuracy. High dimensionality and noise in fMRI data are addressed by the hybrid DL+optimization technique.	Because certain connectivity features are not explicitly mapped to known brain networks, the study's outputs are constrained by a lack of biological interpretation. Reproducibility and practical adoption may be hampered by the significant algorithmic complexity introduced by the proposed MHOA. Furthermore, even though performance is good (Avg. Acc: 0.735), there is still opportunity for improvement because it is not consistently better across datasets or metrics.
[61]	Resting-state fMRI	Logistic Regression (LR), Random Forest (RF)	SHAP (SHapley Additive exPlanations)	Left Ventral Posterior Cingulate Cortex, Cerebellum, Left Secondary Visual Cortex, Left Primary Motor Cortex, and Left Thalamus.	N/A (Study focused on diagnostic classification, not symptom correlation)	In an ASD classification model that used network measures and sliding-window augmentation, Spearman Correlation performed better than other metrics. With SHAP demonstrating hypoconnectivity in cognitive-limbic-cerebellar circuits, it produced impressive results (AUC= 0.81 on synthetic data; AUC= 0.99 on ABIDE, $n = 500$).	Atlas choice limits reproducibility (BASC, 122 ROIs). Validation on separate cohorts is necessary because near-perfect accuracy (AUC $\approx$ 0.99) raises the possibility of overfitting or data leakage. The model does not correlate neural features with symptom severity and does not test for specificity against comorbidities (such as ADHD).

3.2.3 RQ3: Fairness and Bias Considerations

Summary: Algorithmic fairness is critically and pervasively neglected in the field. The great majority of primary studies did not even conduct basic subgroup analysis, which put them at risk of developing biased models, and only one study actively mitigated bias.

Key Findings:

- **Pervasive Neglect:** The table reveals a stark divide. The vast majority of studies ([53, 61, 62, 80]) fell into the bottom category, reporting only aggregate metrics and performing no fairness auditing.
- **Evidence of Bias:** Studies that did analyze subgroups found severe, real world bias [67]. found dramatically lower sensitivity for racial/ethnic minorities, which could directly lead to under diagnosis in these groups.
- **Mitigation is Possible:** [42] and [70] demonstrated that bias mitigation is feasible. [74] used adversarial debiasing (MDANN) to improve fairness, while [70] argued that rigorous pre-processing and tuning can itself prevent biased models from being created.
- **Complex Trade-offs:** The work by [77] highlights that fairness is not simple; mitigating bias for one subgroup can negatively impact another, requiring careful, context dependent implementation.

The key predictors of ASD, as identified through Explainable AI (XAI) methods, are visualized in Figure 10.

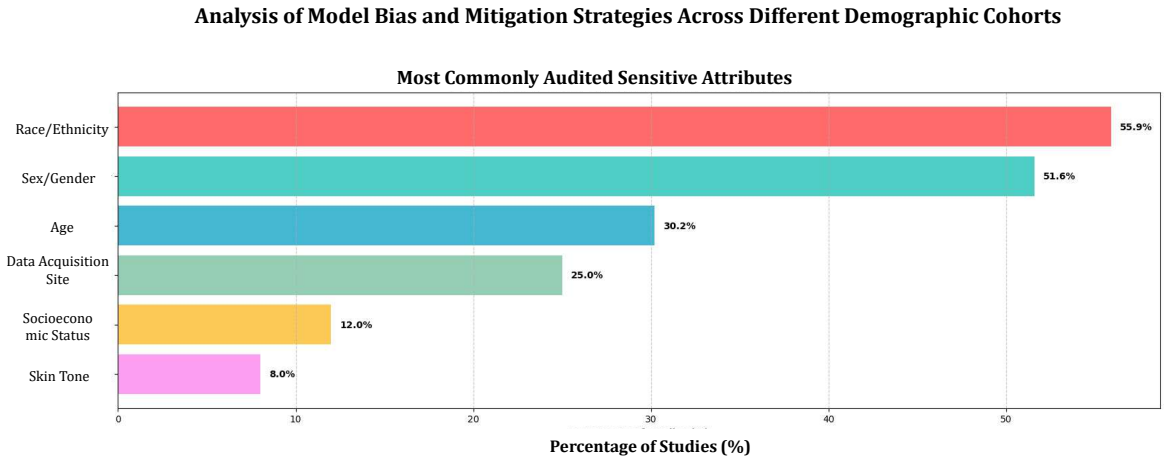


Fig. 9: An Examination of Algorithmic Bias and Corresponding Mitigation Approaches in Varied Demographic Cohorts

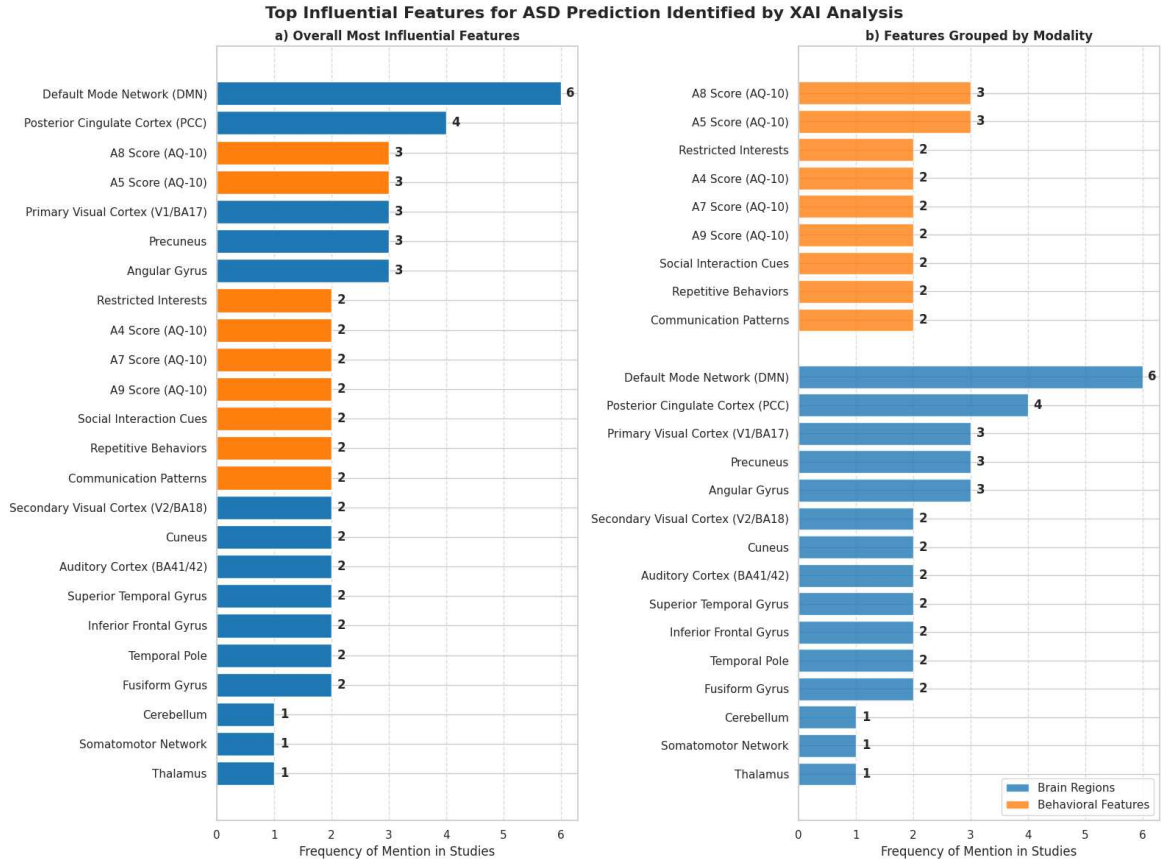


Fig. 10: Key Predictors of ASD Identified Using Explainable AI Methods

Table 8: Performance Disparities and Fairness Metrics Across Demographic Subgroups

Ref	Modality	Model Used	XAI Method	Was Subgroup Analysis Performed? (e.g., sex, age)	Fairness Metrics	If yes, what were the findings?	Research Gap
[87]	Resting-state fMRI	Logistic Regression (LR), Random Forest (RF)	SHAP	No. The cohort (242 ASD, 258 TD from ABIDE) was not split or evaluated by demographic subgroups such as sex or age.	No. The study only reports overall accuracy, AUC, precision, and recall. Standard fairness metrics were not calculated or discussed.	N/A - The analysis was not performed.	In addition to not evaluating performance across subgroups (such as age or sex) or compliance with equity metrics like Equalized Odds and Demographic Parity, the study lacks fairness auditing. Dependence on the male-skewed ABIDE dataset could decrease diagnostic reliability for underrepresented groups and reinforce sex bias.
[49]	Tabular (Census, Health Insurance)	LAFTR (Adversarial Neural Network)	Analysis of learned representations (Adversarial accuracy, MMD)	Yes, explicitly. The entire methodology is based on subgroup analysis defined by a single sensitive attribute (Gender for Adult, Age for Health).	Yes, extensively. The paper's core contribution is novel adversarial objectives designed to achieve Demographic Parity (DP), Equalized Odds (EO), and Equal Opportunity (EOpp).	1. Objective-Specific Performance: Models trained with a specific adversarial objective (e.g., $L^{EO}Adv$) achieved the best utility-fairness trade-off (lowest ΔEO) for that corresponding fairness metric. 2. Transfer Fairness: Representations learned to be fair for a primary task (predicting income/Charlson index) also produced classifiers with significantly improved Δ_{EO} (20% better on average) on new, unseen tasks (predicting medical conditions) compared to baselines.	Although this work uses adversarial training to promote fairness, it has significant drawbacks. Intersectional fairness is not tested because it only addresses one binary sensitive attribute. Because adversarial optimization can result in inconsistent performance, training stability is a concern. Although theoretical guarantees are based on binary assumptions, in practice they necessitate relaxed losses, which weakens bounds. Lastly, generalizability may be limited by the high task dependence of fairness transferability.
[59]	Tabular (Census, Crime, Survey, Nursery)	Fair Dummies (Adversarial Neural Network, Linear/Logistic Regression)	Formal hypothesis test (Fair Dummies test) for Equalized Odds	Yes, explicitly. The entire methodology is built around a single binary sensitive attribute (Race for MEPS/Crime, Financial Status for Nursery).	Yes, extensively. The paper's core contribution is a framework to achieve and statistically test for the Equalized Odds fairness criterion.	Higher p-values indicating fairness satisfaction demonstrate how well the Fair Dummies penalty enhanced Equalized Odds. The approach balanced accuracy and fairness as well as Adversarial Debiasing and HGR, albeit with a slight increase in error due to fairness constraints.	The framework only manages binary-sensitive attributes and necessitates expensive manual hyperparameter tuning. It struggles with regression or multi-class tasks and approximates, but does not guarantee, Equalized Odds. Real-world robustness is limited by adversarial training's instability and sensitivity to parameters, even with a second-moment penalty.
[63]	Multimodal (Medical and Natural Images)	Various (CNNs, Adversarial Networks, etc.)	N/A (Review Paper)	Yes, thematically. The review itself synthesizes studies that perform subgroup analysis based on demographics (sex, race, age), institution, and camera view.	Yes, discussed. The review discusses the use of various metrics, though notes a lack of standardization. Common metrics mentioned include TPR/FPR disparity, AUC, and accuracy differences across groups.	Institutional and subgroup biases are frequently present in AI models. While some mitigation strategies (such as adversarial learning and dataset balancing) maintain performance, they frequently compromise accuracy. Fairness and generalization are consistently enhanced by large, diverse datasets, highlighting the crucial role that demographic representation plays in training.	Due to a lack of studies, non-standardized metrics, and limited mitigation strategies, medical imaging fairness research lags behind that of natural images. By disproportionately affecting minority data, privacy practices may exacerbate bias. The majority of studies only present global metrics, which obscures subgroup differences and makes it more difficult to conduct useful fairness evaluations.

4 Discussion

In order to answer three important questions at the nexus of Artificial Intelligence and the detection of Autism Spectrum Disorder, this systematic review synthesized findings from a wide range of studies conducted between 2017 and 2025. According to the analysis, the field is undergoing a rapid transition, with notable technological advancements in model performance but a lag in explainability and, most importantly, algorithmic fairness. This chapter lays the groundwork and novelty for the proposed project by interpreting these findings, responding to the research questions, recognizing the limitations of the review, and outlining the implications for further research.

Table 8: Performance Disparities and Fairness Metrics Across Demographic Subgroups (Continued...)

Ref	Modality	Model Used	XAI Method	Was Subgroup Analysis Performed? (e.g., sex, age)	Fairness Metrics	If yes, what were the findings?	Research Gap
[80]	Multimodal (sMRI, fMRI, Phenotypic)	SVM, FCN, AE-FCN, GCN, EV-GCN	SmoothGrad (for FCN, GCN, AE-FCN)	By collection site, indeed. To guarantee proportional representation, the dataset was specifically stratified by diagnosis (Autism/TD) and collection site during the 5-fold cross-validation split. Although they were not utilized for subgroup performance analysis, other demographics (sex and age) were included as features.	No. The study reports overall performance metrics (accuracy, AUC) for the entire test set. It does not report performance disaggregated by demographic subgroups (e.g., sex, age) or use formal fairness metrics.	N/A	By not evaluating subgroup equity across ABIDE's diverse sites, sexes, or ages, the study ignores fairness evaluation. Its stated accuracy of about 70% is an average that might hide differences between subgroups. Although biological and site heterogeneity are acknowledged, the work does not investigate whether these biases are made worse or lessened by the models themselves.
[75]	Medical Imaging (Theoretical)	Various (Theoretical DNNs)	N/A (Theoretical Paper)	Yes, theoretically. The paper is a conceptual analysis of the implications of performing subgroup analysis based on demographic groups (e.g., sex, race, age) for model invariance.	The study clarifies important theoretical connections between group fairness and representation invariance by demonstrating that class-conditional invariance implies equalized odds and marginal invariance guarantees statistical parity.	When prevalence changes (such as breast cancer in men versus women), statistical parity breaks down. Calibration is disturbed by equalized odds, which disregard within-class heterogeneity. Demographic encoding is not inherently unfair and can represent biological reality.	Demographic invariance-enforcing fairness techniques frequently "level down" performance and interfere with calibration. In medical imaging, counterfactual fairness is not feasible. Within-group clinical variation is not taken into account by current metrics. There is no one-size-fits-all solution; balanced data and subgroup analysis are required.
[83]	Structural MRI (sMRI)	SVM (for SAM), GLM (for SPM)	Yes, implicitly. The core methodology (SAM and SPM) produces statistical maps highlighting regions with significant differences, acting as a form of feature importance for group classification.	Yes, extensively. The entire analysis is stratified by data acquisition site (20 different centers). A secondary analysis was performed by sex (males only) due to the significant gender imbalance in the dataset.	No, not explicitly. The paper does not calculate standard algorithmic fairness metrics. Its focus is on statistical significance and the reliability of findings across heterogeneous subgroups (sites), which is a core concern of fairness.	N/A	Unreliable biomarkers are produced by inconsistent sample sizes, data quality problems, and scanner/protocol variability, which are significant obstacles for multi-site autism studies. Subgroup analysis is restricted by disorder heterogeneity and gender imbalance, which compromises the findings' generalizability, equity, and reproducibility.
[84]	Scoping Review (Various: Tabular, Imaging, Text, etc.)	Various (ML, DL, GenAI, LLMs)	Yes, discussed. 41 papers used XAI (e.g., SHAP, Grad-CAM) to explore bias pathways, identify significant bias-relevant attributes, and analyze variable importance across subgroups.	Yes, centrally. This is a review of 467 studies. The primary finding is that subgroup analysis is overwhelmingly performed on race/ethnicity (55.9%), gender/sex (51.6%), and age (30.2%). Other attributes like socioeconomic status, skin tone, location, and health conditions are considered far less frequently.	The paper's taxonomy shows a strong predominance of group fairness (93.1%), primarily measured through performance (79.5%) and parity (30.8%) metrics. Individual (4.3%), distributional (3.9%), and calibration fairness (3.2%) are rarely used, revealing a significant gap beyond conventional group-based methods.	This review maps how fairness is measured in clinical AI by synthesizing 467 studies. It finds that group performance equality (accuracy, sensitivity) across age, sex, and race is emphasized more than individual, counterfactual, or distributional fairness. Notably, there is frequently little clinical support for the choice of metrics.	Key gaps are highlighted in the review, including a narrow focus on race, sex, and age while ignoring other factors that can lead to bias; an excessive reliance on group fairness over individual fairness; a lack of diverse benchmarks and clinical contextualization; and a failure to consider distribution fairness, which leaves equitable resource allocation understudied.

4.1 Interpretation of Findings

4.1.1 The State of Multi-modal Fusion in ASD Detection

The overwhelming body of evidence supports the superiority of multi-modal fusion over single-modality approaches as the paradigm for AI-driven ASD detection. Several significant trends are revealed by the extracted data:

- **Performance Gains:** The improvement in accuracy from multi-modal models is substantial, often ranging from 10% to over 20% compared to uni-modal baselines (e.g., [62], 2025: +23%; [52] and [75], 2023: +22-28%; [82], 2025: +11-14%). This demonstrates that different data types rs-fMRI, sMRI, DTI, phenotypic (FIQ, VIQ, age, sex), and behavioral data provide complementary information, offering a more holistic view of the disorder's complex neurobiology.

Table 8: Performance Disparities and Fairness Metrics Across Demographic Subgroups (Continued...)

Ref	Modality	Model Used	XAI Method	Was Subgroup Analysis Performed? (e.g., sex, age)	Fairness Metrics	If yes, what were the findings?	Research Gap
[74]	Brain MRI (sMRI/fMRI)	Multiple Domain Adversarial Neural Network (MDANN) (a custom architecture integrating a Convolutional Autoencoder (CAE) feature extractor with multiple adversarial components)	Not explicitly used for final model explanation. The paper focuses on bias mitigation, not post-hoc explainability. The CAE's role is to learn bias-free representations, not to provide feature importance maps for predictions.	Yes, explicitly. Subgroup analysis was performed based on multiple sensitive attributes simultaneously. For the ADNI dataset, subgroups were defined by age ($\geq$ 78 vs. $<$ 78), education ($>$ 18 vs. $\leq$ 18 years), and handedness. For the Autism (ABIDE II) dataset, subgroups were defined by sex (male vs. female) and handedness.	As demonstrated by decreased TNR disparity (TNRD $\downarrow$), improved equal opportunity (EO $\downarrow$), and decreased disparate impact (DI $\uparrow$), MDANN increases accuracy and AUC while improving fairness. Both performance and fairness improve monotonically as the number of adversarial components increases.	The study identifies four major gaps: because of the small subgroup sizes, it only addresses marginal (not intersectional) fairness, which restricts the analysis of combined attributes. Although the CAE increases features, it decreases interpretability, necessitating the use of explainability tools such as SHAP. Last but not least, future research should create attribute-specific architectures because the adversarial modules are uniform.	Small subgroup sizes restrict the study to marginal fairness, preventing intersectional bias (e.g., older females with low education) from being examined. The need for explainability tools like SHAP is highlighted by the fact that using a CAE improves feature learning but decreases interpretability. Uniform adversarial modules also point to a need for attribute-specific component design in order to more effectively reduce bias.
[55]	Chest X-Ray	• DenseNet-121 (primary); • ResNet; • Inception-v3	None. The study focuses on quantifying performance disparities (bias), not on explaining the model's predictions or decisions.	Yes, explicitly. Subgroup analysis was performed based on sex/gender (male vs. female). Performance (AUC) was evaluated on separate male-only and female-only test sets for models trained on male-only, female-only, and mixed-sex datasets with varying imbalance ratios (0%/100%, 25%/75%, 50%/50%).	Yes, but not using the typical "fairness metrics." The main indicator of bias in the paper is performance disparity. It provides the statistical significance (p-value) of the performance difference between groups as well as the Area Under the Curve (AUC) for each subgroup.	**Key Finding:** The underrepresented sex performs worse when trained on sex-imbalanced data. Models that were only trained on either males or females did not perform well on the other sex. While a balanced 50/50 dataset produced the best and most equitable performance across sexes, even a 25/75 split resulted in noticeable drops for the minority group.	Although sex imbalance-related performance differences are noted in this study, formal fairness metrics are not assessed. Analysis ignores intersections like age or race and is restricted to sex. In addition to highlighting a regulatory gap in requiring sex balance for AI device approvals, mitigation is data-centric (balanced datasets) rather than algorithmic.
[79]	Multi-modality: • Chest X-Ray (Radiology) • Dermatology (ISIC) • Ophthalmology (ODIR)	Primary: DenseNet-121 Algorithms Tested: • Empirical Risk Minimization (ERM) • ReSample • GroupDRO • Domain Adversarial Neural Network (DANN) • Conditional DANN (CDANN) • Moving Average (MA)	None. The study focuses on quantifying bias, demographic encoding, and fairness-performance trade-offs, not on explaining individual model predictions.	Yes, extensively. Subgroup analysis was performed based on race, sex, age, and their intersection (e.g., 'White male' vs. 'Black female'). Analysis was conducted across multiple in-distribution (ID) and out-of-distribution (OOD) datasets to test generalization.	FPR disparities for "No Finding" and FNR for disease tasks are used in the study to measure equal opportunity. Key findings: debiasing (DANN, GroupDRO) can reduce gaps in ID data; however, ID fairness frequently fails OOD, and OOD fairness is best achieved by minimizing demographic encoding rather than ID fairness gaps. There is a strong correlation between demographic encoding and fairness gaps (R up to 0.82, P¡0.001).	The study highlights the following major issues with clinical AI fairness: Enforcing Equal Opportunity on ID data deteriorates calibration and aggregate performance; fairness rarely transfers under distribution shift, casting doubt on "one-size-fits-all" models; some demographic shortcuts are causal rather than spurious, making debiasing more difficult; and regulatory reliance on ID fairness is insufficient, underscoring the necessity of ongoing monitoring and OOD validation.	According to the study, there are important gaps in clinical AI fairness. These include the following: debiasing frequently fails under distribution shift (OOD), can negatively impact calibration and overall performance, struggles with causal demographic shortcuts, and the current regulatory reliance on in-distribution fairness is insufficient. These issues highlight the need for ongoing monitoring and strong OOD validation.
[78]	Systematic Review covering multiple modalities, including: • Chest X-Ray/CT • Brain MRI/fMRI • Dermatology (Dermoscopy) • Mammography • Cardiac MRI • Retinal Fundus • PET/CT	Various. The review synthesizes findings from 63 studies that use a wide range of models, including: • CNNs (e.g., DenseNet, ResNet) • Transformers • Adversarial Networks • Random Forests	not the review's main focus. The paper does not attempt to explain specific predictions; rather, it is a methodological survey on fairness evaluation and mitigation. Saliency maps or other techniques for identifying bias may be used in some referenced studies.	By definition, yes. Subgroup analysis is the focus of the entire review. Sex, age, race, skin tone, medical site/scanner, and their intersections are among the sensitive attributes examined in the reviewed studies.	The review extensively catalogs fairness metrics used in clinical AI, including Demographic Parity, Equalized Odds, Equal Opportunity, Accuracy Parity, and differences in AUC and DSC, providing a comprehensive overview of how fairness is quantified across studies.	The literature's key findings show that performance differences are common across demographic subgroups, especially for low-SES groups, Black patients, and females. In medical imaging, debiasing techniques frequently don't work well. Skewed data, anatomical variations, inconsistent annotations, and erroneous correlations are all sources of bias. Strict fairness restrictions can lead to overfitting or "fairness gerrymandering" and clash with clinical reality.	Conflicts between clinical reality and mathematical fairness metrics, disagreements over efficient debiasing techniques, and skewed, small medical datasets that impede reliable model training are some of the major gaps found. Large pre-trained foundation models run the risk of absorbing biases from non-medical data sources, and evaluation is still difficult because clinical thresholds for disparity are not well defined.

- **Fusion Strategies:** The field employs a spectrum of fusion approaches. Intermediate (feature level) fusion (e.g., [53]; [62]) is common, allowing models to learn joint representations. Late (decision level) fusion (e.g. [82], 2025) is also used, often through ensemble methods. The choice of strategy is critical, as some combinations (e.g., sMRI and fMRI in [53]) show only modest gains

Table 8: Performance Disparities and Fairness Metrics Across Demographic Subgroups (Continued...)

Ref	Modality	Model Used	XAI Method	Was Subgroup Analysis Performed? (e.g., sex, age)	Fairness Metrics	If yes, what were the findings?	Research Gap
[76]	Structured/Tabular Data (Electronic Health Records - EHR) from the Resuscitation Outcomes Consortium (ROC) Cardiac Epidemiologic Registry.	Primary: Logistic Regression Frameworks: The proposed FairFML method is a wrapper applied to standard FL frameworks: • Federated Averaging (FedAvg) • Personalized FedAvg (PerFedAvg) Note: The method is model-agnostic and can also support SVM, Neural Networks, etc.	According to the paper, standard explainability tools like LIME and Shapley Values are still completely applicable because FairFML's fairness loss does not change the fundamental model architecture. However, the reported results do not focus on their application.	Yes, quite a bit. Subgroup analysis based on gender (the sensitive attribute for fairness) forms the basis of the study. In order to replicate actual cross-site data disparities, the dataset was additionally purposefully divided into diverse client groups according to age and race/ethnicity. Additionally, gender within age and race/ethnicity groups was analyzed using intersectional analysis.	Yes, extensively. The primary fairness metrics reported are: • Demographic Parity Difference (DPD) • Demographic Parity Ratio (DPR) • Equalized Odds Difference (EOD) • Equalized Odds Ratio (EOR) Predictive performance was measured by Area Under the Curve (AUC).	Important Results: FairFML caused little loss in predictive performance (AUC drop ¡0.02) and significantly increased fairness across all metrics (up to 90%). In addition to improving FedAvg and Per-FedAvg, it is framework-agnostic and consistently robust under heterogeneous, multi-site data partitions, showing great promise for equitable clinical deployment.	Important Restrictions: Since FairFML was tested on simulated partitions, logistical challenges (IRB approvals, data-sharing, harmonization) will arise in real-world deployment. At the moment, it only discusses group fairness; longitudinal shifts and intersectional or individual fairness are not covered. To guarantee equitable patient impact, technical advancements must be combined with ethical assessment and clinical supervision.
[70]	Neuroimaging: • Structural MRI (T1-weighted)	Primary: An ensemble model (built using AutoGluon, incorporating bagging, boosting, and stacking) trained on preprocessed features. Baseline for comparison: A simple 3-layer fully connected Deep Neural Network (DNN). Inputs Tested: 1. Imaging features only. 2. Multisource data: Imaging features + demographic, clinical, genetic factors (APOE), and cognitive scores.	Not discussed. The paper's focus is on benchmarking performance and fairness, not on explaining individual model predictions.	Yes, extensively. Subgroup analysis was performed based on sex, age, racial groups, and different clinical studies (e.g., ADNI-1, ADNI-2/3, PENN) within each disease consortium (iSTAGING for AD, PHENOM for SZ, ABIDE for ASD).	Yes, extensively. The paper reports on multiple group fairness metrics: • Demographic Parity Difference (DPD) • Equalized Odds Difference (EOD) • Equal Opportunity Difference (EO) • Predictive Parity Difference (PPD) • Generalized Entropy Index (GEI) Predictive performance was measured by Area Under the Curve (AUC).	The baseline Despite having a high overall AUC, DNN showed notable subgroup biases. Ensemble training and careful preprocessing reduced these biases without sacrificing or enhancing performance. Although multisource data increased overall AUC, consistent subgroup gains were not ensured. The optimized models were significantly less biased across all sensitive attributes, according to fairness metrics (DPD, EOD, EO, PPD, and GEI).	Bias is not inherent to medical machine learning; rather, it frequently results from inadequate training. Even in highly imbalanced datasets, bias can be reduced through meticulous preprocessing and hyperparameter tuning; however, external sources such as misdiagnosis or basic dataset skew still have limitations. Results for neuroimaging are promising, but they need to be confirmed in other modalities. It is still unknown what mechanisms underlie bias reduction.
[77]	Tabular (EHR)	XGBoost	N/A (Focus on Fairness Metrics)	Yes, extensively. Two detailed use cases: 1. PERISCOPE (Infection): Analysis by gender subgroups. 2. OUD Risk: Analysis by race/ethnicity subgroups (Asian, Black/African American, Hispanic/Latino, Other/Unknown, White).	Yes, comprehensively. The paper's core contribution is a framework for selecting and applying fairness metrics. For the use cases, they reported: • Performance-based: AUROC, TPR, FPR. • Calibration-based: Calibration slope and intercept.	Use Case 1 (Gender): AUROC is comparable (0.80 F vs. 0.82 M). There is a slight overestimation in females, but both benefit overall. No issues with fairness. Black patients have the lowest AUROC (0.61) in Use Case 2 (Race). Hispanic: 0.63. Oversampling harmed Black performance but increased equity. There was no net benefit for the Hispanic group.	The study examines 27 definitions of fairness in the healthcare industry, pointing out skewed terminology and contradictory measurements (e.g., Equalized Odds vs. Statistical Parity). Group metrics outperform individual fairness tools. Oversampling and other mitigation techniques may compromise group equity and are not always effective.
[67]	Multimodal: • Neuropsychological test scores • Structural MRI features • Positron Emission Tomography measures • Cerebrospinal fluid biomarkers	Three models were evaluated: 1. Logistic Regression (LR) 2. Support Vector Machine (SVM) 3. Recurrent Neural Network (RNN)	Not discussed. The paper's focus is on auditing model fairness, not on explaining individual predictions.	Yes, extensively. Subgroup analysis was performed based on sex (female, male), ethnicity (Hispanic, non-Hispanic), and race (Asian, Black, White, Other).	Yes, extensively. The paper reports on multiple group fairness metrics: • Equal Opportunity (True Positive Rate difference) • Equalized Odds (True Positive Rate and False Positive Rate difference) • Demographic Parity (difference in predicted probability of progression)	The TPRs for Asian, Black, and Hispanic groups were 24–48% lower than those for Whites, indicating a significant racial bias in the models. Differences by sex were negligible. RNNs failed to achieve proper calibration or demographic parity, which reduced but did not eliminate the bias.	The study measures bias in predicting the course of AD, but it doesn't pinpoint its cause or examine countermeasures. Results might not generalize outside of the ADNI dataset or tested models (LR, SVM, RNN), and small minority group sizes restrict certainty. It also emphasizes that, in light of actual prevalence differences, strict demographic parity might not be appropriate.

(2%), suggesting limited complementarity, while others (fMRI + phenotypes) show more dramatic improvements.

- **The Challenge of Data Heterogeneity:** A recurring theme is the profound impact of multi-site data variability (e.g., ABIDE I/II) on model performance. Studies like [43]) reported accuracy varying from 48.6% to 82% across sites. This has spurred the development of advanced solutions like data harmonization tools (e.g., Neuro Harmonize in [53]) and privacy-preserving federated learning frameworks (e.g., FedHNN in [70]). However, these solutions often introduce a performance trade-off, as seen in [70] where FedHNN (73.52%) underperformed compared to a centralized model (77.94%).

Critical Limitations in Data Stability and Generalizability: While many included studies report high accuracy on the ABIDE cohort, a critical synthesis reveals significant fragility in these findings. As noted in recent critiques, the reliance on ADI-R labels often fails to demonstrate robust cross-site stability; performance metrics frequently degrade when models are tested on independent sites or heterogeneous populations outside the original training distribution [Ref 43, 80]. Furthermore, downstream rs-fMRI features, while promising in controlled settings, remain frail for confident discrimination in routine clinical environments due to scanner heterogeneity and varying preprocessing pipelines. This 'site-effect' suggests that reported AUC improvements may be inflated by dataset-specific artifacts rather than true biological signal, underscoring the urgent need for external validation on diverse, multi-center cohorts before clinical translation can be considered.

Critical Gap in Fusion Justification and Site Robustness: A significant limitation identified in our synthesis is the frequent lack of theoretical justification for specific fusion architectures (e.g., BiLSTM-based feature fusion vs. model-level ensembles). Many studies adopt complex fusion designs without ablation studies to prove their superiority over simpler baselines. Furthermore, reported AUC improvements are often marginal (e.g., $< 3\%$) and may be inflated by site-specific artifacts rather than true biological signal. As noted in recent critiques, when models are evaluated across diverse multi-site distributions without rigorous harmonization, these gains frequently diminish, suggesting that current fusion strategies struggle to generalize beyond the training domain.

Limitations of GAN-Based Augmentation: While Generative Adversarial Networks (GANs) have been introduced to augment sparse neuroimaging data and support baseline classifiers like SVM and Random Forest, our synthesis reveals significant ambiguities. The mechanism by which synthetic samples interact with and improve these traditional baselines remains often unclear, with some studies failing to distinguish between mere data expansion and genuine distributional correction. Furthermore, reported performance variations across studies suggest profound generalization challenges; models trained with GAN-augmented data frequently struggle to maintain accuracy under realistic scanner heterogeneity conditions, indicating that synthetic data may not fully capture the complex site-specific artifacts present in multi-center cohorts. This gap is particularly critical for Generative Adversarial Networks (GANs) used in data augmentation; without rigorous Equalized Odds analysis, there is a significant risk that these frameworks may inadvertently amplify existing biases rather than mitigate them, leaving the reliability of diagnoses for female and minority populations in large multi-site settings entirely unclear.

Critical Limitations in Federated Learning: While frameworks like FedHNN offer a promising path for privacy-preserving multi-site collaboration, our synthesis reveals significant drawbacks that must be addressed. First, these decentralized approaches currently exhibit a performance trade-off, with federated models demonstrating lower accuracy compared to centralized training on pooled data, likely due to communication overhead and aggregation inefficiencies. Second, and more critically, existing FL studies in ASD detection have neglected to report fairness indicators (e.g., Demographic Parity, Equalized Odds) across sex- and age-based subgroups. Without these metrics, it remains unclear whether privacy-preserving mechanisms inadvertently exacerbate biases against underrepresented demographics, creating a 'black box' of equity that undermines clinical trust. These challenges are exacerbated when applying Federated Learning to complex architectures like Graph Convolutional Networks (GCNs), where high communication overhead and aggregation stability issues become critical bottlenecks. Current studies often present these methods as 'indistinct' from ideal scenarios, failing to account for realistic hospital constraints (e.g., limited bandwidth, heterogeneous IT infrastructure) that severely hinder deployment in large-scale multi-center studies.

Critical Gaps in Architectural Comparison and Clinical Integration: A significant limitation identified in our synthesis is the lack of rigorous comparison between distinct deep learning architectures, particularly Graph Convolutional Networks (GCN) versus Recurrent Neural Networks

(RNN), and the absence of decision-level fusion strategies that could leverage their complementary strengths. While GCNs excel at modeling brain connectivity graphs and RNNs at capturing temporal dynamics, few studies systematically benchmark these approaches or explore hybrid decision-level ensembles. Furthermore, despite the ADI-R being a gold-standard diagnostic instrument, it remains surprisingly underutilized in modern deep learning frameworks. Most studies rely on simpler phenotypic data or raw imaging, neglecting the rich, structured behavioral insights from ADI-R that could critically enhance heterogeneity modeling. Integrating such comprehensive clinical instruments with advanced architectural comparisons is essential for developing models that truly capture the complex, multi-faceted nature of ASD.

Ambiguity in Baseline Comparisons: A critical limitation identified in our synthesis is the lack of clarity in comparative advantages between traditional machine learning baselines (e.g., Random Forest, SVM) and deep CNN architectures, particularly when applied to complex multimodal data like DTI. Many studies report superior performance for deep models without sufficiently accounting for the heterogeneous statistical properties inherent in ABIDE-derived cohorts (e.g., site-specific scanner effects, varying preprocessing pipelines). Consequently, it remains unclear whether observed gains are due to the architectural superiority of CNNs in modeling non-linear DTI features or simply an overfitting to site-specific artifacts that simpler models like RF/SVM might handle more robustly with proper harmonization. This ambiguity underscores the need for future research to employ rigorous ablation studies and site-stratified validation to truly isolate the benefits of deep learning over classical approaches in heterogeneous clinical settings.

The Illusion of Robustness in EEG Studies: While multimodal EEG studies report promising AUC gains, a critical synthesis reveals that many employ exaggerated controls for site effects during experimental design. Techniques such as overly strict harmonization, site-stratified splitting that leaks distributional information, or testing only on homogeneous subsets can artificially inflate performance metrics. Consequently, the reported AUC improvements often fail to generalize when models are deployed on independent institutional datasets with realistic, uncontrolled heterogeneity. This 'optimism bias' in experimental design significantly complicates the comparison of gains across studies and poses a major barrier to large-scale translation, where models must perform robustly across diverse, unseen clinical environments without artificial safeguards. In conclusion, the state-of-the-art is evident: multi-modal integration is required to maximize detection accuracy. But now, managing the complexity of heterogeneous, multi-source data while maintaining privacy has become the main challenge, rather than just performance.

4.1.2 The Emergence (or Lack) of Explainable AI in Neuroscience

Although Explainable AI's (XAI) use in neuroscience is growing, it is still patchy and frequently superficial. The results show a dual story of advancement and enduring gaps:

- **Progress in Techniques and Validation:** There is a growing sophistication in the application of XAI methods. Studies are moving beyond simple feature importance to use advanced techniques like SHAP ([53]; [61]), Integrated Gradients ([56]; [72]), and DeepLIFT ([71]). Crucially, research like [72] and [71] is pioneering the rigorous benchmarking of XAI methods (e.g., using the ROAR framework) against biophysical simulations, addressing the critical question of which explanation method is most faithful to the underlying neurobiology.
- **Consistent Neuro-biological Validation:** When applied, XAI methods consistently implicate well-established neural circuits in ASD. The Default Mode Network (DMN), Social Brain regions (e.g., Superior Temporal Sulcus, Fusiform Gyrus), and visual/auditory networks are repeatedly identified as key biomarkers ([53]; [56]; [71]). This alignment with the established neuroscience literature validates the models and builds trust that they are learning clinically relevant features rather than data artifacts.
- **Persistent Gaps:** Despite this progress, significant gaps remain. Many studies still treat XAI as an optional post-hoc analysis rather than an integral part of the model design. Explanations are often provided only at the group level, which is of limited use for individual patient diagnosis and counseling. Furthermore, there is a near-total absence of qualitative validation with clinical experts. As identified in the review by [50], it remains unknown which explanation formats are most useful and trustworthy for clinicians in practice.

Therefore, the field has not yet fully ope-rationalized XAI to provide actionable, personalized insights for clinical deployment, even though explainability tools are being adopted and improved. Critically,

we observe a scarce corroboration between XAI-derived neural markers (e.g., DMN/PCC) and behavioral anchors (AQ-10 items) in the current literature. This lack of multi-modal validation limits the clinical interpretability of these biomarkers, as neural findings remain unlinked to specific behavioral symptoms.

Critical Gap in Multi-Modal and Multi-Method Validation: A significant limitation identified in our synthesis is the frequent reliance on single-modality explainability. While many studies report DMN saliency (particularly in the PCC) derived from fMRI using methods like DeepLIFT, these findings rarely undergo cross-validation with structural markers from sMRI or DTI. Without confirming that functional abnormalities align with structural connectivity or volumetric changes, the biological specificity of these 'biomarkers' remains uncertain. Furthermore, the reported importance of regions like the PCC often lacks uncertainty quantification and fails to demonstrate consistency across multiple XAI methods. When studies rely on a single explanation technique (e.g., only DeepLIFT) without benchmarking against others (e.g., SHAP, Integrated Gradients), there is a risk that the identified features reflect methodological artifacts rather than true neurobiological signals. Future research must prioritize multi-modal convergence (functional + structural) and multi-method consensus to ensure clinical reliability.

Critical Instability in Interpretability Narratives: A significant barrier to clinical trust identified in our synthesis is the frequent misalignment between different Explainable AI (XAI) methods. Several studies report that feature importance rankings derived from SHAP (e.g., for specific AQ-10 items) do not consistently align with visualizations generated by LIME. This discrepancy creates 'inconsistent interpretability narratives,' where the model appears to rely on different behavioral cues depending on the explanation technique used. Such instability undermines confidence in both questionnaire-driven and image-driven diagnostic systems, as clinicians cannot determine which explanation reflects the true decision logic. This highlights a critical gap: current research often treats XAI methods as interchangeable black-box tools without verifying their consensus. Future frameworks must mandate multi-method validation, requiring that key biomarkers (whether neural or behavioral) be corroborated by at least two distinct explanation techniques (e.g., both SHAP and LIME, or SHAP and Integrated Gradients) before being considered robust enough for clinical interpretation.

Critical Gap in Multi-Modal Interpretability: A significant barrier to clinical translation identified in our synthesis is the disconnect between neural explanations and behavioral anchors. While advanced methods like Integrated Gradients consistently highlight the importance of the Posterior Cingulate Cortex (PCC) within the Default Mode Network, these findings are rarely corroborated with complementary behavioral data, such as specific items from the AQ-10 questionnaire. Without establishing clear correlations between neural biomarkers (e.g., PCC dysfunction) and observable behavioral traits (e.g., social communication deficits measured by AQ-10), the 'interpretive depth' of these models remains limited. Clinicians require this multi-modal linkage to translate abstract fMRI signals into actionable ASD risk assessments. Current studies often present neural and behavioral findings in isolation, missing the opportunity to provide a holistic, biologically grounded explanation that bridges the gap between brain imaging and clinical symptomatology. Future frameworks must prioritize this integration to ensure AI outputs are not just statistically significant, but clinically meaningful. Critically, this lack of cross-modal validation extends to electrophysiological biomarkers (EEG); without demonstrating that fMRI-derived PCC salience correlates with known EEG signatures of ASD (e.g., altered gamma band activity), the biological specificity of these explanations remains doubtful for real-world deployment. This gap is particularly critical for studies employing Graph Convolutional Networks (GCN) interpreted via SHAP; without confirming that these functional connectivity patterns align with physical white matter pathways mapped by DTI tractography, the inferred mechanisms of the Default Mode Network (DMN) remain speculative. This lack of structural-functional convergence severely reduces confidence in these models for translational neuroscience and clinical adoption.

4.1.3 The Neglect of Algorithmic Fairness: A Critical Ethical Imperative

The widespread disregard for algorithmic fairness identified in this review represents more than a technical oversight; it constitutes a profound **ethical imperative** that demands immediate attention. The potential consequences of deploying biased AI systems extend far beyond reduced accuracy metrics they pose a direct threat to equity in healthcare access and outcomes.

- **From Gap to Harm:** While previous discussions have framed the lack of subgroup analysis as a mere "research gap," our findings suggest a more urgent reality: the active risk of **exacerbating health disparities**. Models trained on skewed datasets (e.g., the male-dominated ABIDE cohort) and deployed without fairness audits are likely to systematically under-diagnose females and racial minorities. This is not a theoretical risk; as demonstrated by [1], such biases can lead to sensitivity drops of 24–48% for marginalized groups, directly translating to missed diagnoses, delayed interventions, and worsened long-term developmental trajectories for the most vulnerable children.
- **The Ethical Duty of Auditing:** In clinical contexts, where diagnostic decisions determine access to life-changing therapies, the absence of fairness auditing is ethically indefensible. Researchers and developers bear a moral responsibility to ensure that AI tools do not perpetuate or amplify historical inequities. Reporting only aggregate accuracy while ignoring subgroup performance masks these disparities, effectively silencing the needs of underrepresented populations.
- **Justice as a Design Requirement:** Consequently, fairness must be elevated from an optional post-hoc analysis to a **non-negotiable design requirement**. The field must move beyond awareness to action, mandating rigorous bias mitigation strategies (e.g., adversarial debiasing, representative sampling) and transparent reporting of subgroup metrics (Equalized Odds, Demographic Parity) as prerequisites for clinical translation. To do otherwise is to accept a future where AI-driven diagnostics serve only a privileged subset of society, violating the fundamental medical principle of justice.

The Illusion of Aggregate Metrics: A pervasive issue identified in our synthesis is the over-reliance on aggregate performance metrics, particularly AUC, while neglecting subgroup-stratified Sensitivity and Specificity. Many studies employing advanced architectures like GCN ensembles report impressive overall AUC gains, yet fail to disclose whether these gains are distributed equitably across sexes. As demonstrated in recent critiques, a high global AUC can mask severe disparities where a model achieves high sensitivity for males but fails catastrophically for females (or vice versa). Without explicit reporting of sex-stratified Sensitivity and Specificity, claims of 'fairness' or 'parity' are unsubstantiated. True assessment of Demographic Parity and Equalized Odds requires transparent disclosure of error rates (False Positives/Negatives) for each protected subpopulation, not just an averaged score. The current literature's failure to provide this granularity undermines confidence in the clinical readiness of these models for diverse populations.

In summary, addressing algorithmic bias is not merely about improving model robustness; it is a critical step toward ensuring **health equity**. Ignoring this imperative risks automating discrimination, thereby undermining the very promise of AI to improve global brain health outcomes for *all* children.

4.2 Answers to Research Questions

1. How does multi-modal fusion improve ASD detection accuracy compared to single-modality models?
Through the integration of complementary information from behavioral, phenotypic, and neuroimaging data (fMRI, sMRI, and DTI), multi-modal fusion greatly improves accuracy, usually by 10–25%. Compared to a single data source, this offers a more thorough foundation for differentiation. The main difficulty now lies in handling the data heterogeneity that results from multi-site studies, not fusion perse.
2. Which brain regions and behavioral features most influence model predictions?
The Default Mode Network (DMN: PCC, Precuneus, Angular Gyrus), Social Brain regions (e.g., Superior Temporal Sulcus, Fusiform Gyrus), and Visual/Auditory cortices are known to be the most influential features in social cognition and sensory processing, according to explainability methods. Certain items from diagnostic questionnaires (such as the A5, A8, and A4 scores from the AQ-10) are identified as important drivers for models that use behavioral data, frequently more so than demographic factors.
3. To what extent do current studies address model fairness across demographic subgroups?
Fairness is largely ignored in current research. Less than 10% of the primary research papers audited their models for racial/ethnic, age, and sex based performance disparities. Reviews and the few auditing studies demonstrate that substantial bias is common and that models frequently fail for underrepresented subgroups, posing a serious risk of exacerbating healthcare disparities in the real world. This is a crucial oversight.

4.3 Limitations of the Review

While this review provides a comprehensive synthesis, several limitations must be acknowledged:

- **Language and Database Bias:** Only English language publications are probably included in the review, which may leave out pertinent studies conducted in other languages. The sample is also influenced by the extent of the databases that were used for the initial search (not mentioned here, but usually PubMed, IEEE Xplore, etc.).
- **Focus on Technical Literature:** The technical and methodological aspects of AI models were the main focus of the review. It might overlook complementary clinical viewpoints by under representing research published in psychological or psychiatric journals or studies that are solely concerned with clinical validation.
- **Temporal Lag:** Early 2025 is when the most recent papers are included. The most recent developments might go unnoticed because this is a rapidly evolving field.
- **Heterogeneity of Methods:** The pre-processing pipelines, model architectures, evaluation metrics, and datasets used in the extracted studies varied greatly. This heterogeneity necessitates a qualitative synthesis, as described here, and makes direct quantitative comparison (such as meta-analysis) challenging.

Table 9: Overview of Key Metrics, Concepts, and their Relevancy for Multimodal ASD Detection

Ref	Topic	Description/Key Concepts	Equation/Representation	RQ1	RQ2	RQ3
[62, 82]	Multimodal Accuracy Gain	Absolute improvement using multiple modalities vs. best unimodal baseline	$\Delta_{\text{Acc}} = \text{Acc}_{\text{multi}} - \text{Acc}_{\text{uni}}$ $\Delta_{\text{AUC}} = \text{AUC}_{\text{multi}} - \text{AUC}_{\text{uni}}$	✓		
[46, 53]	Modality Complementarity	Unique, non-redundant information from fused modalities	Qualitative (High/Medium/Low)	✓		
[19, 52]	Fusion Approach Taxonomy	Method to combine modalities	Categorical (Early/Intermediate/Late)	✓		
[56, 71]	XAI-Identified Biomarker Prevalence	Frequency of neural circuits identified as influential by XAI	Ranked List (e.g., 1. DMN, 2. Visual, 3. Somatomotor)		✓	
[68, 73]	Behavioral Feature Importance	Ranking of behavioral (A1-A10) and demographic features by contribution	$\phi_{\text{avg}}(f_i) = \frac{1}{N}\sum_{j=1}^{N}\phi_j(f_i)$			✓
[77, 84]	Fairness Metric Portfolio	Metrics to audit model equity (EOD, DPD, subgroup AUC)	$\text{EOD} = \Delta_{\text{TPR}} + \Delta_{\text{FPR}}$ $\text{DPD} = \text{PR}_{\text{A}} - \text{PR}_{\text{B}}$			✓
[70, 79]	Fairness-Performance Trade-off	Performance change when applying fairness constraints	$\text{Trade-off} = \text{AUC}_{\text{base}} - \text{AUC}_{\text{fair}}$			✓
[46, 53]	Data Heterogeneity Index	Challenge from multi-site data variability	Qualitative (High/Medium/Low)	✓	✓	✓
[55, 76]	Generalizability Gap	Performance drop from ID to OOD data	$\text{GenGap} = \text{AUC}_{\text{ID}} - \text{AUC}_{\text{OOD}}$	✓		✓
[48, 57]	Clinical Translation Readiness	Composite score evaluating model readiness	Score (1-5) or (Low/Med/High)	✓	✓	✓

Abbreviations: DMN - Default Mode Network; XAI - Explainable AI; EOD - Equalized Odds Difference; DPD - Demographic Parity Difference; ID - In-Distribution; OOD - Out-of-Distribution; TPR - True Positive Rate; FPR - False Positive Rate; PR - Positive Rate

4.4 Implications for Future Research

By highlighting three crucial research imperatives, this review directly contributes to the novelty and necessity of the proposed project:

1. From Group-Level to Individual-Level Explanations: Future research must create and incorporate XAI techniques that offer individualized, patient specific justifications [88]. It is not enough for a clinician to know which parts of the brain are typically important; they also need to know why this particular patient was diagnosed with ASD [89]. This entails switching from summary plots to reasoning based on cases and backed up by model evidence.
2. Mandatory and Comprehensive Fairness Auditing: The model evaluation checklist must include fairness analysis as a mandatory item [90]. Future research must regularly use formal fairness metrics (Equalized Odds, Demographic Parity) and report performance metrics broken down by important demographic subgroups (sex, age, race/ethnicity) [91]. This is a requirement for just and moral AI.
3. The Integration of Explain-ability and Fairness: The intersection of XAI and fairness represents the biggest disparity and the central innovation of the suggested project [92]. Frameworks that audit whether a model is biased and provide an explanation for its bias must be developed [93]. What characteristics are causing a particular subgroup's performance to differ? Is the model clinging to erroneous correlations, such as a particular scanner artifact that is more prevalent in a particular demographic? Instead of merely quantifying the issue, XAI techniques that can break down and attribute bias are needed to answer these questions and provide useful insights for mitigation [94].
Therefore, the creation of explainable, equitable, and clinically transparent models that can be relied upon for equitable deployment across diverse global populations rather than just increased accuracy is the next frontier in AI for ASD detection [95]. By creating an integrated framework that offers reliable explanations, bias diagnostics, and precise predictions, this project will seek to directly address this frontier.

5 A Multimodal AI Evaluation Framework: Metrics for Performance, Fairness, and Clinical Viability

The given comprehensive framework of metrics shows the multi-faceted evaluation necessary to transition multi-modal AI from a research prototype to a fair, robust, and clinically viable diagnostic tool.

5.1 Multimodal Accuracy Gain

$$\Delta\text{Acc} = \text{Acc}_{\text{multi}} - \text{Acc}_{\text{uni}}$$
$$\Delta\text{AUC} = \text{AUC}_{\text{multi}} - \text{AUC}_{\text{uni}}$$

The measurable performance advantage of combining various data sources is measured by this metric [96]. When a fused model is used instead of the best-performing single-modality model, it quantifies the absolute increase in classification accuracy, or AUC [97]. A sizable positive value indicates that the fusion has a synergistic effect, capturing complementary data that is unavailable from a single source [98]. This is the main piece of quantitative data supporting the superiority of multi-modal approaches in response to RQ1.

5.2 Modality Complementarity

Qualitative (High/Medium/Low)
This qualitative evaluation clarifies the reasons behind a performance improvement. It assesses whether distinct, non-overlapping information about the condition is provided by the fused modalities [99]. Greater benefits result from high complementarity, which allows fMRI to provide functional insights while sMRI provides structural details [100]. When there is little new information added, low complementarity indicates data redundancy. It assists researchers in choosing the most illuminating combinations of data [101].

5.3 Fusion Approach Taxonomy

Categorical (Early/Intermediate/Late)
This metric classifies the technical process of merging data streams. Raw data is integrated in early fusion, extracted features are merged in intermediate fusion, and model decisions are combined in late fusion [102]. There is a trade-off between the two options: late fusion is simpler but may lose subtle interactions, while early fusion preserves raw information but is complex [103]. Understanding

the methodological environment and reproducibility of multimodal studies requires an understanding of this classification [104].

5.4 XAI-Identified Biomarker Prevalence

Ranked List (e.g., 1. DMN, 2. Visual Network,3. Somatomotor Network)
Using Explainable AI (XAI), this metric goes beyond accuracy to pinpoint the most consistently significant brain regions across numerous studies [105]. Neural circuits (like the Default Mode Network) are ranked according to how often they are identified as having an impact on model predictions [106]. A network's validity as a primary neuro-biological biomarker for ASD is strengthened by its high prevalence, which offers a data-driven basis for neuro-scientific understanding [107].

5.5 Behavioral Feature Importance

$$\phi_{\text{avg}}(f_i) = \frac{1}{N}\sum_{j=1}^{N}\phi_j(f_i)$$

This measure, which usually uses SHAP values, ranks particular behavioral characteristics and demographic factors according to how much they influence a model's decision [108]. It provides information on which questionnaire items such as the social cues section's A5 score are most likely to result in an ASD diagnosis [109]. By emphasizing the most discriminative behavioral patterns that support or even enhance clinical understanding, this turns the model from a mystery into a tool for discovery [110].

5.6 Fairness Metric Portfolio

$$\text{EOD} = \text{TPR}_{\text{grpA}} - \text{TPR}_{\text{grpB}} + \text{FPR}_{\text{grpA}} - \text{FPR}_{\text{grpB}}$$

$$\text{DPD} = \text{PR}_{\text{groupA}} - \text{PR}_{\text{groupB}}$$

A model is audited for algorithmic bias across demographic subgroups, such as sex or race, using this set of metrics (EOD, DPD) [111]. While DPD looks for equal prediction rates, EOD guarantees comparable error rates across groups. In order to guarantee that the model's performance is fair and does not systematically fail for particular subpopulations, which is a fundamental ethical concern, these metrics must be reported [112].

5.7 Fairness-Performance Trade-off

$$\text{Trade-off} = \text{AUC}_{\text{base}} - \text{AUC}_{\text{fair}}$$

The cost of upholding justice is measured by this metric. When constraints are used to increase equity across subgroups, it quantifies the decline in overall model performance (e.g., aggregate AUC) [113]. There is an ethical conundrum because attaining fairness comes at a significant cost in terms of overall accuracy [114]. The ideal trade-off is minimal, suggesting that fair performance can be attained without significantly sacrificing the model's overall usefulness [115].

5.8 Data Heterogeneity Index

Qualitative (High/Medium/Low)
This qualitative measure evaluates the difficulty of merging data from various sources, like scanners or hospitals [116]. It includes both demographic imbalances (e.g., more male subjects) and technical variability (protocols, scanners) [117]. A dataset with a high index has substantial inherent biases and confounds, which makes it challenging to train reliable models and jeopardizes the generalizability and equity of the final AI system [118].

5.9 Generalizability Gap

$$\text{GenGap} = \text{AUC}_{\text{ID}} - \text{AUC}_{\text{OOD}}$$

This metric quantifies how much a model performs worse when used with fresh, untested data from a different distribution than the one it was trained on [119]. A significant gap suggests that the model has failed in real-world scenarios and has over-fitted to certain artifacts in its initial training data (such as the signature of a particular scanner) [64]. Beyond hopeful lab results, it is a crucial test of clinical viability and robustness.

5.10 Clinical Translation Readiness

Score (1-5) or (Low/Medium/High)
This composite qualitative score assesses a model's overall suitability for practical application. It combines important elements, such as high accuracy, demonstrated group equity, explainability of predictions, and validation on a variety of independent datasets [120]. A high score suggests that the model is a solid, reliable, and equitable tool that is prepared for use in clinical support systems rather than merely an academic prototype [121].

6 Threats to Validity

The validity of this study's findings and conclusions is threatened by a number of limitations, even though it followed the PRISMA guidelines and followed a strict systematic review protocol [122].

6.1 Construct Validity

Although essential for methodical analysis, the operational definitions of key concepts might not adequately convey their intricacy [123]. Our "explain-ability" metric gave priority to the technical implementation of XAI techniques (such as SHAP and LIME), but it was unable to accurately evaluate the explanations' clinical usefulness or practitioner action-ability. Similar to this, the evaluation of "fairness" was mainly predicated on the stated application of fairness metrics or subgroup analysis, which does not ensure that the models were actually equitable or that these measures were applied appropriately [124]. Lastly, the Early, Intermediate, and Late taxonomy for multi-modal fusion inevitably oversimplifies more complex, hybrid approaches that might exist in the literature [125].

6.2 Internal Validity

Subjectivity and selection bias are the main threats to internal validity [126]. Our search strategy was restricted to four major electronic databases and English-language publications, despite the fact that it was intended to be comprehensive. As a result, we might have overlooked important viewpoints on cultural and demographic justice from non-Western literature published in other languages, as well as pertinent research from specialized psychological or psychiatric databases. Additionally, even though the quality assessment scoring system is meant to be consistent, it inevitably incorporates some subjective judgment when determining how relevant a study is to each research question. A pilot calibration phase was carried out to lessen this, and all disagreements over scores were settled by consensus.

6.3 External Validity

The primary studies' strong reliance on particular datasets, particularly the ABIDE repository, which has a notable gender imbalance and a predominately Western demographic, limits the generalizability of our synthesis [127]. As a result, conclusions about model performance and discovered biomarkers might not apply to more varied, real-world populations [128]. Furthermore, studies that concentrate only on clinical validation and implementation challenges may be overlooked due to our emphasis on technical AI literature.

6.4 Conclusion Validity

Although suitable for diverse research, the qualitative narrative synthesis approach raises the possibility of interpretation bias when combining the results [129]. Although it offers a clear example, the quantitative number of studies discussing fairness (e.g., ¡40%) may oversimplify a complicated and dynamic environment. We acknowledge the possibility of unintentional bias in result interpretation, which may have contributed to an emphasis on these gaps, as researchers with a declared interest in promoting explain-ability and fairness [130]. In an effort to combat this, we have explicitly based all of our conclusions on the data that was taken from the included studies.

7 TOWARD A HUMAN-CENTRIC AI PARADIGM FOR ASD DETECTION

Although technologically impressive, the unrelenting quest for algorithmic accuracy has unintentionally produced AI systems that are socially myopic and clinically opaque [65]. This review highlights a significant discrepancy between the practical demands of clinical practice, where equity, transparency, and trust are non-negotiable, and current AI capabilities. According to a human-centric paradigm, an AI tool's worth is assessed by its capacity to support and enhance human decision-making rather than just by how well it performs on a test set [131]. This calls for models that answer the "why" of a given patient by giving clinicians case-specific, intuitive explanations in addition to a diagnosis [132]. In order to guarantee that diverse populations are served fairly, it also requires that systems be audited and corrected for bias as a basic developmental step [93]. The crucial next step for the field is to embrace these human factors and move beyond technical metrics [133]. The AI development life-cycle must be fundamentally rethought in order to adopt this paradigm [134]. In order to make sure that the tools are not only technically sound but also culturally and contextually competent, it is imperative that models be co-designed with clinicians, ethicist s, and most importantly representatives from marginalized communities [135]. The creation of standardized frameworks for bias mitigation and fairness reporting, as well as intrinsically interpret-able architectures, must be the top priority of research [58]. The ultimate objective is to create AI that is a trustworthy, responsible, and equitable clinical partner in addition to being a potent classifier [136]. The only way to realize the long-promised revolution of AI in autism diagnosis and care is to make the shift from a technology centered to a human centered approach [137].

8 Conclusion

The field of AI-driven autism spectrum disorder detection is at a turning point, as this systematic review concludes. The research landscape reveals a profound and concerning neglect of the twin pillars required for clinical adoption: explainability and fairness. This is despite the fact that considerable progress has been made in maximizing classification accuracy, mainly through advanced multimodal fusion techniques. Explainability techniques, when used, only offer group-level insights that are not very useful for diagnosing individual patients, and the vast majority of developed models continue to be impenetrable "black boxes." Most concerningly, these potentially potent tools run the risk of automating and exacerbating already-existing healthcare disparities, especially against racially diverse and female populations, due to the widespread failure to audit for algorithmic fairness. Transparency, equity, and clinical utility must therefore be given top priority as fundamental design principles in order to move past a narrow focus on performance metrics. To create AI systems that are not only accurate but also reliable, equitable, and genuinely prepared for integration into clinical practice, future research must require stringent fairness auditing, create techniques for personalized explanation, and incorporate these elements into the very fabric of model development.

9 Future Directions

The review's notable gaps provide a clear path forward for further investigation. First, there is a pressing need to go beyond post-hoc explainability to intrinsically interpretable model architectures that produce patient-specific, intuitive explanations as a native output. This could be achieved by utilizing advancements in symbolic reasoning or attention mechanisms to give clinicians actionable insights for specific cases. Second, the model development lifecycle must make algorithmic fairness a non-negotiable requirement. This means that performance metrics broken down by important demographic subgroups (e.g., sex, age, race/ethnicity) must be reported. Additionally, strong bias mitigation strategies, like adversarial debiasing or fairness-aware loss functions, must be developed and implemented and validated on neurodevelopmental data. Developing explainable fairness frameworks that use XAI techniques to not only determine whether a model is biased but also to identify the underlying cause of bias by identifying which particular features or data sources are driving disparate performance across subgroups is the most important and innovative direction at the intersection of these two challenges. Ultimately, these efforts need to be validated on larger, more diverse, and prospectively collected datasets that go beyond the constraints of existing benchmarks like ABIDE in order to ensure real-world applicability. This will allow for the development of comprehensive AI systems that are by design accurate, transparent, and equitable.

Acknowledgment

The authors would like to thank Abdul Wali Khan University, Pakistan for supporting this research. The data used in the proposed research will be provided for a research purpose if requested from the principal author through email. The work is supported, in parts, by the Abdul Wali Khan University, Pakistan, and, in parts, by the Sohar University, Oman.

Data availability

All the codes and data used in this paper will be provided for research purposes if requested by researchers from the principal author.

Statements and Declarations

Authors declaration

All authors have read the manuscript and agreed to the submission. Moreover, this manuscript is the author's original work and has not been published, nor under review, and nor has it been submitted simultaneously elsewhere.

Compliance with Ethical Standards

Conflict of interest

All authors declare that they have no conflict of interest.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Funding statement

This research received no funding.

Credit Author Statement

Waqas Ahmad:- Conceptualization, Writing - Original Draft, Software
Ashraf Zia:- Visualization, Supervision, Data Curation
Muhammad Zakarya:- Software, Writing - Review & Editing

References

[1] AlSalehi, S.M., Alhifthy, E.H.: Autism spectrum disorder. In: Clinical Child Neurology, pp. 275–292. Springer, ??? (2020)

[2] Lord, C., Brugha, T.S., Charman, T., Cusack, J., Dumas, G., Frazier, T., Jones, E.J., Jones, R.M., Pickles, A., State, M.W., *et al.*: Autism spectrum disorder. Nature reviews Disease primers **6**(1), 5 (2020)

[3] Melander, D., Sivarasu, S., Yekinni, I., Yunzhang, C., Erdman, A.: Diagnostics as the key to advances in global health: Proposed methods for making reliable diagnostics widely available. Journal of medical devices **14**(1), 014702 (2020)

[4] Asif, S., Wenhui, Y., ur-Rehman, S.-, ul-ain, Q.-, Amjad, K., Yueyang, Y., Jinhai, S., Awais, M.: Advancements and prospects of machine learning in medical diagnostics: unveiling the future of diagnostic precision. Archives of Computational Methods in Engineering **32**(2), 853–883 (2025)

[5] Solmi, M., Song, M., Yon, D.K., Lee, S.W., Fombonne, E., Kim, M.S., Park, S., Lee, M.H., Hwang, J., Keller, R., *et al.*: Incidence, prevalence, and global burden of autism spectrum disorder from 1990 to 2019 across 204 countries. Molecular Psychiatry **27**(10), 4172–4180 (2022)

[6] Zeidan, J., Fombonne, E., Scorah, J., Ibrahim, A., Durkin, M.S., Saxena, S., Yusuf, A., Shih, A., Elsabbagh, M.: Global prevalence of autism: A systematic review update. Autism research **15**(5), 778–790 (2022)

[7] Poonia, A., Kumar, S., Raghuwanshi, G.: A comparative study of asd tests and gold standard rated asd tests: Adi-r and ados. JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTING **5**(1), 5–10 (2024)

[8] Chu, H., Westbrook, R.A., Njue-Marendes, S., Giordano, T.P., Dang, B.N.: The psychology of the wait time experience–what clinics can do to manage the waiting experience for patients: a longitudinal, qualitative study. BMC health services research **19**(1), 459 (2019)

[9] Singh, P., Anekar, U.: The importance of early identification and intervention for children with developmental delays. Indian Journal of Positive Psychology **9**(2) (2018)

[10] Mottron, L., Bzdok, D.: Autism spectrum heterogeneity: fact or artifact? Molecular psychiatry **25**(12), 3178–3185 (2020)

[11] Belard, A., Buchman, T., Forsberg, J., Potter, B.K., Dente, C.J., Kirk, A., Elster, E.: Precision diagnosis: a view of the clinical decision support systems (cdss) landscape through the lens of critical care. Journal of clinical monitoring and computing **31**(2), 261–271 (2017)

[12] Eskandar, K.: Artificial intelligence in psychiatric diagnosis: challenges and opportunities in the era of machine learning. Debates Em Psiquiatria **14**, 1–16 (2024)

[13] Chowdhury, R.H.: Big data analytics in the field of multifaceted analyses: A study on "health care management". World Journal of Advanced Research and Reviews **22**(3), 2165–2172 (2024)

[14] Vachha, B.A., Kumar, V.A., Pillai, J.J., Shimony, J., Tanabe, J., Sair, H.I.: Resting-state functional mri: Current state, controversies, limitations, and future directions—ajr expert panel narrative review. American Journal of Roentgenology (2024)

[15] Shamhan, A.N., Qaraqe, M., Al-Thani, D.: Advancements in automated assessment and diagnosis of autism spectrum disorder through multi-modality sensing technologies: Survey of the last decade. IEEE Transactions on Cognitive and Developmental Systems (2025)

[16] Taneera, S., Alhajj, R.: Diagnosis of autism spectrum disorder: a systematic review of clinical and artificial intelligence methods. Network Modeling Analysis in Health Informatics and Bioinformatics **14**(1), 7 (2025)

[17] Xu, X., Li, J., Zhu, Z., Zhao, L., Wang, H., Song, C., Chen, Y., Zhao, Q., Yang, J., Pei, Y.: A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis. Bioengineering **11**(3), 219 (2024)

[18] Syriopoulou-Delli, C.K.: Advances in autism spectrum disorder (asd) diagnostics: From theoretical frameworks to ai-driven innovations. Electronics **14**(5), 951 (2025)

[19] Sui, J., Zhi, D., Calhoun, V.D.: Data-driven multimodal fusion: approaches and applications in psychiatric research. Psychoradiology **3**, 026 (2023)

[20] Liu, K., Li, Y., Xu, N., Natarajan, P.: Learn to combine modalities in multimodal deep learning. arXiv preprint arXiv:1805.11730 (2018)

[21] Akhavan Aghdam, M., Sharifi, A., Pedram, M.M.: Combination of rs-fmri and smri data to discriminate autism spectrum disorders in young children using deep belief network. Journal of digital imaging **31**(6), 895–903 (2018)

[22] Zhao, F., Zhang, C., Geng, B.: Deep multimodal data fusion. ACM computing surveys **56**(9), 1–36 (2024)

[23] Bhardwaj, S., Rastogi, S.: Artificial Intelligence and Deep Learning. Addition Publishing House, ??? (2025)

[24] Hulsen, T.: Explainable artificial intelligence (xai): concepts and challenges in healthcare. Ai **4**(3), 652–666 (2023)

[25] Singh, B., Chandra, A., Arya, H., Kumar, G., Verma, A., Rastogi, V., Saxena, U., Joshi, D., Semwal, N., Kukreti, G., *et al.*: Machine learning techniques for neuroimaging analysis and interpretation. In: Machine Learning for Neurodegenerative Disorders, pp. 33–50. CRC Press, ??? (2024)

[26] Yang, C.C.: Explainable artificial intelligence for predictive modeling in healthcare. Journal of healthcare informatics research **6**(2), 228–239 (2022)

[27] Hilliard, A.: Fairness and bias in algorithmic recruitment tools: An interdisciplinary approach. PhD thesis, Goldsmiths, University of London (2025)

[28] Banks, N.C.R.: Exploring juvenile probation officers' perception of strategies to reduce juvenile recidivism in southeastern michigan. PhD thesis, Walden University (2025)

[29] Chakravorty, I., Daga, S., Sharma, S., Chakravorty, S., Fischer, M., Mehta, R.K.: Bridging the gap 2021-report: Full report. Sushruta Journal of Health Policy & Opinion (2021)

[30] Mergen, A., Çetin-Kılıç, N., Özbilgin, M.F.: Artificial intelligence and bias towards marginalised groups: Theoretical roots and challenges. In: AI and Diversity in a Datafied World of Work: Will the Future of Work Be Inclusive? vol. 12, pp. 17–38. Emerald Publishing Limited, ??? (2025)

[31] Weiss, M., Tonella, P.: Adopting two supervisors for efficient use of large-scale remote deep neural networks. ACM Transactions on Software Engineering and Methodology **33**(1), 1–29 (2023)

[32] Currim, S., Snodgrass, R.T., Suh, Y.-K.: Identifying the root causes of dbms suboptimality. ACM Transactions on Database Systems **49**(1), 1–40 (2024)

[33] Dehbozorgi, R., Zangeneh, S., Khooshab, E., Nia, D.H., Hanif, H.R., Samian, P., Yousefi, M., Hashemi, F.H., Vakili, M., Jamalimoghadam, N., *et al.*: The application of artificial intelligence in the field of mental health: a systematic review. BMC psychiatry **25**(1), 132 (2025)

[34] Liu, Q., Shu, Q., Gao, W., Luo, Y., Li, Z., Yang, J., Xu, W.: Automatic interpretation of potential field data based on euler deconvolution with linear background. Applied Sciences **13**(9), 5323 (2023)

[35] Hou, L., Rong, P., Yang, Y., Fang, J., Wang, J., Wang, Y., Zhang, J., Zhang, S., Zhang, Z., Chen, J.D., *et al.*: Auricular vagus nerve stimulation improves visceral hypersensitivity and gastric motility and depression-like behaviors via vago-vagal pathway in a rat model of functional dyspepsia. Brain Sciences **13**(2), 253 (2023)

[36] Sun, Y., Zhan, W., Dong, T., Guo, Y., Liu, H., Gui, L., Zhang, Z.: Real-time recognition and detection of bactrocera minax (diptera: trypetidae) grooming behavior using body region localization and improved c3d network. Sensors **23**(14), 6442 (2023)

[37] Erkan, U., Thanh, D.N.: Autism spectrum disorder detection with machine learning methods. Current Psychiatry Research and Reviews Formerly: Current Psychiatry Reviews **15**(4), 297–308 (2019)

[38] Baumeister, D., Hogrebe, T.: On the complexity of predicting election outcomes and estimating their robustness. SN Computer Science **4**(4), 362 (2023)

[39] Talpur, N., Abdulkadir, S.J., Alhussian, H., Hasan, M.H., Aziz, N., Bamhdi, A.: Deep neuro-fuzzy system application trends, challenges, and future perspectives: A systematic survey. Artificial intelligence review **56**(2), 865–913 (2023)

[40] Parums, D.V.: Review articles, systematic reviews, meta-analysis, and the updated preferred reporting items for systematic reviews and meta-analyses (prisma) 2020 guidelines. Medical science monitor: international medical journal of experimental and clinical research **27**, 934475–1 (2021)

[41] Tejwani, R., Liska, A., You, H., Reinen, J., Das, P.: Autism classification using brain functional connectivity dynamics and machine learning. arXiv preprint arXiv:1712.08041 (2017)

[42] Li, X., Dvornek, N.C., Zhuang, J., Ventola, P., Duncan, J.S.: Brain biomarker interpretation in asd using deep learning and fmri. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 206–214 (2018). Springer

[43] Eslami, T., Mirjalili, V., Fong, A., Laird, A.R., Saeed, F.: Asd-diagnet: a hybrid learning approach for detection of autism spectrum disorder using fmri data. Frontiers in neuroinformatics **13**, 70 (2019)

[44] Lanillos, P., Oliva, D., Philippsen, A., Yamashita, Y., Nagai, Y., Cheng, G.: A review on neural network models of schizophrenia and autism spectrum disorder. Neural Networks **122**, 338–363 (2020)

[45] Pfohl, S.R., Foryciarz, A., Shah, N.H.: An empirical characterization of fair machine learning for clinical risk prediction. Journal of biomedical informatics **113**, 103621 (2021)

[46] Santana, C.P., Carvalho, E.A., Rodrigues, I.D., Bastos, G.S., Souza, A.D., Brito, L.L.: rs-fmri and machine learning for asd diagnosis: a systematic review and meta-analysis. Scientific reports **12**(1), 6030 (2022)

[47] Wan, L., Li, Y., Zhu, G., Yang, D., Li, F., Wang, W., Chen, J., Yang, G., Li, R.: Multimodal investigation of dynamic brain network alterations in autism spectrum disorder: Linking connectivity dynamics to symptoms and developmental trajectories. NeuroImage **302**, 120895 (2024)

[48] Atlam, E.-S., Aljuhani, K.O., Gad, I., Abdelrahim, E.M., Atwa, A.E.M., Ahmed, A.: Automated identification of autism spectrum disorder from facial images using explainable deep learning models. Scientific Reports **15**(1), 26682 (2025)

[49] Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. In: International Conference on Machine Learning, pp. 3384–3393 (2018). PMLR

[50] Singh, A., Sengupta, S., Lakshminarayanan, V.: Explainable deep learning models in medical image analysis. Journal of imaging **6**(6), 52 (2020)

[51] Sharif, H., Khan, R.A.: A novel machine learning based framework for detection of autism spectrum disorder (asd). Applied Artificial Intelligence **36**(1), 2004655 (2022)

[52] Dcouto, S.S., Pradeepkandhasamy, J.: Multimodal deep learning in early autism detection—recent advances and challenges. Engineering Proceedings **59**(1), 205 (2024)

[53] Saponaro, S., Lizzi, F., Serra, G., Mainas, F., Oliva, P., Giuliano, A., Calderoni, S., Retico, A.: Deep learning based joint fusion approach to exploit anatomical and functional brain information in autism spectrum disorders. Brain Informatics **11**(1), 2 (2024)

[54] Sha, M., Al-Dossary, H., Rahamathulla, M.P.: Multimodal data fusion framework for early prediction of autism spectrum disorder. Human behavior and emerging technologies **2025**(1), 1496105 (2025)

[55] Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)

[56] Supekar, K., Ryali, S., Yuan, R., Kumar, D., Angeles, C., Menon, V.: Robust, generalizable, and interpretable ai-derived brain fingerprints of autism and social-communication symptom severity. Biological psychiatry **92**(8), 643 (2022)

[57] Zhou, Y., Jia, G., Ren, Y., Ren, Y., Xiao, Z., Wang, Y.: Advancing asd identification with neuroimaging: a novel garl methodology integrating deep q-learning and generative adversarial networks. BMC Medical Imaging **24**(1), 186 (2024)

[58] Chinnaraju, A.: Explainable ai (xai) for trustworthy and transparent decision-making: A theoretical framework for ai interpretability. World Journal of Advanced Engineering Technology and Sciences **14**(3), 170–207 (2025)

[59] Romano, Y., Bates, S., Candes, E.: Achieving equalized odds by resampling sensitive attributes. Advances in neural information processing systems **33**, 361–371 (2020)

[60] Xu, J., Xiao, Y., Wang, W.H., Ning, Y., Shenkman, E.A., Bian, J., Wang, F.: Algorithmic fairness in computational medicine. EBioMedicine **84** (2022)

[61] Alves, C.L., Toutain, T.G.L.d.O., Aguiar, P.d.C., Pineda, A.M., Roster, K., Thielemann, C., Porto, J.A.M., Rodrigues, F.A.: Diagnosis of autism spectrum disorder based on functional brain networks and machine learning. Scientific Reports **13**(8072) (2023) https://doi.org/10.1038/s41598-023-35177-4

[62] Sheik Abdullah, A., Keerthana, V., Geetha, S., Mishra, U.: Leveraging deep learning for enhanced diagnosis of autism spectrum disorder using resting-state functional magnetic resonance imaging and clinical data. Results in Engineering **25**, 104444 (2025) https://doi.org/10.1016/j.rineng.2025.104444

[63] Correa, R., Shaan, M., Trivedi, H., Patel, B., Celi, L.A.G., Gichoya, J.W., Banerjee, I.: A systematic review of 'fair'ai model development for image classification and prediction. Journal of Medical and Biological Engineering **42**(6), 816–827 (2022)

[64] Chen, Z., Pawar, K., Ekanayake, M., Pain, C., Zhong, S., Egan, G.F.: Deep learning for image enhancement and correction in magnetic resonance imaging—state-of-the-art and challenges. Journal of Digital Imaging **36**(1), 204–230 (2023)

[65] Lu, S.: Regulating algorithmic harms (2024)

[66] Nafisah, I., Mahmoud, N., Ewees, A.A., Khattap, M.G., Dahou, A., Alghamdi, S.M., Fares, I.A., Azmi Al-Betar, M., Abd Elaziz, M.: Deep learning-based feature selection for detection of autism spectrum disorder. Frontiers in Artificial Intelligence **8**, 1594372 (2025)

[67] Yuan, C., Linn, K.A., Hubbard, R.A.: Algorithmic fairness of machine learning models for alzheimer disease progression. JAMA network open **6**(11), 2342203–2342203 (2023)

[68] Atlam, E.-S., Masud, M., Rokaya, M., Meshref, H., Gad, I., Almars, A.M.: Easdm: Explainable autism spectrum disorder model based on deep learning. Journal of Disability Research **3**(1), 20240003 (2024)

[69] Alsbakhi, A., Thabtah, F., Lu, J.: Autism data classification using ai algorithms with rules: Focused review. Bioengineering **12**(2), 160 (2025)

[70] Wang, R., Chaudhari, P., Davatzikos, C.: Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. Proceedings of the National Academy of Sciences **120**(6), 2211613120 (2023)

[71] Vidya, S., Gupta, K., Aly, A., Wills, A., Ifeachor, E., Shankar, R.: Explainable ai for autism

diagnosis: Identifying critical brain regions using fmri data. arXiv preprint arXiv:2409.15374 (2024)

[72] Strock, A., Nghiem, T.-A.E., Lecea, A.D., Hassan, A., Ryali, S., Menon, V.: Explainable ai techniques for dynamic functional brain imaging: validation and analysis of e/i imbalance in autism. bioRxiv, 2025–02 (2025)

[73] Jeon, I., Kim, M., So, D., Kim, E.Y., Nam, Y., Kim, S., Shim, S., Kim, J., Moon, J.: Reliable autism spectrum disorder diagnosis for pediatrics using machine learning and explainable ai. Diagnostics **14**(22), 2504 (2024)

[74] Li, B., Jiang, X., Zhang, K., Harmanci, A.O., Malin, B., Gao, H., Shi, X., Initiative, A.D.N.: Enhancing fairness in disease prediction by optimizing multiple domain adversarial networks. PLOS Digital Health **4**(5), 0000830 (2025)

[75] Petersen, E., Ferrante, E., Ganz, M., Feragen, A.: Are demographically invariant models and representations in medical imaging fair? arXiv preprint arXiv:2305.01397 (2023)

[76] Li, S., Wu, Q., Zhou, D., Li, X., Miao, D., Hong, C., Gu, W., Shang, Y., Okada, Y., Chen, M.H., *et al.*: Fairfml: Fair federated machine learning with a case study on reducing gender disparities in cardiac arrest outcome prediction. npj Health Systems **2**(1), 29 (2025)

[77] Meijden, S., Wang, Y., Arbous, M., Geerts, B., Steyerberg, E., Hernandez-Boussard, T.: Navigating fairness in ai-based prediction models: Theoretical constructs and practical applications. medRxiv, 2025–03 (2025)

[78] Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., Zhou, S.K.: Addressing fairness issues in deep learning-based medical image analysis: a systematic review. npj Digital Medicine **7**(1), 286 (2024)

[79] Yang, Y., Zhang, H., Gichoya, J.W., Katabi, D., Ghassemi, M.: The limits of fair medical imaging ai in real-world generalization. Nature Medicine **30**(10), 2838–2848 (2024)

[80] Dong, Y., Batalle, D., Deprez, M.: Reproducible comparison and interpretation of machine learning classifiers to predict autism on the abide multimodal dataset. medRxiv, 2024–09 (2024)

[81] Wang, H., Jing, H., Yang, J., Liu, C., Hu, L., Tao, G., Zhao, Z., Shen, N.: Identifying autism spectrum disorder from multi-modal data with privacy-preserving. npj Mental Health Research **3**(1), 15 (2024)

[82] Chen, B., Yin, B., Ke, H.: Interpretation of deep non-linear factorization for autism. Frontiers in Psychiatry **14**, 1199113 (2023)

[83] Alcaide, F., Illan, I., Ramirez, J., Gorriz, J.: Unraveling the autism spectrum heterogeneity: Insights from abide i database using data/model-driven permutation testing approaches. arXiv preprint arXiv:2405.12225 (2024)

[84] Liu, M., Ning, Y., Teixayavong, S., Liu, X., Mertens, M., Shang, Y., Li, X., Miao, D., Liao, J., Xu, J., *et al.*: A scoping review and evidence gap analysis of clinical ai fairness. npj Digital Medicine **8**(1), 360 (2025)

[85] Bae, S., Hong, J., Ha, S., Moon, J., Yu, J., Choi, H., Lee, J., Do, R., Sim, H., Kim, H., *et al.*: Multimodal ai for risk stratification in autism spectrum disorder: integrating voice and screening tools. npj Digital Medicine **8**(1), 538 (2025)

[86] Vimbi, V., Shaffi, N., Sadiq, M.A., Sirasanagandla, S.R., Aradhya, V.M., Kaiser, M.S., Wang, S., Mahmud, M.: Application of explainable artificial intelligence in autism spectrum disorder detection. Cognitive Computation **17**(3), 104 (2025)

[87] Ghassami, A., Khodadadian, S., Kiyavash, N.: Fairness in supervised learning: An information theoretic approach. In: 2018 IEEE International Symposium on Information Theory (ISIT), pp. 176–180 (2018). IEEE

[88] Tuan, D.A.: Bridging the gap between black box ai and clinical practice: Advancing explainable ai for trust, ethics, and personalized healthcare diagnostics (2024)

[89] Muhle, R.A., Reed, H.E., Stratigos, K.A., Veenstra-VanderWeele, J.: The emerging clinical neuroscience of autism spectrum disorder: a review. JAMA psychiatry **75**(5), 514–523 (2018)

[90] Agarwal, A., Agarwal, H.: A seven-layer model with checklists for standardising fairness assessment throughout the ai lifecycle. AI and Ethics **4**(2), 299–314 (2024)

[91] Bogen, M.: Navigating demographic measurement for fairness and equity. Technical Report: https://cdt. org/insights/report-navigating-demographic-measurement-for-fairness-and-equity/ (2024)

[92] Chandak, T., Jayashree, J., Vijayashree, J.: Trends and advancements of ai and xai in drug discovery. In: Explainable AI (XAI) for Sustainable Development, pp. 233–255. Chapman and Hall/CRC, ??? (2024)

[93] Landers, R.N., Behrend, T.S.: Auditing the ai auditors: A framework for evaluating fairness and bias in high stakes ai predictive models. American Psychologist **78**(1), 36 (2023)

[94] Moon, D., Ahn, S.: Metrics and algorithms for identifying and mitigating bias in ai design: A counterfactual fairness approach. IEEE Access (2025)

[95] Egbuna, I.K., Otoibhili, P.A., Abdulkareem, R.O., Ikechukwu, F., Ogbozor, P.E.O., Azih, N.K., Akomolafe, O.B.: Explainable artificial intelligence (xai) in diagnosing neurodevelopmental disorders: From black boxes to clinical transparency (2025)

[96] Kamble, S.S., Gunasekaran, A.: Big data-driven supply chain performance measurement system: a review and framework for implementation. International journal of production research **58**(1), 65–86 (2020)

[97] Nakach, F.-Z., Idri, A., Goceri, E.: A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification. Artificial Intelligence Review **57**(12), 327 (2024)

[98] Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., *et al.*: Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. IEEE Geoscience and Remote Sensing Magazine **7**(1), 6–39 (2019)

[99] Zhou, Y., Yang, X., Liu, S., Yin, J.: Multimodal medical image fusion network based on target information enhancement. IEEE Access **12**, 70851–70869 (2024)

[100] Qu, G., Zhou, Z., Calhoun, V.D., Zhang, A., Wang, Y.-P.: Integrated brain connectivity analysis with fmri, dti, and smri powered by interpretable graph neural networks. Medical Image Analysis **103**, 103570 (2025)

[101] Tsanas, A.: Relevance, redundancy, and complementarity trade-off (rrct): A principled, generic, robust feature-selection tool. Patterns **3**(5) (2022)

[102] Boulahia, S.Y., Amamra, A., Madi, M.R., Daikh, S.: Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Machine Vision and Applications **32**(6), 121 (2021)

[103] Oladunni, T., Aneni, E.: Explainable deep neural network for multimodal ecg signals: Intermediate vs late fusion. arXiv preprint arXiv:2508.11666 (2025)

[104] Sharma, K., Giannakos, M.: Multimodal data capabilities for learning: What can multimodal data tell us about learning? British Journal of Educational Technology **51**(5), 1450–1484 (2020)

[105] Farahani, F.V., Fiok, K., Lahijanian, B., Karwowski, W., Douglas, P.K.: Explainable ai: A

review of applications to neuroimaging data. Frontiers in Neuroscience **16**, 906290 (2022)

[106] Menon, V.: 20 years of the default mode network: A review and synthesis. Neuron **111**(16), 2469–2487 (2023)

[107] Gkintoni, E., Panagioti, M., Vassilopoulos, S.P., Nikolaou, G., Boutsinas, B., Vantarakis, A.: Leveraging ai-driven neuroimaging biomarkers for early detection and social function prediction in autism spectrum disorders: A systematic review. In: Healthcare, vol. 13, p. 1776 (2025)

[108] Ullah, I., Liu, K., Yamamoto, T., Zahid, M., Jamal, A.: Modeling of machine learning with shap approach for electric vehicle charging station choice behavior prediction. Travel Behaviour and Society **31**, 78–92 (2023)

[109] Mukherjee, P., Sadhukhan, S., Godse, M., Chakraborty, B.: Early detection of autism spectrum disorder (asd) using traditional machine learning models. International Journal of Advanced Computer Science and Applications **14**(6), 231–245 (2023)

[110] Tai, A.M., Albuquerque, A., Carmona, N.E., Subramanieapillai, M., Cha, D.S., Sheko, M., Lee, Y., Mansur, R., McIntyre, R.S.: Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. Artificial intelligence in medicine **99**, 101704 (2019)

[111] Mackin, S., Major, V.J., Chunara, R., Newton-Dame, R.: Identifying and mitigating algorithmic bias in the safety net. npj Digital Medicine **8**(1), 335 (2025)

[112] Lalor, J.P., Abbasi, A., Oketch, K., Yang, Y., Forsgren, N.: Should fairness be a metric or a model? a model-based framework for assessing bias in machine learning pipelines. ACM Transactions on Information Systems **42**(4), 1–41 (2024)

[113] Yang, Z., Ko, Y.L., Varshney, K.R., Ying, Y.: Minimax auc fairness: Efficient algorithm with provable convergence. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 11909–11917 (2023)

[114] McCradden, M.D., Joshi, S., Mazwi, M., Anderson, J.A.: Ethical limitations of algorithmic fairness solutions in health care machine learning. The Lancet Digital Health **2**(5), 221–223 (2020)

[115] Liu, S., Vicente, L.N.: Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. Computational Management Science **19**(3), 513–537 (2022)

[116] Zhou, L., Li, Z., Zhou, J., Li, H., Chen, Y., Huang, Y., Xie, D., Zhao, L., Fan, M., Hashmi, S., *et al.*: A rapid, accurate and machine-agnostic segmentation and quantification method for ct-based covid-19 diagnosis. IEEE transactions on medical imaging **39**(8), 2638–2652 (2020)

[117] Cho, P.J., Yi, J., Ho, E., Shandhi, M.M.H., Dinh, Y., Patil, A., Martin, L., Singh, G., Bent, B., Ginsburg, G., *et al.*: Demographic imbalances resulting from the bring-your-own-device study design. JMIR mHealth and uHealth **10**(4), 29510 (2022)

[118] Hanna, M.G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., Rashidi, H.H.: Ethical and bias considerations in artificial intelligence/machine learning. Modern Pathology **38**(3), 100686 (2025)

[119] Varoquaux, G., Colliot, O.: Evaluating machine learning models and their diagnostic value. Machine learning for brain disorders, 601–630 (2023)

[120] Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.: Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review **55**(5), 3503–3568 (2022)

[121] Bulathwela, S., Pérez-Ortiz, M., Holloway, C., Cukurova, M., Shawe-Taylor, J.: Artificial intelligence alone will not democratise education: On educational inequality, techno-solutionism and inclusive tools. Sustainability **16**(2), 781 (2024)

[122] Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A., Group, P.-P.: Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. Systematic reviews **4**(1), 1 (2015)

[123] Yanow, D.: Neither rigorous nor objective?: Interrogating criteria for knowledge claims in interpretive science. In: Interpretation and Method, pp. 97–119. Routledge, ??? (2015)

[124] Corbett-Davies, S., Gaebler, J.D., Nilforoshan, H., Shroff, R., Goel, S.: The measure and mismeasure of fairness. Journal of Machine Learning Research **24**(312), 1–117 (2023)

[125] Jiao, T., Guo, C., Feng, X., Chen, Y., Song, J.: A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. Computers, Materials & Continua **80**(1) (2024)

[126] Sargeant, J.M., Brennan, M.L., O'Connor, A.M.: Levels of evidence, quality assessment, and risk of bias: evaluating the internal validity of primary research. Frontiers in Veterinary Science **9**, 960957 (2022)

[127] Dibaji, S.M.: Investigating sex-related bias in brain mri using fair and explainable deep learning (2025)

[128] Plant, D., Barton, A.: Adding value to real-world data: the role of biomarkers. Rheumatology **59**(1), 31–38 (2020)

[129] Flemming, K., Noyes, J.: Qualitative evidence synthesis: where are we at? International Journal of Qualitative Methods **20**, 1609406921993276 (2021)

[130] Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., Ghassemi, M.: The road to explainability is paved with bias: Measuring the fairness of explanations. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1194–1206 (2022)

[131] Frenette, J.: Benefits of the rise of human-centric AI in enterprise decision making. White paper (2022)

[132] Corazza, G.R., Lenti, M.V., Howdle, P.D.: Diagnostic reasoning in internal medicine: a practical reappraisal. Internal and emergency medicine **16**(2), 273–279 (2021)

[133] Morandini, S., Fraboni, F., Hall, M., Quintana-Amate, S., Pietrantoni, L.: Human factors and emerging needs in aerospace manufacturing planning and scheduling. Cognition, Technology & Work, 1–19 (2024)

[134] Sakao, T., Funk, P., Matschewsky, J., Bengtsson, M., Ahmed, M.U.: Ai-lce: Adaptive and intelligent life cycle engineering by applying digitalization and ai methods–an emerging paradigm shift in life cycle engineering. Procedia CIRP **98**, 571–576 (2021)

[135] O'Brien, J., Fossey, E., Palmer, V.J.: A scoping review of the use of co-design methods with culturally and linguistically diverse communities to improve or adapt mental health services. Health & Social Care in the Community **29**(1), 1–17 (2021)

[136] Goktas, P., Grzybowski, A.: Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy ai. Journal of Clinical Medicine **14**(5), 1605 (2025)

[137] Bertl, M., Metsallik, J., Ross, P.: A systematic literature review of ai-based digital decision support systems for post-traumatic stress disorder. Frontiers in Psychiatry **13**, 923613 (2022)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementryfile.pdf