SCIENTIFIC
RESEARCH ARCHIVES

International Journal of Biological and Pharmaceutical Sciences Archive

ISSN: 0799-6616 (Online)

Journal homepage: https://ijbpsa.com/

(REVIEW ARTICLE)

Check for updates

Explainable Artificial Intelligence (XAI) in Diagnosing Neurodevelopmental Disorders: From Black Boxes to Clinical Transparency

Ifeanyi Kingsley Egbuna [1, *], Precious Airebanmen Otoibhili [2], Rofiat Oyiza Abdulkareem [3], Festus Ikechukwu Ogbozor [4], Praise Etinosa Oyegue [5], Nnaemeka Kelsey Azih [6] and Oluwaseyi Blessing Akomolafe [7]

[1] Department of Supply Chain Management, Marketing, and Management, Wright State University, United States.
[2] Department of Anatomy, Ambrose Alli University, Ekpoma, Edo State, Nigeria.
[3] Department of Pharmacology, Bayero University, Kano, Nigeria.
[4] Department of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Poland.
[5] Department of Anatomy, Ambrose Alli University, Ekpoma, Edo State, Nigeria.
[6] Department of Paediatrics, Margaret Lawrence University Teaching Hospital Abuja, Nigeria.
[7] Department of Master of Arts in Psychology, School of Health Sciences, Mapúa University, Manila, Philippines.

International Journal of Biological and Pharmaceutical Sciences Archive, 2025, 10(01), 031-057

Publication history: Received on 18 May 2025; revised on 30 June 2025; accepted on 03 July 2025

Article DOI: https://doi.org/10.53771/ijbpsa.2025.10.1.0052

Abstract

Neurodevelopmental disorders (NDDs), including autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD), affect millions of children globally, posing immense clinical, social, and economic burdens. While early diagnosis remains critical for improving long-term outcomes, traditional assessment methods often suffer from subjectivity, late detection, and limited scalability. The advent of artificial intelligence (AI) has ushered in a new era of data-driven precision in mental health diagnostics, yet the opaque, "black-box" nature of most AI models has hindered their acceptance in high-stakes clinical settings. In response, explainable AI (XAI) has emerged as a crucial bridge between computational performance and clinical interpretability. This review critically explores the foundations, applications, and limitations of XAI in the early detection and diagnosis of NDDs. We examine core XAI paradigms—ranging from SHAP and LIME to attention-based and saliency-driven techniques—and illustrate their capacity to illuminate AI decision-making in real-world diagnostic workflows. Case studies on the use of XAI in analyzing fMRI, EEG, and behavioral data for ASD and ADHD offer compelling evidence of its transformative potential. Yet, challenges persist, including the inconsistency of explanation reliability, trade-offs between model transparency and accuracy, and the risks posed by data bias, particularly in underrepresented pediatric populations. Looking forward, we chart future directions involving the fusion of XAI with digital biomarkers, federated learning for multicenter collaboration, and clinician-in-the-loop systems to ensure ethical, trustworthy, and context-sensitive deployment. By integrating interpretability into the very fabric of AI systems, this review advocates for a future where transparency and technological innovation coalesce to advance pediatric neuropsychiatric care.

Keywords: Explainable AI; Neurodevelopmental Disorders; Autism Diagnosis; ADHD Clustering; Interpretable Models; Pediatric Neuroscience; Clinical Transparency; Digital Biomarkers

1. Introduction

The diagnosis and management of neurodevelopmental disorders (NDDs) remain one of the most complex challenges in contemporary clinical neuroscience. These disorders, which include autism spectrum disorder (ASD), attention-deficit/hyperactivity disorder (ADHD), intellectual disabilities, and specific learning disorders, typically emerge early in development and persist throughout life, significantly impairing social, academic, and occupational functioning [1].

[*] Corresponding author: Ifeanyi Kingsley Egbuna; Email: egbuna.2@wright.edu

Timely and accurate diagnosis is essential for improving long-term outcomes, but conventional diagnostic approaches are often labor-intensive, subjective, and prone to variability across clinicians and cultures [1,2]. In recent years, artificial intelligence (AI) has demonstrated immense potential to transform how we approach the diagnosis of these disorders, particularly by leveraging vast and multidimensional datasets such as neuroimaging, electrophysiological signals, and behavioral assessments [3].

However, as AI continues to penetrate clinical decision-making, concerns about the transparency and interpretability of its models have become increasingly prominent. Clinicians, caregivers, and regulators are rightfully wary of relying on systems that cannot clearly explain how diagnostic conclusions are reached. This opacity, often referred to as the "black-box" problem, is a major barrier to the integration of AI into sensitive domains such as pediatric mental health [4]. To bridge this gap, researchers are now turning to Explainable Artificial Intelligence (XAI), a paradigm that seeks to make the decision-making processes of AI models more interpretable, reliable, and clinically acceptable [5]. This review explores the current landscape of AI in diagnosing neurodevelopmental disorders, critically examines the role of XAI in enhancing clinical transparency, and outlines emerging directions for research and application.

1.1. The Burden of Neurodevelopmental Disorders

Neurodevelopmental disorders affect a significant proportion of the global pediatric population and represent a major public health concern. According to the World Health Organization (WHO), an estimated 317 million children and adolescents were living with developmental disabilities in 2019, with the highest burdens found in low- and middle-income countries (LMICs) [6]. These conditions disrupt various domains of development—including cognition, language, motor skills, and social interaction—and often require lifelong support and intervention. Early diagnosis is critical, as timely interventions have been shown to markedly improve outcomes in language acquisition, adaptive behavior, and academic success [7].

Despite growing awareness, underdiagnosis and delayed diagnosis remain pervasive challenges, particularly in resource-constrained settings. A lack of trained professionals, insufficient screening tools, and cultural stigmatization contribute to late identification of affected individuals. Moreover, diagnostic heterogeneity—manifested in the overlapping symptoms across different NDDs—makes the process inherently complex even in high-resource environments. This landscape underscores the urgent need for more objective, scalable, and data-driven diagnostic tools that can facilitate early detection and personalized care [8,9].

1.2. AI's Rise in Early Detection and Diagnosis

The application of artificial intelligence in medicine, particularly in the domain of neurodevelopmental diagnostics, has grown rapidly over the past decade. Machine learning (ML) algorithms, and more recently, deep learning architectures, have demonstrated high accuracy in identifying patterns in complex and high-dimensional datasets. These include structural and functional neuroimaging data (e.g., MRI, fMRI), electrophysiological recordings (e.g., EEG), and behavioral video or audio data—modalities that are rich in diagnostic potential but difficult for human experts to interpret consistently [10,11].

For example, studies have shown that AI can detect atypical neural connectivity patterns in fMRI scans of children with ASD, often before behavioral symptoms fully manifest [12]. Similarly, attention-based models have been used to classify ADHD subtypes using resting-state EEG data with promising results [13]. AI has also enabled scalable and non-invasive screening approaches using video analysis of eye gaze, facial expression, and motor behavior in young children [14]. These developments point to a future where AI can serve as a powerful adjunct to clinical judgment, aiding early detection and stratification of NDDs.

However, despite these promising advances, widespread clinical adoption remains limited. A primary reason is the opaque nature of many AI systems, which often fail to provide understandable justifications for their decisions—making clinicians hesitant to rely on them for high-stakes diagnoses. This concern is particularly acute in pediatrics, where decisions affect not just the child, but also family dynamics, educational pathways, and therapeutic interventions.

1.3. The Black-Box Problem in Clinical AI

The "black-box" nature of many AI systems—especially those based on deep neural networks—presents a serious challenge to their clinical use. These models, while capable of achieving high predictive performance, do not readily reveal how input features contribute to their outputs. In clinical neuroscience, where decisions must be evidence-based, interpretable, and justifiable, this opacity undermines trust, safety, and accountability [15]. For instance, a clinician might hesitate to alter a child's treatment plan based on an AI model that cannot explain why it flagged a particular diagnosis. Moreover, regulatory bodies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) are increasingly emphasizing transparency and explainability as prerequisites for approving AI tools in healthcare. The ethical concerns are equally pressing: if a model makes a misclassification, clinicians and families must be able to understand the contributing factors and seek recourse or correction [16,17]. In the absence of explainability, there is a risk that AI could perpetuate biases embedded in training data, particularly affecting underrepresented populations[18].

In response to these challenges, the field of Explainable Artificial Intelligence (XAI) has emerged to bridge the gap between model complexity and human interpretability. Techniques such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and attention visualization offer insights into which features the model considers most influential. By translating abstract model decisions into understandable reasoning, XAI has the potential to enhance trust, enable clinical oversight, and pave the way for ethically responsible AI deployment in neurodevelopmental diagnostics [19,20].

2. Overview of AI in Neurodevelopmental Diagnostics

Artificial Intelligence (AI) has emerged as a transformative catalyst in the early detection and diagnosis of neurodevelopmental disorders (NDDs), promising to revolutionize how clinicians approach complex developmental conditions. NDDs, including autism spectrum disorder (ASD), attention-deficit/hyperactivity disorder (ADHD), and intellectual disabilities, are characterized by atypical brain development that manifests in cognitive, behavioral, and social impairments. Diagnosing these conditions early is critical for maximizing the benefits of therapeutic interventions, yet traditional diagnostic pathways are often time-consuming, subjective, and limited by human perceptual constraints. These limitations underscore the urgent need for objective, data-driven methods to aid early identification, particularly during sensitive developmental windows [21,22].

In response, researchers and clinicians have increasingly turned to AI tools capable of analyzing vast, multimodal datasets—including neuroimaging, behavioral assessments, and genetic information—to uncover latent patterns predictive of NDDs. To provide a comprehensive snapshot of how AI is transforming neurodevelopmental diagnostics, we present Table 1, which summarizes the primary modalities, their data sources, key AI techniques, and their applications in diagnosing NDDs. This table serves as a roadmap for understanding the diverse ways AI is applied across neuroimaging, behavioral analysis, and digital phenotyping, highlighting the potential for both standalone and integrative approaches. Through the integration of machine learning (ML), deep learning (DL), and natural language processing (NLP), AI systems can model high-dimensional relationships between biological signals and clinical outcomes. This not only accelerates the diagnostic process but also facilitates personalized predictions about developmental trajectories and treatment responsiveness [23]. The current section explores how AI is reshaping neurodevelopmental diagnostics by focusing on key areas such as neuroimaging, behavioral and speech analysis, digital phenotyping, and multimodal integration.

Table 1 AI Modalities and Applications in Diagnosing Neurodevelopmental Disorders

Modality	Data Source	AI Techniques	NDD Applications	Key Studies	Advantages	Challenges
Structural MRI (sMRI)	Brain morphology (e.g., cortical thickness, volume)	Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs)	ASD classification, biomarker identification (e.g., precuneus, corpus callosum)	Heinsfeld et al. [24], Sherkatghanad et al. [25]	High-resolution anatomical insights, non-invasive	High computational cost, preprocessing variability
Functional MRI (fMRI)	Resting-state connectivity, BOLD signals	Graph Neural Networks (GNNs), Attention-based Models	ASD, ADHD diagnosis via connectivity patterns (e.g., posterior	Khosla et al. [27], Ahmadi et al. [28]	Captures dynamic brain activity	Noisy data, complex interpretation

			cingulate cortex)			
Diffusion Tensor Imaging (DTI)	White matter tract integrity (e.g., fractional anisotropy)	Random Forests, Autoencoders	DLD, ASD detection via tract connectivity (e.g., arcuate fasciculus)	Lou et al. [31], Zhang et al. [30]	Reveals microstructural changes	Limited by small sample sizes, preprocessing sensitivity
EEG	Brain electrical activity (e.g., alpha/theta rhythms)	Recurrent Neural Networks (RNNs), SHAP-enhanced Models	ASD, ADHD classification via frequency bands	Rogala et al. [98], Chakladar et al. [13]	Cost-effective, high temporal resolution	Susceptible to artifacts, requires expertise for interpretation
Behavioral Analysis	Video/audio of facial expressions, gaze, motor behavior	CNNs, RNNs, Transformers	Early ASD detection via gaze patterns, motor abnormalities	Perochon et al. [36], Leo et al. [40]	Non-invasive, scalable for home use	Cultural variability, data standardization issues
Speech and Prosody Analysis	Audio recordings (e.g., pitch, pausing, lexical diversity)	Transformers, SVMs with Mel-frequency cepstral coefficients	ASD, language disorder detection via prosodic features	Bone et al. [42], Liu et al. [43]	Scalable via mobile apps, non-invasive	Requires large annotated datasets, privacy concerns
Digital Phenotyping	Smartphone sensors, wearables (e.g., GPS, heart rate, sleep data)	Deep Learning, Time-series Models	ASD, ADHD monitoring via social, motor, and physiological patterns	Saeb et al. [54], Goodwin et al. [59]	Ecologically valid, continuous monitoring	Privacy issues, device variability
Multimodal Integration	Combined sMRI, fMRI, EEG, behavioral, and sensor data	Late-fusion Models, Transformers	Enhanced ASD/ADHD detection via synergistic patterns	Akhavan Aghdam et al. [32], Chen et al. [46]	Comprehensive view of neurodevelopment	Data harmonization, computational complexity

2.1. Neuroimaging-Based AI Models

2.1.1. Structural MRI and AI

Structural Magnetic Resonance Imaging (sMRI) is one of the most extensively used modalities in neuroscience, providing high-resolution insights into brain morphology. In the context of neurodevelopmental diagnostics, AI-driven analysis of sMRI data has become a prominent approach for identifying early neuroanatomical alterations associated with NDDs. Convolutional neural networks (CNNs), a subset of deep learning models, have shown remarkable success in differentiating between individuals with and without ASD by extracting spatial hierarchies of features from volumetric sMRI data. According to Heinsfeld et al. [24], a deep learning model trained on the ABIDE (Autism Brain Imaging Data Exchange) dataset was able to classify ASD with an accuracy of 70%, outperforming traditional ML algorithms in detecting subtle morphological differences in regions such as the precuneus and the inferior parietal lobule.

More recently, advances in explainable AI (XAI) have begun to bridge the gap between diagnostic accuracy and clinical interpretability. Models using saliency maps and layer-wise relevance propagation can now highlight specific brain regions contributing to classification decisions, thereby allowing clinicians to interpret the neurobiological basis of the algorithm's output. A study by Sherkatghanad et al. [25] employed a 3D CNN architecture to analyze structural differences in ASD patients and identified significant features in the corpus callosum and cerebellum, regions previously implicated in ASD etiology. These findings not only validate the model's predictions but also support its utility in revealing candidate biomarkers for further investigation.

2.1.2. Functional MRI and AI

Functional MRI (fMRI), which measures brain activity by detecting changes in blood oxygenation, offers complementary insights into neural connectivity and functional organization. AI algorithms have been particularly effective in analyzing resting-state fMRI (rs-fMRI) data, which captures spontaneous brain activity patterns that are altered in many NDDs [26]. Functional connectivity matrices derived from rs-fMRI are typically high-dimensional and noisy, making them ideal candidates for dimensionality reduction and pattern recognition via AI. For instance, Khosla et al. [27] developed a graph-based deep learning model that classified ASD from rs-fMRI data with over 75% accuracy by modeling inter-regional connectivity as graph features.

Furthermore, attention-based deep learning models have been introduced to highlight functional networks that are most predictive of disorder status. Recent work by Ahmadi et al. [28] utilized a temporal attention mechanism in their deep learning architecture to focus on dynamic changes in connectivity patterns across time, significantly improving ADHD diagnosis. Such models are invaluable not only for their predictive capability but also for advancing our understanding of dynamic functional connectivity and its role in developmental psychopathology. This paradigm shift—from static to dynamic brain network modeling—could help identify transient but clinically relevant disruptions in neural coordination, especially in early stages of development.

2.1.3. Diffusion Tensor Imaging and AI

Diffusion Tensor Imaging (DTI), which measures the directionality and integrity of white matter tracts, has also been integrated with AI to detect microstructural changes indicative of NDDs. Alterations in white matter connectivity are hallmarks of disrupted neurodevelopment and often precede overt behavioral symptoms [29,30]. Traditional methods of analyzing DTI data—such as fractional anisotropy (FA) or mean diffusivity (MD)—are now being enhanced by machine learning classifiers, including support vector machines (SVMs) and ensemble tree methods. A study by Lou et al. [31] used DTI-based features with a random forest classifier to distinguish children with developmental language disorder (DLD), achieving an accuracy of 82% and revealing connectivity deficits in arcuate and superior longitudinal fasciculi.

Beyond conventional classification, unsupervised learning approaches such as autoencoders are now being used to discover latent white matter phenotypes in mixed diagnostic populations. These techniques can cluster patients into subtypes based on neural profiles, thus paving the way for a transdiagnostic framework of NDDs. Moreover, explainable DTI-AI models are now enabling clinicians to visualize which tracts contribute most to classification, enhancing the models' clinical utility and trustworthiness [26,30].

2.1.4. Multimodal Neuroimaging and AI Integration

While single-modality analyses provide valuable insights, integrating multiple neuroimaging modalities has been shown to significantly enhance diagnostic performance. AI models designed for multimodal learning can simultaneously process structural, functional, and diffusion data, capturing a richer representation of the developing brain. For example, Akhavan Aghdam et al. [32] developed a multimodal deep learning model that combined sMRI and rs-fMRI data for early ASD detection, achieving an AUC of 0.87. Their model leveraged late-fusion techniques to combine embeddings from both modalities, revealing synergistic patterns of anatomical and functional abnormalities.

Multimodal AI models also allow for longitudinal tracking of neurodevelopment, offering predictive insights into symptom progression and treatment outcomes. More recently, transformer-based architectures have been introduced to align temporally asynchronous imaging modalities, enabling end-to-end learning of developmental trajectories. Integrating neuroimaging data with clinical and behavioral features via AI represents a frontier in precision psychiatry, allowing for truly individualized diagnosis and care planning [22, 28]. However, such integration raises new challenges in data harmonization, model generalizability, and interpretability—issues that ongoing research is actively addressing through federated learning and cross-site validation protocols [31].

2.2. Behavioral and Speech Analysis via AI

The behavioral phenotype of neurodevelopmental disorders often precedes detectable structural brain abnormalities, making behavioral and speech analysis crucial for early diagnosis. Advances in artificial intelligence have enabled the extraction of granular, high-dimensional features from video, audio, and sensor data, revealing subtle behavioral and linguistic patterns that are imperceptible to the human eye or ear [33]. Such models have shown particular promise in pediatric populations, where non-invasive and low-burden diagnostic tools are critically needed. AI-based behavioral analytics can enhance early screening, stratify risk levels, and supplement clinician-administered diagnostic tools such as the Autism Diagnostic Observation Schedule (ADOS) [33,34]. This section explores four key areas where AI has significantly advanced behavioral and speech analysis: facial expression and gaze tracking, motor behavior and gesture analysis, speech and prosody modeling, and the integration of multimodal behavioral cues.

2.2.1. Facial Expression and Gaze Tracking

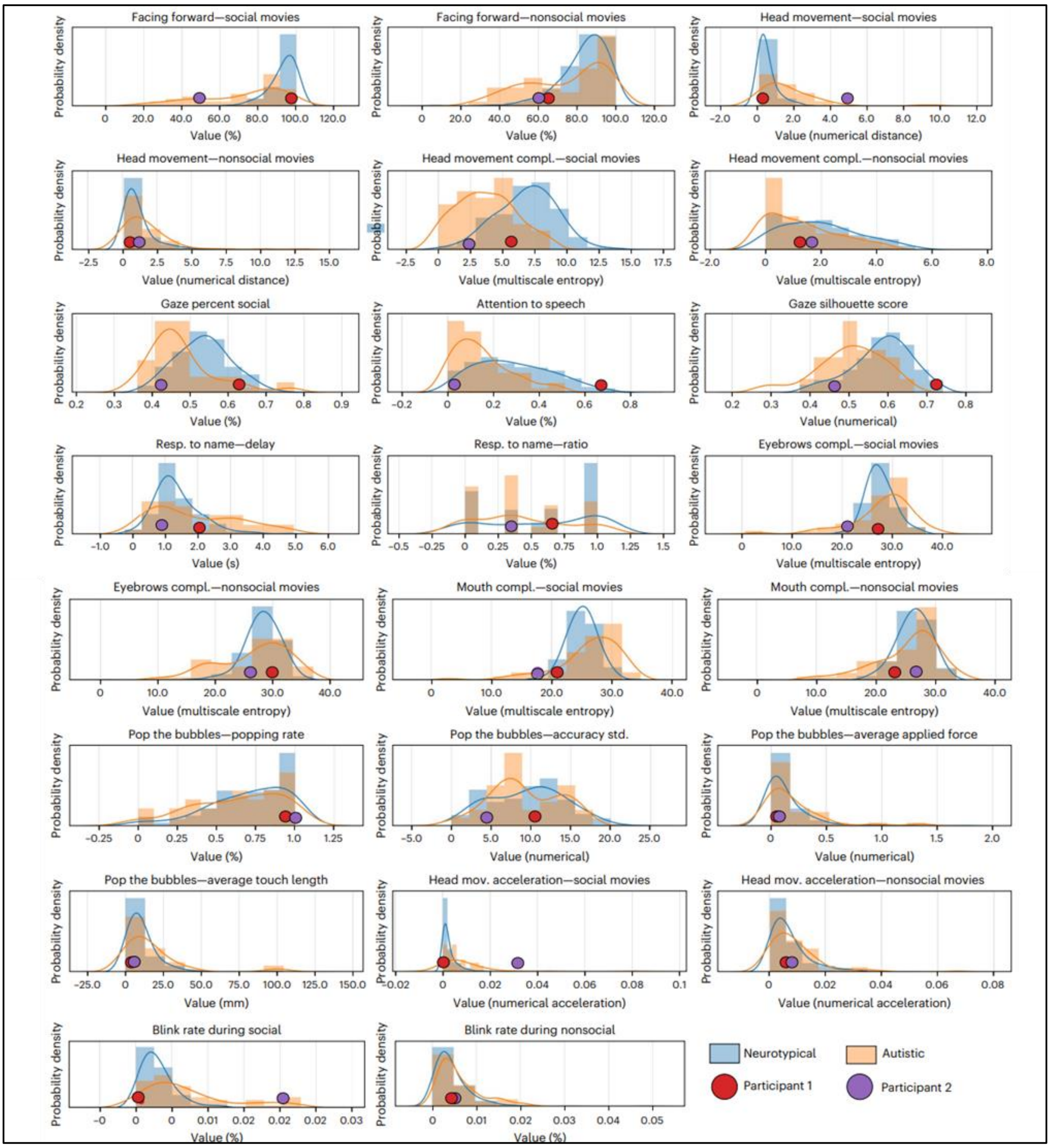

Figure 1 Distributions of app-derived behavioral variables, such as gaze fixation and facial expression intensity, for ASD and non-ASD groups, highlighting discriminative features for diagnosis. Adapted from Perochon et al. [36], with permission. Copyrighth, NatureMedicine 2023, licensed under a Creative Commons Attribution 4.0 International License

AI-powered computer vision systems have transformed facial expression and eye-gaze tracking into powerful diagnostic tools for neurodevelopmental disorders. Children with ASD, for example, often show reduced eye contact, atypical facial affect, and delays in the development of social reciprocity—behaviors that can now be quantitatively measured using AI. Deep convolutional networks trained on large datasets such as AffectNet or FER2013 can detect facial action units and microexpressions in real-time video, distinguishing between neurotypical and atypical emotional responses [35].

A notable study by Perochon et al. [36] used computer vision to track facial expressions and gaze patterns in infants and toddlers during structured play. Their model could detect early signs of ASD with a sensitivity of over 80%, well before a formal clinical diagnosis would typically occur. The system identified shorter gaze durations and lower frequency of shared attention episodes as significant predictors. Digital phenotyping via mobile apps enables the collection of rich behavioral data, such as gaze and facial expressions, for ASD diagnosis. Figure 1 illustrates the distributions of key app-derived variables, including gaze fixation and facial expression intensity, which differentiate children with ASD from non-ASD peers. By visualizing these feature distributions, clinicians can better understand the behavioral markers driving AI predictions, enhancing diagnostic transparency. Similarly, Alvari et al. [37] leveraged AI models to classify emotional responses from facial micro-expressions in high-risk infants, finding distinct temporal patterns associated with later ASD outcomes. These results underscore the feasibility of using non-intrusive, AI-enhanced video analysis for preclinical risk assessment.

2.2.2. Motor Behavior and Gesture Analysis

Motor abnormalities—including delays in gross and fine motor skills, stereotypic movements, and poor postural control—are increasingly recognized as early indicators of NDDs [38]. Traditional motor assessments are often qualitative and subjective, but AI systems using pose estimation and time-series modeling have revolutionized how motor behavior is analyzed. By extracting joint trajectories from video or wearable sensors, machine learning algorithms can quantify gait, reach patterns, hand flapping, and other repetitive behaviors with high temporal precision [39].

For example, the work of Leo et al. [40] utilized an AI pipeline that combined OpenPose-based skeletal tracking with recurrent neural networks to identify atypical movement patterns in infants as young as 6 months. The model achieved classification accuracies of 85% for ASD vs. control groups and showed strong predictive power for future developmental delays. Similarly, Crippa et al. [41] applied a logistic regression model to motion capture data and were able to distinguish children with ASD from controls based on kinematic features like variability in trunk posture and hand trajectory smoothness. These approaches are particularly promising for use in home environments, where smartphone-recorded videos could replace clinical motion labs, reducing both cost and access barriers.

2.2.3. Speech, Prosody, and Language Modeling

Speech and language disruptions are hallmark features of many neurodevelopmental conditions, particularly ASD and language delay disorders. AI has enabled the fine-grained analysis of speech and prosody—including pitch variation, speech rate, pause duration, and vocal intensity—features that are challenging for clinicians to assess with consistency. Recent models use recurrent neural networks and transformers to analyze time-series speech signals, often achieving impressive diagnostic sensitivity and specificity.

A landmark study by Bone et al. [42] demonstrated that prosodic abnormalities in children with ASD—such as flat intonation and abnormal pausing—could be accurately modeled using support vector machines fed with Mel-frequency cepstral coefficients (MFCCs). The system identified ASD with over 75% accuracy, indicating that acoustic features alone hold strong diagnostic value. More recently, Liu et al. [43] trained a transformer-based language model to parse semi-structured narrative tasks from children aged 3–6 years, revealing that lexical diversity and syntactic complexity were significantly reduced in those later diagnosed with ASD or language impairment. These systems have the added benefit of being scalable and non-invasive, potentially allowing for speech-based screening tools deployable via mobile apps or telehealth platforms.

2.2.4. Multimodal Behavioral Integration and AI Fusion Models

One of the most promising directions in behavioral AI is the integration of multimodal data—combining visual, auditory, and sensor-derived features to build comprehensive behavioral profiles. Multimodal fusion models can capture cross-domain interactions that are often missed in unimodal systems, such as the coordination between speech timing and facial affect or between gaze and gesture synchrony [44,45]. Fusion techniques range from early integration (where features are concatenated before modeling) to late fusion (where decisions from separate models are combined).

Advanced architectures like hybrid LSTM-CNN or attention-based fusion models are now widely used in this domain [45].

A study by Chen et al. [46] from the Autism Behavior Imaging project demonstrated that a fusion model combining facial expressions, vocal prosody, and head orientation could outperform any unimodal model in predicting ASD status, achieving 88% accuracy. Furthermore, this approach enhanced the interpretability of the AI output, allowing clinicians to visualize cross-modal inconsistencies—for instance, a smiling face paired with flat prosody—which are often clinically meaningful in ASD. Another emerging area is the use of self-supervised learning to reduce dependence on labeled data, enabling models to learn joint representations of behavior from raw videos without extensive human annotation [47]. These multimodal systems have the potential to become core components of next-generation diagnostic platforms, especially as they are increasingly integrated into mobile and home-based settings. However, challenges remain in terms of data standardization, real-time processing, and ensuring model robustness across cultural and demographic variability. Addressing these limitations will be critical for transitioning these tools from research prototypes to clinical applications.

2.3. Digital Phenotyping and Passive Monitoring

Digital phenotyping refers to the real-time, continuous quantification of human behavior and physiology using data collected passively from digital devices such as smartphones, wearables, and Internet-of-Things (IoT) systems [48,49]. In the context of neurodevelopmental disorders (NDDs), this paradigm offers transformative possibilities by capturing ecologically valid indicators of social interaction, motor activity, sleep, communication patterns, and emotional regulation in naturalistic settings [50,51]. Unlike episodic clinical evaluations, passive monitoring allows for longitudinal tracking of behavioral fluctuations, contextual responsiveness, and treatment effects, providing a high-resolution digital mirror of individual functioning [49]. As mobile penetration increases globally, these methods also hold potential for bridging the diagnostic gap in under-resourced settings.

2.3.1. Smartphone-Based Sensing of Social and Behavioral Patterns

Smartphones, equipped with sensors such as GPS, accelerometers, microphones, and touchscreen logs, have emerged as powerful tools for detecting behavioral irregularities associated with NDDs [52,53]. AI algorithms can process call frequency, text complexity, screen usage, mobility patterns, and geospatial routines to infer social withdrawal, communication deficits, and restricted interests—common in conditions such as autism spectrum disorder (ASD) and attention-deficit/hyperactivity disorder (ADHD) [53].

Saeb et al. [54] demonstrated that passive data from GPS and phone usage strongly correlated with depressive symptom severity, with later studies extending these methods to developmental disorders [55]. Other research has shown passive smartphone data can predict related mental health outcomes, such as social anxiety, with up to 85% accuracy [56].

2.3.2. Wearable Devices for Sleep, Activity, and Physiological Monitoring

Wearables such as smartwatches and fitness bands enable continuous measurement of physiological and behavioral states relevant to neurodevelopmental pathology. Features such as heart rate variability, galvanic skin response, step count, circadian rhythm regularity, and sleep architecture are now being captured with increasing accuracy [57]. These biosignals can serve as proxies for anxiety regulation, sensory sensitivities, motor restlessness, and attention variability.

Research by Goodwin et al. [59] used data from Empatica E4 wristbands to model autonomic arousal and stereotypic behaviors in children with ASD. Their system, integrating electrodermal activity and motion data, predicted behavioral escalation events (e.g., meltdowns) with 84% accuracy. Similarly, Kim et al. [60] trained deep learning models on heart rate and activity data from wearable devices to classify children with ADHD versus neurotypical peers, achieving 78% accuracy. These results suggest that AI-enhanced wearables can serve not only as diagnostic adjuncts but also as dynamic trackers of treatment efficacy and environmental triggers, enabling personalized intervention timing.

2.3.3. Voice and Communication Metadata from Passive Sensing

Beyond active speech analysis, passive voice monitoring via ambient microphones or smartphone sensors provides a low-burden method to capture patterns in vocalization frequency, prosodic variability, and interaction latency [61]. These features reflect social engagement, verbal initiative, and affective state—core domains affected in NDDs. AI tools can analyze turn-taking behavior, duration of silence, and pitch contours to detect deviations from typical developmental trajectories.

A study by Oller et al. [62] utilized the LENA (Language Environment Analysis) system to capture day-long audio recordings in natural home settings. They found that children with ASD exhibited significantly lower conversational reciprocity and spontaneous vocalizations compared to their neurotypical peers. Machine learning classifiers trained on LENA features achieved over 90% accuracy in distinguishing high-risk toddlers from controls. This passive approach has since been adapted to smartphone-integrated apps, allowing for longitudinal vocal development tracking without the need for structured testing environments.

2.3.4. Ecological Momentary Assessment (EMA) and Digital Diaries

Digital phenotyping also includes EMA tools and digital self-report diaries that prompt caregivers or patients to report symptoms, contexts, and mood states in real time or at randomized intervals [63,64]. While not strictly passive, EMA minimizes recall bias and aligns with the temporal granularity of passive sensing. When combined with AI, EMA data enhances context interpretation, enabling joint modeling of subjective experiences and sensor-based observations.

Lindhiem et al. [65] demonstrated that EMA reports of irritability and attention lapses, combined with passive activity and location data, improved the prediction of ADHD symptom severity by 15% compared to sensor data alone. Additionally, EMA responses have been linked with physiological data to detect stress patterns, creating a comprehensive bio-behavioral profile [66]. In the context of NDDs, where symptoms fluctuate based on environment and time of day, EMA can guide clinicians in tailoring intervention timing and monitoring therapy response [67].

Despite their promise, digital phenotyping approaches face key challenges including data privacy concerns, variability in device use across populations, and the need for regulatory frameworks that ensure clinical robustness and equity. However, the integration of these technologies into digital health ecosystems marks a major step toward scalable, personalized, and continuous neurodevelopmental monitoring.

3. Foundations of Explainable AI (XAI)

The rapid integration of artificial intelligence into critical fields such as medicine, finance, and policy-making has sparked a vital demand for transparency in machine learning decision-making. In healthcare, particularly in the context of neurodevelopmental disorders (NDDs), clinical experts must be able to understand how AI models derive conclusions from patient data, especially when such decisions could significantly affect diagnostic and therapeutic outcomes [68]. Explainable Artificial Intelligence (XAI) arises as a response to this need, referring to a set of techniques and principles that make the outcomes of AI models interpretable and understandable to humans [69]. This interpretability is not merely a technical luxury but a functional necessity for ethical compliance, legal accountability, and human trust in AI-powered decisions. It aims to bridge the gap between model performance and human cognition, ensuring that clinicians, data scientists, and even patients can scrutinize the rationale behind AI predictions [68]. As AI models, especially deep neural networks, grow in complexity and opacity, the importance of robust, reliable explainability mechanisms becomes even more critical to mitigate risks such as bias, overfitting, or unpredictable behavior in real-world clinical settings.

3.1. What Is Explainability in AI?

Explainability in AI refers to the ability to make a model’s internal logic, decision pathways, and feature relevance comprehensible to human users. It enables stakeholders to trace back decisions to the features or data patterns that most influenced the outcome [70,71]. This concept is foundational to the development and application of AI systems in sensitive domains where decisions must be justified, audited, and trusted. In the medical sciences, for example, explainability allows clinicians to interrogate how and why a model has suggested a certain diagnosis or predicted a risk score, thereby facilitating informed decision-making rather than blind reliance on algorithmic output [70]. Researchers such as Chander et al. [72] emphasize that explainability not only aids in trust calibration and user engagement but also plays a crucial role in detecting model biases and vulnerabilities, making AI systems more robust and reliable in practice. Moreover, the demand for explainability is not merely philosophical but driven by pragmatic and regulatory pressures. Legal frameworks such as the General Data Protection Regulation (GDPR) in Europe and growing calls for algorithmic transparency in the United States demand that AI systems be auditable and interpretable [73]. In clinical neuroscience, explainable systems help identify whether a model's predictions align with pathophysiological understandings of disorders such as autism spectrum disorder (ASD) or attention deficit hyperactivity disorder (ADHD), ultimately enhancing the credibility and applicability of computational diagnostics [74]. Without explainability, these models become "black boxes" whose predictions, however accurate, may remain untrusted, unused, or even potentially dangerous if misinterpreted or applied inappropriately. Thus, explainability is not a peripheral add-on but a central pillar in the design and deployment of trustworthy and human-centric AI systems [74].

3.2. Types of Explanations: Post Hoc vs. Intrinsic

Explainability in AI systems can be classified into two broad paradigms: intrinsic and post hoc explanations. Intrinsic explanations refer to the transparency that is built into the model itself. These models are constructed with interpretability as a core feature, such that their decision-making process is directly observable from their structure. Linear regression models, for instance, offer coefficients that represent the exact contribution of each input feature to the predicted outcome, while decision trees utilize a sequence of human-readable decision rules to guide prediction [75]. These models are often favored in clinical contexts where clarity is paramount, even if their simplicity may sometimes come at the cost of predictive accuracy. Ribeiro et al. [76] observed that while such models may underperform compared to more complex architectures like deep learning networks, their interpretability can make them more suitable for certain high-stakes applications, including early childhood diagnostic assessments.

In contrast, post hoc explanations are applied to models that are not intrinsically interpretable. These include deep neural networks, ensemble methods, and other complex architectures that, despite their predictive power, operate in a manner that is opaque to human observers [77]. Post hoc explanation techniques are developed to analyze these models after training, offering insights into which input features were most influential or how a particular prediction was made [77,78]. This paradigm allows researchers to extract human-understandable rationales from otherwise inscrutable algorithms, enabling the use of high-performance models in domains where interpretability is a prerequisite. However, post hoc techniques come with limitations. They may offer approximations rather than faithful representations of the model's internal reasoning, and they can sometimes yield explanations that are misleading or unstable. As explained by Lipton [79], this raises important questions about the epistemic value of such explanations and the extent to which they should be relied upon in sensitive contexts such as clinical decision-making.

In the domain of neurodevelopmental disorders, where early detection may depend on subtle behavioral or neuroimaging markers, both intrinsic and post hoc models offer different advantages. Intrinsic models can support transparent clinical protocols, while post hoc techniques enable the use of sophisticated deep learning methods without entirely sacrificing interpretability [80,81]. Consequently, the current trajectory of XAI research involves not just the development of new explanation tools but also the careful selection and combination of explanatory approaches that balance interpretability with model performance, depending on the context of use and the stakes involved.

3.3. Common XAI Techniques: SHAP, LIME, Grad-CAM, Attention Maps

A growing toolkit of explanation techniques has emerged to operationalize XAI in practice, each with its own strengths, limitations, and ideal use cases. SHAP (SHapley Additive exPlanations) stands out as one of the most theoretically grounded methods, drawing from cooperative game theory to assign each feature a Shapley value that quantifies its contribution to the model’s output [82,83].

Table 2 Common XAI Techniques in Neurodevelopmental Disorder Diagnostics

XAI Technique	Mechanism	NDD Application	Key Studies	Strengths	Limitations
SHAP (Shapley Additive Explanations)	Assigns feature importance based on game theory Shapley values	Identifies key EEG frequency bands in ASD, ADHD feature attribution	Lundberg & Lee [82], Rogala et al. [98]	Model-agnostic, consistent feature importance	Computationally intensive, sensitive to input perturbations
LIME (Local Interpretable Model-agnostic Explanations)	Perturbs inputs to create local surrogate models	Explains individual ASD risk scores from behavioral data	Ribeiro et al. [76], Perochon et al. [36]	Case-specific, intuitive for clinicians	Local approximations may not reflect global model behavior
Grad-CAM (Gradient-weighted Class Activation Mapping)	Uses gradients to highlight influential regions in visual inputs	Visualizes brain regions (e.g., prefrontal cortex) in fMRI-based ASD diagnosis	Lin et al. [97], Hussain & Shouno [84]	Effective for imaging data, visually intuitive	Limited to convolutional architectures, less effective for non-visual data

Attention Maps	Visualizes attention weights in transformer models	Highlights temporal sequences in EEG or behavioral data for ADHD clustering	Ahmadi et al. [28], Jacobson et al. [104]	Intuitive for sequential data, aligns with temporal dynamics	Can be unstable, requires validation for reliability
Decision Trees	Provides rule-based decision paths	Explains behavioral feature contributions (e.g., gaze, gestures) in ASD detection	Perochon et al. [36], Arnett et al. [100]	Highly interpretable, clinician-friendly	Limited to simpler models, may miss complex patterns
Layer-wise Relevance Propagation	Propagates relevance scores backward through neural layers	Identifies critical brain regions in sMRI for ASD classification	Sherkatghanad et al. [25]	Fine-grained feature attribution	Complex to implement, model-specific

This method is model-agnostic and ensures consistency in feature importance estimates across different models and inputs. Lundberg and Lee [82], who introduced SHAP, demonstrated its utility in medical diagnostics, showing that it could effectively explain black-box predictions of patient risk factors in a way that aligned with clinical intuition. In the context of neurodevelopmental disorders, SHAP has been applied to analyze complex models built on behavioral data or neuroimaging inputs, offering insight into which features—such as gaze fixation patterns or cortical thickness measures—most influence the model's classification [82].

Another widely used technique is LIME (Local Interpretable Model-agnostic Explanations), which generates local approximations of the model's behavior around a specific prediction. LIME works by perturbing the input features and observing the resulting changes in output, fitting an interpretable model—such as a linear regressor—to the neighborhood of the original instance [76,83]. This approach provides case-specific explanations, making it particularly useful for individual-level predictions, such as determining why a particular patient was flagged as high risk for ASD. Ribeiro et al. [76] illustrated that LIME could be instrumental in settings where clinicians must justify or interrogate single-instance model outputs, enhancing transparency without sacrificing model complexity.

For models trained on visual data, such as neuroimaging scans or facial behavior videos, techniques like Grad-CAM (Gradient-weighted Class Activation Mapping) have become essential. Grad-CAM works by leveraging the gradients of the output layer with respect to the convolutional layers to produce heatmaps that highlight areas of the input image most influential to the prediction [84,85]. This has been especially impactful in studies using fMRI or EEG scans to classify neurodevelopmental conditions, where it is crucial to identify which brain regions the model is attending to. Similarly, attention maps derived from transformer-based architectures provide interpretability for models processing sequential data, such as time-series signals or language transcripts [86]. These maps show how much "attention" the model allocates to each input token or segment, offering a visual and quantitative insight into the temporal and contextual weighting mechanisms employed by the model. To elucidate the practical utility of XAI techniques in neurodevelopmental diagnostics, Table 2 provides a detailed comparison of common XAI methods, their mechanisms, applications in NDD contexts, and their strengths and limitations. This table underscores how these techniques enhance interpretability in complex AI models, facilitating their integration into clinical workflows.

Collectively, these XAI techniques are enabling researchers and clinicians to peer into the inner workings of AI systems, converting complex predictive processes into understandable narratives. Their use in the context of neurodevelopmental disorders is growing, particularly as AI tools are increasingly deployed to analyze high-dimensional, multimodal datasets. However, the use of these tools must be tempered by a critical understanding of their limitations, as even the most elegant explanations can mislead if they fail to faithfully represent the underlying model behavior.

4. Applications of XAI in Neurodevelopmental Disorder Diagnostics

The application of explainable AI (XAI) within the realm of neurodevelopmental disorder (NDD) diagnostics marks a significant turning point in computational medicine [87]. Historically, the diagnosis of conditions like autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), and developmental language disorder has relied heavily on clinical observation, caregiver reports, and time-consuming psychometric evaluations [87-89]. These methods, though valuable, are prone to subjectivity, inter-rater variability, and delays that can critically impact early intervention outcomes [88]. With the advent of machine learning, particularly deep learning models applied to neuroimaging, behavioral data, and multi-omics, significant improvements in predictive performance have emerged [90]. However, the opaque nature of these models has limited their direct integration into clinical practice. XAI fills this gap by translating complex model inferences into interpretable insights, thereby empowering clinicians with not only accurate but also transparent diagnostic tools.

4.1. Enhancing Trust and Clinical Adoption

One of the most direct impacts of XAI in the context of NDD diagnostics is the enhancement of clinician trust and the promotion of clinical adoption of AI tools. The adoption of AI-based systems in psychiatry and pediatric neurology remains relatively cautious, often hindered by skepticism regarding how these systems reach their conclusions. Studies such as those by Holzinger et al. [91] emphasize that interpretability is not only about technical explanation but also about aligning machine predictions with domain knowledge, which is essential for building trust among healthcare professionals. For instance, when an AI model identifies a child as being at high risk for ASD, clinicians are more likely to accept the recommendation if the model provides a transparent rationale, such as highlighting atypical eye-tracking patterns or reduced connectivity in specific brain regions known to be associated with ASD [92,93]. These explanations not only validate the model's prediction but also enhance the clinician's diagnostic confidence and willingness to use AI tools as complementary supports.

Moreover, explainability aids in identifying edge cases where the model's prediction might be unreliable. For example, Saporta et al. [94] demonstrated that XAI tools could reveal when an AI system's decision was based on spurious correlations, such as image artifacts or irrelevant demographic features. This ability to scrutinize and challenge the model's reasoning is crucial in ensuring safe deployment, especially in pediatric populations where ethical and legal standards are particularly stringent. In sum, XAI not only serves as a safeguard against algorithmic opacity but also as a bridge that connects advanced computational methods with the clinical intuition and accountability required in real-world practice.

4.2. Case Studies of XAI in Diagnosing Neurodevelopmental Disorders (NDDs)

The integration of explainable artificial intelligence (XAI) into diagnostic pipelines for neurodevelopmental disorders (NDDs) has transitioned from theoretical exploration to practical implementation [74]. With the increasing availability of neuroimaging and behavioral data, researchers are leveraging XAI not only to enhance diagnostic accuracy but to illuminate the reasoning behind AI-generated predictions. This section explores three key domains where XAI has significantly impacted the diagnosis and characterization of NDDs: autism spectrum disorder (ASD), attention-deficit/hyperactivity disorder (ADHD), and behavioral pattern analysis in children. Each domain highlights how explainability tools enable more transparent and clinically grounded decisions, helping bridge the gap between complex machine learning models and real-world healthcare needs. To consolidate the practical impact of XAI in NDD diagnostics, Table 3 presents a selection of case studies that demonstrate how specific XAI techniques have been applied to different data modalities for ASD, ADHD, and behavioral analysis. These examples highlight the clinical relevance and interpretability of XAI-driven insights, reinforcing their potential to transform diagnostic practices.

Table 3 Case Studies of XAI Applications in NDD Diagnosis

Study	NDD	Data Modality	XAI Technique	Key Findings	Clinical Impact	Reference
Lin et al. (2022)	ASD	fMRI	Grad-CAM	Highlighted posterior cingulate and medial prefrontal cortex in ASD classification	Validated neurobiological markers, enhanced clinician trust	[97]
Rogala et al. (2023)	ASD	EEG	SHAP	Identified alpha/theta rhythm abnormalities in	Improved EEG-based diagnostic precision,	[98]

				frontal/temporal lobes	actionable for early screening	
Arnett et al. (2022)	ADHD	Neuropsychological Tests	Decision Trees	Clustered ADHD subtypes based on impulsivity and attention deficits	Informed tailored treatment plans, reduced diagnostic subjectivity	[100]
Agoalikum et al. (2023)	ADHD	fMRI	Attention Maps	Highlighted dorsolateral prefrontal cortex in impulsive symptom prediction	Clarified neural basis of ADHD subtypes, supported targeted interventions	[102]
Perochon et al. (2023)	ASD	Behavioral (Video)	Decision Trees, SHAP	Identified gaze and gesture deficits as key ASD predictors	Enabled non-invasive, scalable screening, actionable for preclinical risk assessment	[36]
Jacobson et al. (2022)	Anxiety/ Behavioral Dysregulation	Wearable Sensors	Attention Maps	Highlighted sleep restlessness as a predictor of anxiety	Facilitated continuous monitoring, personalized intervention timing	[104]

4.2.1. XAI in ASD Diagnosis Using fMRI or EEG

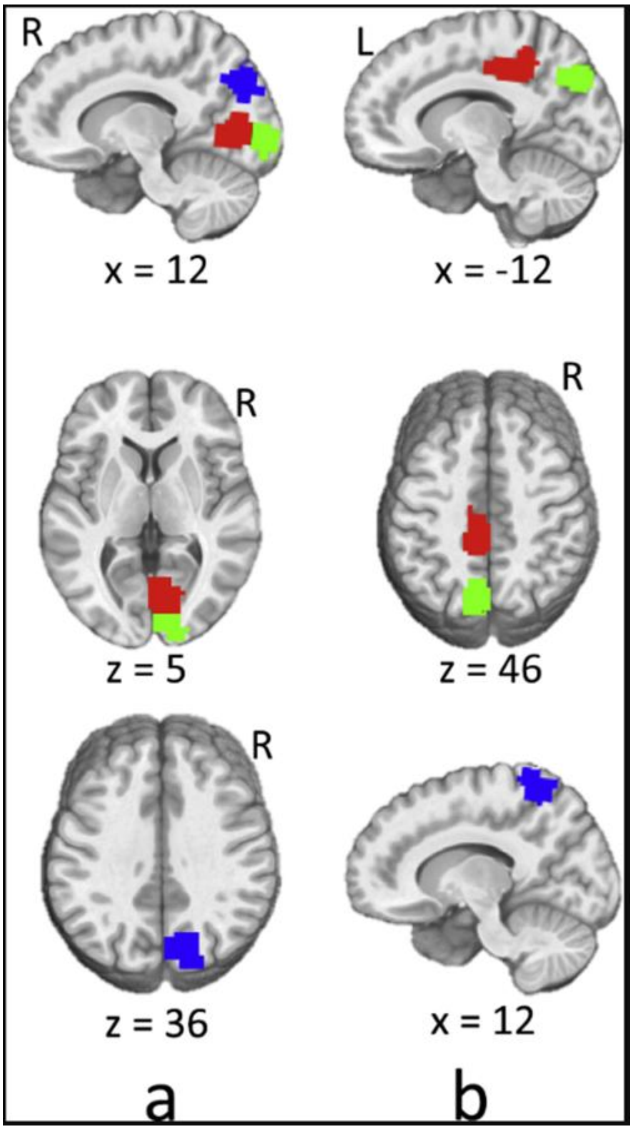

Figure 2 Class activation map highlighting brain regions critical for ASD classification in fMRI data, including the superior temporal sulcus and prefrontal cortex. Adapted from Heinsfeld et al. [24] with permission. Copyright, Elsevier 2018

Autism spectrum disorder (ASD) presents unique challenges in diagnosis due to its heterogeneous manifestations and the absence of definitive biological markers. Recent efforts have focused on leveraging functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) data within deep learning frameworks to identify neural signatures of ASD [95,96]. However, the opaque nature of many of these models has limited their clinical acceptance. To address this, researchers have increasingly incorporated XAI techniques into ASD classification workflows. For example, Lin et al. [97] applied a convolutional neural network (CNN) to resting-state fMRI data and used Grad-CAM (Gradient-weighted Class Activation Mapping) to visualize the regions of interest that influenced the model's predictions. Their analysis revealed that altered connectivity in the posterior cingulate cortex and medial prefrontal cortex were significant contributors to ASD classification—insights that aligned with prior neurobiological research on ASD-related network disruptions. Class activation maps, such as Grad-CAM, enable clinicians to visualize the neural underpinnings of AI-driven ASD diagnoses. Figure 2 presents a class activation map overlaid on an fMRI scan, highlighting regions like the superior temporal sulcus and prefrontal cortex that drive ASD classification. Such interpretable visualizations enhance the clinical utility of AI by aligning predictions with neurobiological evidence.

Similarly, EEG-based models have employed SHAP (SHapley Additive exPlanations) values to highlight the most informative frequency bands and channels. In a study by Rogala et al. [98], interpretable deep learning models were trained on EEG recordings from toddlers with suspected ASD. SHAP analysis revealed that abnormalities in alpha and theta rhythms, particularly in the frontal and temporal lobes, were pivotal for ASD discrimination. These findings not only validated known electrophysiological differences in ASD but also helped clinicians understand the neural dynamics captured by the model. These case studies illustrate how XAI enables the alignment of computational inferences with neurobiological knowledge, enhancing the trustworthiness and interpretability of AI-driven ASD diagnostic tools.

4.2.2. Interpretable AI for ADHD Symptom Clustering

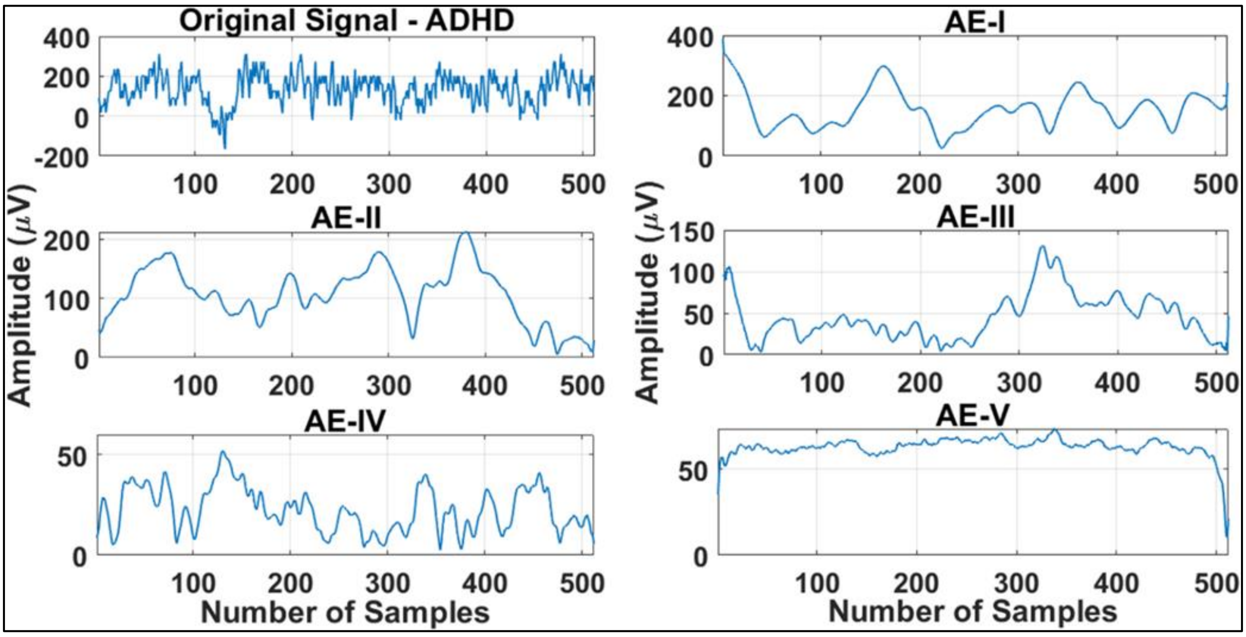

Figure 3 SHAP summary plot illustrating the contribution of EEG features to ADHD classification, highlighting alpha and theta band power as key predictors. Adapted from Khare and Acharya [124] with permission.Copyright, Elsevier 2023.

In contrast to ASD, attention-deficit/hyperactivity disorder (ADHD) is often diagnosed based on behavioral checklists and clinician observations, which introduces subjectivity and inter-rater variability [99]. To mitigate these issues, machine learning approaches are being employed to detect data-driven symptom clusters. Yet, the black-box nature of many clustering and classification algorithms remains a barrier to clinical adoption.

Recent studies have utilized interpretable machine learning frameworks to deconstruct ADHD symptoms into coherent subtypes with clear explanatory factors. For instance, Arnett et al. [100] used hierarchical clustering on neuropsychological test scores and incorporated decision tree algorithms to elucidate how features like impulsivity, response inhibition, and sustained attention differentiated subgroups. The resulting tree structures allowed clinicians to trace back the classification logic, thereby offering interpretable symptom groupings that could inform individualized treatment strategies. In another example, attention-based deep learning models were applied to fMRI data from children with ADHD. Researchers employed attention maps to highlight regions of the brain that contributed most to symptom categorization [101]. According to the findings of Agoalikum et al. [102], the dorsolateral prefrontal cortex and anterior cingulate cortex showed elevated attention weights when predicting impulsive versus inattentive symptom clusters. These results were presented in a visual format that clinicians could readily interpret, aligning well with existing neurological models of ADHD. By offering transparent logic and neurobiologically grounded interpretations, XAI is aiding the redefinition of ADHD not just as a unitary diagnosis but as a spectrum with distinct, explainable subtypes. To illustrate the practical utility of XAI in ADHD diagnostics, consider the application of SHAP to EEG-based classification models. Figure 3 presents a SHAP summary plot, which quantifies the impact of EEG features, such as alpha and theta band power, on the model's prediction of ADHD. This visualization not only highlights key neurophysiological markers but also enhances clinician trust by providing a transparent rationale for diagnostic decisions.

4.2.3. Behavioral Data Analysis in Children Using Interpretable Decision Trees or Attention-Based Models

Behavioral data—ranging from motor movement patterns to social interaction cues—represent a valuable, non-invasive source of information for early detection of neurodevelopmental disorders [103]. However, these data are often high-dimensional and temporally complex, making traditional statistical approaches inadequate. Interpretable AI models have emerged as promising tools to extract meaningful patterns while retaining transparency. A compelling application is the use of decision tree classifiers to analyze video-based behavioral data from preschool-aged children. In a landmark study by Perochon et al. [36], features such as gaze direction, gesture frequency, and joint attention episodes were encoded into a structured dataset. A CART (Classification and Regression Trees) algorithm was then trained to distinguish children at risk for ASD. The model's branching structure revealed that failure to respond to name calls and reduced gesture use were the most discriminative features—findings that were both interpretable and clinically actionable. In parallel, attention-based transformers have been applied to longitudinal behavioral datasets. In one recent study, Jacobson et al. [104] developed a model that processed wearable sensor data capturing motor activity and sleep patterns. The attention mechanism highlighted temporal sequences—such as prolonged restlessness during sleep—that contributed to anxiety or behavioral dysregulation diagnoses. These attention scores were visualized as heat maps, offering an intuitive explanation for clinicians and caregivers. These applications show how XAI can transform raw behavioral data into interpretable diagnostic cues, thereby empowering early intervention strategies while upholding transparency and accountability.

4.3. Personalizing Diagnostic Pathways

Another transformative role of XAI lies in its capacity to support personalized diagnostic pathways for individuals with neurodevelopmental conditions. Neurodevelopmental disorders are notoriously heterogeneous, often overlapping in symptomology and manifesting along diverse developmental trajectories [105]. Traditional diagnostic systems, though standardized, struggle to account for such individual variation. AI models trained on multimodal data—such as genomics, neuroimaging, and behavioral assessments—have demonstrated an ability to capture this heterogeneity [105,106]. However, XAI techniques are what make this granularity accessible and actionable for clinicians. For instance, in the study by Perochon et al. [36], SHAP was used to parse individualized diagnostic predictions for children undergoing ASD screening. The system highlighted specific behavioral indicators—such as joint attention deficits or speech delays—that were particularly influential in each child's risk score, allowing clinicians to tailor follow-up assessments accordingly. Rather than offering a one-size-fits-all classification, the XAI-enhanced system provided a transparent, patient-specific reasoning path, empowering a more nuanced and targeted diagnostic process [107]. This capability is particularly beneficial for children who fall into diagnostic "gray zones," where traditional criteria may not be fully met, but early intervention could still yield significant benefits.

Moreover, explainability supports shared decision-making between clinicians, caregivers, and patients. By clearly communicating the basis of a model's judgment, XAI fosters collaborative discussions and helps manage expectations. Parents, in particular, are often more receptive to AI-generated assessments when accompanied by understandable rationales, thereby improving compliance with recommended interventions and follow-up care [108]. As personalized medicine continues to evolve, the ability of XAI to illuminate the unique constellation of features driving each diagnostic decision will be key to realizing precision psychiatry for neurodevelopmental disorders.

5. Limitations and Current Challenges

Explainable AI (XAI) represents a pivotal advance in neurodevelopmental disorder diagnostics, promising transparency and interpretability in otherwise opaque machine learning models [109]. Yet, despite significant progress, numerous intrinsic limitations and practical challenges constrain its current utility. These challenges arise from the unstable nature of explanation methods, fundamental compromises between accuracy and interpretability, and pervasive data constraints common in pediatric neuroscience research. Addressing these limitations is crucial to achieving robust, clinically trustworthy AI systems that can be meaningfully integrated into neurodevelopmental care. To provide a structured overview of the challenges facing XAI in NDD diagnostics, Table 4 summarizes the key limitations, their implications, and potential mitigation strategies. This table aims to guide researchers and clinicians in addressing these hurdles to enhance the reliability and adoption of XAI systems in clinical practice.

Table 4 Limitations of XAI in Neurodevelopmental Disorder Diagnostics

Limitation	Description	Implication	Mitigation Strategies	Key References
Inconsistency in Explanation Reliability	Variability in SHAP/LIME outputs due to input noise or preprocessing differences	Undermines clinician trust, risks misinterpretation of feature importance	Standardize metrics for explanation stability, incorporate clinician feedback	[110], [111], [112]

Trade-off Between Accuracy and Interpretability	Complex models (e.g., CNNs) are accurate but opaque; simpler models are less accurate	Limits clinical adoption of high-performing models	Develop hybrid models, use surrogate interpretable models	[113], [114], [115]
Dataset Biases	Underrepresentation of diverse populations in pediatric datasets	Models may perpetuate biases, reducing generalizability	Use federated learning, synthetic data generation, and diverse cohort recruitment	[117], [118], [120]
Small Pediatric Cohort Sizes	Limited sample sizes due to ethical and recruitment constraints	Increases overfitting risk, reduces model robustness	Implement transfer learning, data augmentation, and multi-site data pooling	[117], [119]
Privacy Concerns	Sensitive pediatric data raises ethical and legal issues	Limits data sharing, hinders large-scale studies	Adopt federated learning, differential privacy techniques	[120], [126]
Lack of Standardized Metrics	No consensus on evaluating explanation quality or clinical relevance	Hinders comparison and validation of XAI methods	Develop universal benchmarks, involve clinicians in validation	[112]

5.1. Inconsistency in Explanation Reliability

The reliability of explanations generated by XAI techniques remains a profound concern, particularly given the high-stakes environment of neurodevelopmental disorder diagnosis. Most XAI frameworks rely on post hoc methods, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which approximate feature importance through perturbations or local surrogate models [110]. However, research has demonstrated that these approximations can exhibit significant variability depending on minor fluctuations in input data or model parameters [110]. For instance, in neuroimaging-based diagnostics of autism spectrum disorder (ASD), minor noise or slight preprocessing differences in brain scans have been shown to shift feature attribution maps, altering the interpretation of which brain regions contribute most to classification decisions [111]. This instability raises critical questions about the fidelity of such explanations and their alignment with actual neural correlates of disease pathology. Moreover, XAI methods sometimes yield contradictory explanations for similar cases or identical predictions, undermining confidence in their use as decision support tools.

This inconsistency is compounded by a lack of consensus on standard metrics for evaluating explanation reliability and clinical relevance. While some studies advocate for stability metrics, others emphasize human-centered validation through clinician feedback [112]. However, the absence of universally accepted benchmarks hinders progress, leaving it unclear which XAI approaches provide explanations that truly reflect meaningful and actionable insights. In clinical contexts where decisions may impact therapeutic interventions or prognostic counseling, such uncertainty can diminish the willingness of practitioners to rely on AI-supported diagnoses [112].

5.2. Trade-off Between Accuracy and Interpretability

A central technical dilemma in XAI for neurodevelopmental disorders is the inherent trade-off between model accuracy and interpretability. State-of-the-art diagnostic models often employ deep learning architectures like convolutional neural networks (CNNs) or graph neural networks (GNNs) that excel at capturing complex, nonlinear relationships in heterogeneous neurodevelopmental data, such as functional MRI, EEG, and behavioral assessments [113]. These models frequently outperform simpler, interpretable algorithms but do so at the cost of becoming “black boxes” whose internal decision logic is obscured. On the other hand, interpretable models such as logistic regression or decision trees offer clear, human-understandable decision boundaries but typically lack the nuanced representational capacity to address the multidimensional heterogeneity of neurodevelopmental disorders [114]. This creates a practical tension: prioritizing interpretability risks missing subtle but clinically relevant patterns, while maximizing predictive power sacrifices transparency and, consequently, clinician trust and regulatory approval.

Efforts to reconcile this trade-off include hybrid approaches, such as integrating inherently interpretable modules within deep networks or designing surrogate models that approximate complex models' behavior in an interpretable manner. For example, Li et al. [115] developed interpretable graph convolutional models for brain connectivity data in ASD diagnosis, balancing complexity and clarity. Nevertheless, these approaches remain in early stages and are often limited to specific data modalities or constrained datasets. Moreover, it remains unclear how well surrogate explanations generalize across diverse clinical scenarios. This trade-off also intersects with ethical and legal considerations. Regulatory agencies increasingly require explainability for AI-based medical devices, mandating a level of transparency that many high-performing models struggle to meet [116]. Thus, the challenge is not merely technical but also operational, requiring advances that deliver both rigorous performance and comprehensible explanations that clinicians and patients can trust.

5.3. Dataset Biases and Overfitting in Small Pediatric Cohorts

Data limitations present one of the most pervasive barriers to robust XAI applications in neurodevelopmental disorders. Pediatric neuroimaging and behavioral datasets typically suffer from small sample sizes due to recruitment difficulties, ethical constraints, and the low prevalence of many disorders. This scarcity increases the risk of overfitting, whereby AI models capture idiosyncratic noise or cohort-specific artifacts rather than generalized biomarkers [117].

Moreover, neurodevelopmental disorders display substantial demographic and phenotypic heterogeneity, including variations across age, sex, ethnicity, and socioeconomic background, which are frequently underrepresented or unevenly distributed in available datasets. For example, many prominent neuroimaging databases are predominantly composed of Western, Caucasian participants, limiting generalizability to global populations [118]. This biased data landscape leads to models—and consequently explanations—that reflect confounding variables rather than universal neurobiological features, potentially perpetuating disparities in diagnostic accuracy across subpopulations.

Overfitting is further exacerbated by the high dimensionality of neurodevelopmental data, which often includes tens of thousands of features relative to relatively small cohort sizes. Without adequate regularization or data augmentation strategies, models tend to memorize training data, limiting external validity [119]. Such overfitting undermines the clinical reliability of both model predictions and their explanations, as spurious correlations are misinterpreted as meaningful markers. To address these challenges, emerging methodologies such as federated learning, transfer learning, and synthetic data generation have been proposed to leverage multi-institutional data while preserving patient privacy and increasing dataset diversity [120]. However, these approaches introduce complexities in ensuring that explanations remain valid and interpretable across heterogeneous data sources. Furthermore, careful scrutiny is required to detect and mitigate embedded biases that could adversely impact underrepresented groups.

6. Future Directions

As artificial intelligence systems continue to make inroads into clinical neuroscience, particularly in pediatric neurodevelopmental diagnostics, the need to balance model performance with interpretability becomes even more urgent. Explainable AI (XAI) is not just a supplementary feature but an essential component that ensures transparency, trust, and ethical alignment in clinical contexts. Emerging research suggests that the next wave of advancement in this field will be driven by integrative frameworks that combine explainability with real-world clinical usability, multisite data collaboration, and active involvement of human experts [121,122]. This section explores the most promising directions in the evolution of XAI: integration with digital biomarkers, federated explainable learning, and human-in-the-loop designs. To chart the path forward for XAI in neurodevelopmental diagnostics, Table 5 outlines key future directions, their potential impact, and enabling technologies. This table highlights how emerging paradigms like digital biomarkers, federated learning, and human-in-the-loop systems can advance the field, ensuring ethical and effective deployment of AI in pediatric neuroscience.

Table 5 Future Directions for XAI in Neurodevelopmental Disorder Diagnostics

Future Direction	Description	Potential Impact	Enabling Technologies	Key References
Integration with Digital Biomarkers	Using XAI to interpret behavioral/physiological data from smartphones, wearables	Scalable, non-invasive early detection, personalized monitoring	SHAP, LIME, attention-based models, wearable sensors	[123], [124]

Federated Explainable Learning	Decentralized model training with local explanations across institutions	Improved generalizability, privacy-preserving multicenter studies	Federated SHAP, differential privacy, blockchain for data integrity	[125], [126]
Human-in-the-Loop Models	Incorporating clinician feedback into model training and explanation generation	Enhanced model relevance, improved trust and clinical alignment	Active learning, causal attention maps, co-design frameworks	[127], [128], [129]
Multimodal Data Fusion	Combining neuroimaging, behavioral, and genomic data with XAI	Comprehensive diagnostic profiles, improved accuracy	Transformers, late-fusion models, cross-modal attention mechanisms	[106], [122]
Real-time Monitoring Systems	Continuous tracking of NDD symptoms using XAI-enhanced digital phenotyping	Dynamic intervention timing, longitudinal tracking	Time-series models, edge computing, real-time SHAP	[49], [104]
Ethical and Regulatory Frameworks	Developing standards for XAI transparency and fairness	Ensures safe, equitable deployment, regulatory compliance	GDPR-compliant algorithms, standardized explanation metrics	[73], [116]

6.1. Integrating XAI with Digital Biomarkers

The fusion of XAI with digital biomarkers is poised to revolutionize early detection and individualized monitoring in pediatric neuroscience. Digital biomarkers—quantifiable physiological and behavioral data collected via digital devices—have shown substantial promise for tracking developmental milestones and deviations [123]. However, the integration of such high-dimensional, temporally varying data into AI models demands more than just predictive accuracy; it requires interpretability to be clinically viable. Recent studies have demonstrated the utility of XAI in identifying behavioral signatures from digital phenotyping tools such as smartphone usage, eye-tracking, and wearable sensors. For instance, Khare et al. [124] developed an interpretable AI system that leverages touchscreen interaction patterns to flag early signs of attention-deficit/hyperactivity disorder (ADHD). The model used SHAP (SHapley Additive exPlanations) to highlight which specific digital biomarkers—like reaction time variability or motor consistency—were most influential in the prediction, thereby providing actionable insights for clinicians.

Furthermore, explainability plays a critical role in establishing trust when AI inferences are based on novel biomarkers that have not yet been validated in traditional clinical settings. Clinicians require clear mappings between AI-driven outputs and established pathophysiological frameworks [122]. XAI bridges this gap by visually and semantically aligning algorithmic findings with known clinical markers, such as EEG signal entropy or vocal prosody patterns in autism spectrum disorder. Going forward, robust frameworks are needed to standardize the extraction and interpretation of digital biomarkers across populations and devices. Integration with XAI will not only enhance model transparency but also facilitate regulatory approval and cross-institutional adoption.

6.2. Federated Explainable Learning for Multicenter Studies

One of the central challenges in pediatric neurodevelopmental research is the scarcity of large, diverse datasets due to privacy constraints and ethical regulations. Federated learning (FL), which enables decentralized model training across multiple institutions without direct data sharing, has emerged as a solution. However, integrating explainability into FL frameworks—termed Federated Explainable Learning (FEL)—is a relatively new but transformative direction.

In FEL architectures, each node (institution) computes local explanations using methods like LIME or SHAP and then aggregates them along with the model updates. A notable example is the work by Wang et al. [125], who developed a federated version of SHAP to enable privacy-preserving feature attribution across five hospitals analyzing neuroimaging data for autism classification. This allowed for consistent interpretability of the model's decisions across institutions without exposing patient-level data. Nevertheless, several challenges remain. Explanation consistency across non-identically distributed datasets is a pressing issue, as variations in imaging protocols or behavioral assessments can yield diverging model rationales. Moreover, communicating the reasoning of decentralized models to clinicians demands uniform standards in how explanations are generated, validated, and visualized [126].

The integration of FEL into multicenter neurodevelopmental studies promises not only improved model generalization through broader data access but also the democratization of interpretability tools, ensuring that all participating clinicians understand and trust the AI system's rationale regardless of their institutional affiliation.

6.3. Human-in-the-Loop Models and Co-Design with Clinicians

A truly transformative vision for AI in pediatric neuroscience is one where clinicians are not passive recipients of model outputs but active participants in the development and refinement of AI tools. Human-in-the-loop (HITL) frameworks operationalize this vision by embedding expert feedback directly into the model training or post-prediction phase. This approach not only improves model relevance but also enhances interpretability, as clinicians' domain knowledge guides feature selection and outcome prioritization. For example, the co-design approach used in the CLIN-XAI project (Amann et al., [127]) involved pediatric neurologists in iterative model refinement, using their feedback to adjust both the features included in the model and the way explanations were presented. Their input led to the adoption of causal attention maps over traditional saliency plots, thereby aligning the AI's reasoning process with clinicians' diagnostic heuristics.

Moreover, HITL systems can enable active learning paradigms where models identify uncertain predictions and defer to clinicians for clarification. These interactions not only improve performance in edge cases but also generate high-value labeled data that further enhance interpretability models over time [128]. As pediatric datasets often suffer from limited volume and diversity, incorporating clinician feedback into every stage—from feature engineering to explanation generation—ensures that AI systems remain both precise and clinically aligned [129,130]. The future of explainable AI in neuroscience lies not just in smarter algorithms, but in deeper collaborations between machine learning experts and clinical practitioners.

7. Conclusion

The intersection of artificial intelligence and neurodevelopmental disorder (NDD) diagnosis represents one of the most transformative frontiers in modern neuroscience. As this review has established, the burden of NDDs—particularly autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD)—remains alarmingly high, with early detection proving pivotal for long-term functional outcomes. The emergence of AI, and more specifically explainable AI (XAI), offers a new paradigm for both identifying and understanding these conditions with greater precision, timeliness, and clinical relevance. However, while AI models have begun to outperform traditional diagnostic tools in sensitivity and speed, the opaque nature of many deep learning algorithms has posed a significant barrier to clinical adoption. The quest for transparency and accountability has thus positioned XAI as a vital component in the responsible deployment of AI in pediatric neuropsychiatry.

Throughout this review, we explored how XAI techniques—ranging from SHAP and LIME to Grad-CAM and attention mechanisms—can be applied to neuroimaging, behavioral, and EEG data to generate clinically interpretable insights. These tools have not only improved trust among clinicians but have also uncovered novel biological and behavioral patterns previously obscured by statistical noise or analytical limitations. Case studies have demonstrated how XAI methods can elucidate the neurofunctional correlates of ASD using fMRI, disentangle ADHD symptom clusters through neuropsychological features, and enhance the interpretability of child behavior analytics via decision trees and attention-based models. These examples underscore the growing maturity and impact of XAI in clinical contexts, where transparency, reproducibility, and human interpretability are non-negotiable.

Nonetheless, significant limitations persist. This review has outlined how inconsistencies in explanation reliability, trade-offs between model accuracy and interpretability, and the prevalence of dataset biases—particularly in small, heterogeneous pediatric cohorts—continue to challenge the robustness of current XAI models. Furthermore, ethical questions surrounding data privacy, algorithmic accountability, and clinician oversight remain unresolved. Future directions must therefore prioritize integrated solutions, including the development of federated explainable learning systems, co-designed human-in-the-loop models, and synergy with digital biomarkers to enable continuous, real-time, and ethically grounded monitoring of neurodevelopment. Only by navigating these challenges with interdisciplinary rigor can the neuroscience community fully realize the promise of XAI: a future where powerful AI tools not only diagnose, but also explain, justify, and humanize clinical decision-making for some of the most vulnerable populations.

Compliance with ethical standards

Acknowledgments

The authors wish to acknowledge the collaborative effort of all contributing scholars and colleagues who jointly authored and edited this review paper. This work was conducted entirely through the intellectual and academic contributions of the authoring team, without external funding or assistance from any individual, institution, or organization.

Disclosure of conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Visser, J. C., Rommelse, N. N., Greven, C. U., & Buitelaar, J. K. (2016). Autism spectrum disorder and attention-deficit/hyperactivity disorder in early childhood: A review of unique and shared characteristics and developmental antecedents. *Neuroscience & Biobehavioral Reviews*, *65*, 229-263.

[2] Zhou, J., Park, S., Dong, S., Tang, X., & Wei, X. (2025). Artificial intelligence-driven transformative applications in disease diagnosis technology. *Medical Review*, (0).

[3] Onciul, R., Tataru, C. I., Dumitru, A. V., Crivoi, C., Serban, M., Covache-Busuioc, R. A., ... & Toader, C. (2025). Artificial intelligence and neuroscience: transformative synergies in brain research and clinical applications. *Journal of Clinical Medicine*, *14*(2), 550.

[4] Xu, H., & Shuttleworth, K. M. J. (2024). Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm". *Intelligent Medicine*, *4*(1), 52-57.

[5] Band, S. S., Yarahmadi, A., Hsu, C. C., Biyari, M., Sookhak, M., Ameri, R., ... & Liang, H. W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked, 40, 101286.

[6] World Health Organization, & UNICEF. (2023). Global report on children with developmental disabilities: From the margins to the mainstream. World Health Organization. https://www.unicef.org/media/145016/file/Global-report-on-children-with-developmental-disabilities-2023.pdf

[7] Scherzer, A. L., Chhagan, M., Kauchali, S., & Susser, E. (2012). Global perspective on early diagnosis and intervention for children with developmental delays and disabilities. Developmental Medicine & Child Neurology, 54(12), 1079-1084.

[8] Stewart, L. A., & Lee, L. C. (2017). Screening for autism spectrum disorder in low-and middle-income countries: A systematic review. Autism, 21(5), 527-539.

[9] Hus, Y., & Segal, O. (2021). Challenges surrounding the diagnosis of autism in children. Neuropsychiatric disease and treatment, 3509-3529.

[10] Toki, E. I., Tsoulos, I. G., Santamato, V., & Pange, J. (2024). Machine learning for predicting neurodevelopmental disorders in children. *Applied Sciences*, *14*(2), 837.

[11] Zhao, S., Li, W., Wang, X., Foglia, S., Tan, H., Zhang, B., ... & Gao, Z. (2024). A Systematic Review of Machine Learning Methods for Multimodal EEG Data in Clinical Application. *arXiv preprint arXiv:2501.08585*.

[12] Feng, M., & Xu, J. (2023). Detection of ASD children through deep-learning application of fMRI. *Children*, *10*(10), 1654.

[13] Chakladar, D. D., Shankar, A., Liwicki, F., Barma, S., & Saini, R. (2025). Attention Dynamics: Estimating Attention Levels of ADHD using Swin Transformer. In *International Conference on Pattern Recognition* (pp. 270-283). Springer, Cham.

[14] De Belen, R. A. J., Bednarz, T., Sowmya, A., & Del Favero, D. (2020). Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational psychiatry*, *10*(1), 333.

[15] Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion, 77, 29-52.

[16] U.S. Food and Drug Administration. (2021). Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles. Retrieved from https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles

[17] European Medicines Agency. (2025). Artificial intelligence. Retrieved from https://www.ema.europa.eu/en/about-us/how-we-work/data-regulation-big-data-other-sources/artificial-intelligence

[18] Pham, T. (2025). Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. Royal Society Open Science, 12(5), 241873.

[19] Taiyeb Khosroshahi, M., Morsali, S., Gharakhanlou, S., Motamedi, A., Hassanbaghlou, S., Vahedi, H., ... & Jafarizadeh, A. (2025). Explainable artificial intelligence in neuroimaging of Alzheimer's disease. Diagnostics, 15(5), 612.

[20] Bhati, D., Neha, F., & Amiruzzaman, M. (2024). A survey on explainable artificial intelligence (xai) techniques for visualizing deep learning models in medical imaging. Journal of Imaging, 10(10), 239.

[21] Zhang, S. (2025). AI-assisted early screening, diagnosis, and intervention for autism in young children. Frontiers in Psychiatry, 16, 1513809.

[22] Wolff, N., Kohls, G., Mack, J. T., Vahid, A., Elster, E. M., Stroth, S., ... & Roessner, V. (2022). A data driven machine learning approach to differentiate between autism spectrum disorder and attention-deficit/hyperactivity disorder based on the best-practice diagnostic instruments for autism. *Scientific reports*, *12*(1), 18744.

[23] Huang, W., & Shu, N. (2025). AI-powered integration of multimodal imaging in precision medicine for neuropsychiatric disorders. *Cell Reports Medicine*, *6*(5).

[24] Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, *17*, 16-23.

[25] Sherkatghanad, Z., Akhondzadeh, M., Salari, S., Zomorodi-Moghadam, M., Abdar, M., Acharya, U. R., ... & Salari, V. (2020). Automated detection of autism spectrum disorder using a convolutional neural network. *Frontiers in neuroscience*, *13*, 1325.

[26] Khosla, M., Jamison, K., Ngo, G. H., Kuceyeski, A., & Sabuncu, M. R. (2019). Machine learning in resting-state fMRI analysis. *Magnetic resonance imaging*, *64*, 101-121.

[27] Khosla, M., Jamison, K., Kuceyeski, A., & Sabuncu, M. R. (2019). Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. *NeuroImage*, *199*, 651-662.

[28] Ahmadi, M., Kazemi, K., Kuc, K., Cybulska-Klosowicz, A., Helfroush, M. S., & Aarabi, A. (2021). Resting state dynamic functional connectivity in children with attention deficit/hyperactivity disorder. *Journal of Neural Engineering*, *18*(4), 0460d1.

[29] Zhao, Y., Yang, L., Gong, G., Cao, Q., & Liu, J. (2022). Identify aberrant white matter microstructure in ASD, ADHD and other neurodevelopmental disorders: A meta-analysis of diffusion tensor imaging studies. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *113*, 110477.

[30] Zhang, F., Savadjiev, P., Cai, W., Song, Y., Rathi, Y., Tunç, B., ... & O'Donnell, L. J. (2018). Whole brain white matter connectivity analysis using machine learning: an application to autism. *Neuroimage*, *172*, 826-837.

[31] Lou, C., Duan, X., Altarelli, I., Sweeney, J. A., Ramus, F., & Zhao, J. (2019). White matter network connectivity deficits in developmental dyslexia. *Human Brain Mapping*, *40*(2), 505-516.

[32] Akhavan Aghdam, M., Sharifi, A., & Pedram, M. M. (2018). 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.

[33] Themistocleous, C. K., Andreou, M., & Peristeri, E. (2024). Autism detection in children: Integrating machine learning and natural language processing in narrative analysis. *Behavioral Sciences*, *14*(6), 459.

[34] Syriopoulou-Delli, C. K. (2025). Advances in Autism Spectrum Disorder (ASD) diagnostics: From theoretical frameworks to ai-driven innovations. *Electronics*, *14*(5), 951.

[35] Alvari, G., Coviello, L., & Furlanello, C. (2021). EYE-C: eye-contact robust detection and analysis during unconstrained child-therapist interactions in the clinical setting of autism spectrum disorders. *Brain Sciences*, *11*(12), 1555.

[36] Perochon, S., Di Martino, J. M., Carpenter, K. L., Compton, S., Davis, N., Eichner, B., ... & Dawson, G. (2023). Early detection of autism using digital behavioral phenotyping. *Nature Medicine*, *29*(10), 2489-2497.

[37] Alvari, G., Furlanello, C., & Venuti, P. (2021). Is smiling the key? Machine learning analytics detect subtle patterns in micro-expressions of infants with asd. *Journal of clinical medicine*, *10*(8), 1776.

[38] Colizzi, M., Ciceri, M. L., Di Gennaro, G., Morari, B., Inglese, A., Gandolfi, M., ... & Zoccante, L. (2020). Investigating gait, movement, and coordination in children with neurodevelopmental disorders: is there a role for motor abnormalities in atypical neurodevelopment?. *Brain Sciences*, *10*(9), 601.

[39] McCay, K. D., Ho, E. S., Shum, H. P., Fehringer, G., Marcroft, C., & Embleton, N. D. (2020). Abnormal infant movements classification with deep learning on pose-based features. *IEEE Access*, *8*, 51582-51592.

[40] Leo, M., Bernava, G. M., Carcagnì, P., & Distante, C. (2022). Video-based automatic baby motion analysis for early neurological disorder diagnosis: state of the art and future directions. *Sensors*, *22*(3), 866.

[41] Crippa, A., Salvatore, C., Perego, P., Forti, S., Nobile, M., Molteni, M., & Castiglioni, I. (2015). Use of machine learning to identify children with autism and their motor abnormalities. *Journal of autism and developmental disorders*, *45*, 2146-2156.

[42] Bone, D., Lee, C. C., Black, M. P., Williams, M. E., Lee, S., Levitt, P., & Narayanan, S. (2014). The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *Journal of Speech, Language, and Hearing Research*, *57*(4), 1162-1177.

[43] Liu, D., Liu, Z., Yang, Q., Huang, Y., & Prud'hommeaux, E. (2022, October). Evaluating the performance of transformer-based language models for neuroatypical language. In *Proceedings of COLING. International Conference on Computational Linguistics* (Vol. 2022, p. 3412).

[44] Ghaleb, E., Burenko, I., Rasenberg, M., Pouw, W., Toni, I., Uhrig, P., ... & Fernández, R. (2024). Leveraging Speech for Gesture Detection in Multimodal Communication. *arXiv preprint arXiv:2404.14952*.

[45] Jagnade, G., Sable, S., & Ikar, M. (2023, July). Advancing Multimodal Fusion in Human-Computer Interaction: Integrating Eye Tracking, Lips Detection, Speech Recognition, and Voice Synthesis for Intelligent Cursor Control and Auditory Feedback. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE.

[46] Chen, J., Chen, C., Xu, R., & Liu, L. (2024). Autism identification based on the intelligent analysis of facial behaviors: an approach combining coarse-and fine-grained analysis. *Children*, *11*(11), 1306.

[47] Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, *39*(3), 42-62.

[48] Mendes, J. P., Moura, I. R., Van de Ven, P., Viana, D., Silva, F. J., Coutinho, L. R., ... & Teles, A. S. (2022). Sensing apps and public data sets for digital phenotyping of mental health: systematic review. *Journal of medical Internet research*, *24*(2), e28735.

[49] Edwards, Q., Idoko, B., Idoko, J. E., Ejembi, E. V., & Onuh, E. P. (2024). Remote monitoring of social behavior in children with autism: The role of digital phenotyping in public programs.

[50] Bruni, O., Breda, M., Mammarella, V., Mogavero, M. P., & Ferri, R. (2025). Sleep and circadian disturbances in children with neurodevelopmental disorders. *Nature Reviews Neurology*, 1-18.

[51] Leifler, E. (2022). *Educational inclusion for students with neurodevelopmental conditions*. Karolinska Institutet (Sweden).

[52] Krichen, M. (2021). Anomalies detection through smartphone sensors: A review. *IEEE Sensors Journal*, *21*(6), 7207-7217.

[53] Reddy, K., Taksande, A., Kurian, B., & REDDY, K. (2024). Harnessing the Power of Mobile Phone Technology: Screening and Identifying Autism Spectrum Disorder With Smartphone Apps. *Cureus*, *16*(2).

[54] Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research*, *17*(7), e4273.

[55] Boukhechba, M., Chow, P., Fua, K., Teachman, B. A., & Barnes, L. E. (2018). Predicting social anxiety from global positioning system traces of college students: feasibility study. *JMIR mental health*, *5*(3), e10101.

[56] Abbas, H., Garberson, F., Liu-Mayo, S., Glover, E., & Wall, D. P. (2020). Multi-modular AI approach to streamline autism diagnosis in young children. *Scientific reports*, *10*(1), 5014.

[57] Rykov, Y., Thach, T. Q., Bojic, I., Christopoulos, G., & Car, J. (2021). Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. *JMIR mHealth and uHealth*, *9*(10), e24872.

[58] Presby, D. M., Jasinski, S. R., & Capodilupo, E. R. (2023). Wearable derived cardiovascular responses to stressors in free-living conditions. *PLoS One*, *18*(6), e0285332.

[59] Empatica. (2021, July 20). E4 wristband | Real-time physiological signals | Wearable PPG, EDA, temperature, motion sensors. https://www.empatica.com/research/e4/

[60] Kim, W. P., Kim, H. J., Pack, S. P., Lim, J. H., Cho, C. H., & Lee, H. J. (2023). Machine learning–based prediction of attention-deficit/hyperactivity disorder and sleep problems with wearable data in children. *JAMA network open*, *6*(3), e233502-e233502.

[61] Aung, M. S. H., Alquaddoomi, F., Hsieh, C. K., Rabbi, M., Yang, L., Pollak, J. P., ... & Choudhury, T. (2016). Leveraging multi-modal sensing for mobile health: a case review in chronic pain. *IEEE journal of selected topics in signal processing*, *10*(5), 962-974.

[62] Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., ... & Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, *107*(30), 13354-13359.

[63] Onnela, J. P. (2021). Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology*, *46*(1), 45-54.

[64] Bufano, P., Laurino, M., Said, S., Tognetti, A., & Menicucci, D. (2023). Digital phenotyping for monitoring mental disorders: systematic review. *Journal of medical Internet research*, *25*, e46778.

[65] Lindhiem, O., Goel, M., Shaaban, S., Mak, K. J., Chikersal, P., Feldman, J., & Harris, J. L. (2022). Objective measurement of hyperactivity using mobile sensing and machine learning: pilot study. *JMIR Formative Research*, *6*(4), e35803.

[66] Ernst, H., Scherpf, M., Pannasch, S., Helmert, J. R., Malberg, H., & Schmidt, M. (2023). Assessment of the human response to acute mental stress–An overview and a multimodal study. *PloS One*, *18*(11), e0294069.

[67] Bruni, O., Angriman, M., Calisti, F., Comandini, A., Esposito, G., Cortese, S., & Ferri, R. (2018). Practitioner review: treatment of chronic insomnia in children and adolescents with neurodevelopmental disabilities. *Journal of Child Psychology and Psychiatry*, *59*(5), 489-508.

[68] Abgrall, G., Holder, A. L., Chelly Dagdia, Z., Zeitouni, K., & Monnet, X. (2024). Should AI models be explainable to clinicians?. *Critical Care*, *28*(1), 301.

[69] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.

[70] Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89-106.

[71] Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 1-66.

[72] Chander, B., John, C., Warrier, L., & Gopalakrishnan, K. (2025). Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. *ACM Computing Surveys*, *57*(6), 1-49.

[73] Temme, M. (2017). Algorithms and transparency in view of the new general data protection regulation. *Eur. Data Prot. L. Rev.*, *3*, 473.

[74] Uddin, M., Wang, Y., & Woodbury-Smith, M. (2019). Artificial intelligence for precision medicine in neurodevelopmental disorders. *NPJ digital medicine*, *2*(1), 112.

[75] van Mourik, F., Jutte, A., Berendse, S. E., Bukhsh, F. A., & Ahmed, F. (2024). Tertiary review on explainable artificial intelligence: where do we stand?. *Machine Learning and Knowledge Extraction*, *6*(3), 1997-2017.

[76] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

[77] Retzlaff, C. O., Angerschmid, A., Saranti, A., Schneeberger, D., Roettger, R., Mueller, H., & Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cognitive Systems Research*, *86*, 101243.

[78] Mitros, J., & Mac Namee, B. (2019). A Categorisation of post-hoc explanations for predictive models. *arXiv preprint arXiv:1904.02495*.

[79] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57.

[80] Scheinost, D., Pollatou, A., Dufford, A. J., Jiang, R., Farruggia, M. C., Rosenblatt, M., ... & Westwater, M. L. (2023). Machine learning and prediction in fetal, infant, and toddler neuroimaging: a review and primer. *Biological psychiatry*, *93*(10), 893-904.

[81] Jacob, S., Wolff, J. J., Steinbach, M. S., Doyle, C. B., Kumar, V., & Elison, J. T. (2019). Neurodevelopmental heterogeneity and computational approaches for understanding autism. *Translational psychiatry*, *9*(1), 63.

[82] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

[83] Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

[84] Hussain, T., & Shouno, H. (2023). Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping. *Information*, *14*(12), 642.

[85] Yuen, K. K. F. (2024). A Tutorial on Explainable Image Classification for Dementia Stages Using Convolutional Neural Network and Gradient-weighted Class Activation Mapping. *arXiv preprint arXiv:2408.10572*.

[86] Fantozzi, P., & Naldi, M. (2024). The explainability of transformers: Current status and directions. *Computers*, *13*(4), 92.

[87] Dias, Raquel, and Ali Torkamani. "Artificial intelligence in clinical and genomic diagnostics." *Genome medicine* 11, no. 1 (2019): 70.

[88] Iyama-Kurtycz, T. (2020). *Diagnosing and caring for the child with autism spectrum disorder*. Springer International Publishing.

[89] Cortese, S., Bellato, A., Gabellone, A., Marzulli, L., Matera, E., Parlatini, V., ... & Margari, L. (2025). Latest clinical frontiers related to autism diagnostic strategies. *Cell Reports Medicine*.

[90] Ali, H. (2023). Artificial intelligence in multi-omics data integration: Advancing precision medicine, biomarker discovery and genomic-driven disease interventions. *Int J Sci Res Arch*, *8*(1), 1012-30.

[91] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312.

[92] Ahmed, I. A., Senan, E. M., Rassem, T. H., Ali, M. A., Shatnawi, H. S. A., Alwazer, S. M., & Alshahrani, M. (2022). Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques. *Electronics*, *11*(4), 530.

[93] Supekar, K., Ryali, S., Yuan, R., Kumar, D., de Los Angeles, C., & Menon, V. (2022). Robust, generalizable, and interpretable artificial intelligence–derived brain fingerprints of autism and social communication symptom severity. *Biological Psychiatry*, *92*(8), 643-653.

[94] Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q., Nguyen, C. D., ... & Rajpurkar, P. (2022). Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence*, *4*(10), 867-878.

[95] Feng, M., & Xu, J. (2023). Detection of ASD children through deep-learning application of fMRI. *Children*, *10*(10), 1654.

[96] Giansanti, D. (2023). An Umbrella Review of the Fusion of fMRI and AI in Autism. *Diagnostics*, *13*(23), 3552.

[97] Lin, Q. H., Niu, Y. W., Sui, J., Zhao, W. D., Zhuo, C., & Calhoun, V. D. (2022). SSPNet: An interpretable 3D-CNN for classification of schizophrenia using phase maps of resting-state complex-valued fMRI data. *Medical Image Analysis*, *79*, 102430.

[98] Rogala, J., Żygierewicz, J., Malinowska, U., Cygan, H., Stawicka, E., Kobus, A., & Vanrumste, B. (2023). Enhancing autism spectrum disorder classification in children through the integration of traditional statistics and classical machine learning techniques in EEG analysis. *Scientific Reports*, *13*(1), 21748.

[99] Musullulu, H. (2025). Evaluating attention deficit and hyperactivity disorder (ADHD): a review of current methods and issues. *Frontiers in Psychology*, *16*, 1466088.

[100] Arnett, A. B., & Flaherty, B. P. (2022). A framework for characterizing heterogeneity in neurodevelopmental data using latent profile analysis in a sample of children with ADHD. *Journal of Neurodevelopmental Disorders*, *14*(1), 45.

[101] Cao, M., Martin, E., & Li, X. (2023). Machine learning in attention-deficit/hyperactivity disorder: new approaches toward understanding the neural mechanisms. *Translational Psychiatry*, *13*(1), 236.

[102] Agoalikum, E., Klugah-Brown, B., Wu, H., Hu, P., Jing, J., & Biswal, B. (2023). Structural differences among children, adolescents, and adults with attention-deficit/hyperactivity disorder and abnormal Granger causality of the right pallidum and whole-brain. *Frontiers in human neuroscience*, *17*, 1076873.

[103] Marschik, P. B., Pokorny, F. B., Peharz, R., Zhang, D., O'Muircheartaigh, J., Roeyers, H., ... & BEE-PRI Study Group. (2017). A novel way to measure and predict development: A heuristic approach to facilitate the early detection of neurodevelopmental disorders. *Current neurology and neuroscience reports*, *17*, 1-15.

[104] Jacobson, N. C., & Bhattacharya, S. (2022). Digital biomarkers of anxiety disorder symptom changes: Personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. *Behaviour Research and Therapy*, *149*, 104013.

[105] Thapar, A., Cooper, M., & Rutter, M. (2017). Neurodevelopmental disorders. *The Lancet Psychiatry*, *4*(4), 339-346.

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

[107] Rane, N., Choudhary, S., & Rane, J. (2023). Explainable artificial intelligence (XAI) in healthcare: Interpretable models for clinical decision support. *Available at SSRN 4637897*.

[108] Neugnot-Cerioli, M., & Laurenty, O. M. (2024). The Future of Child Development in the AI Era. Cross-Disciplinary Perspectives Between AI and Child Development Experts. *arXiv preprint arXiv:2405.19275*.

[109] Garg, A., Singh, A. K., & Kumar, A. (2024). Mental disorders management using explainable artificial intelligence (XAI). In *Explainable Artificial Intelligence for Biomedical and Healthcare Applications* (pp. 113-138). CRC Press.

[110] Bhattacharya, A. (2022). *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd.

[111] Triana, A. M., Glerean, E., Saramäki, J., & Korhonen, O. (2020). Effects of spatial smoothing on group-level differences in functional brain networks. *Network Neuroscience*, *4*(3), 556-574.

[112] Benishek, L. E., Kachalia, A., & Biddison, L. D. (2023). Improving clinician well-being and patient safety through human-centered design. *JAMA*, *329*(14), 1149-1150.

[113] Mohammadi, H., & Karwowski, W. (2024). Graph Neural Networks in Brain Connectivity Studies: Methods, Challenges, and Future Directions. *Brain Sciences*, *15*(1), 17.

[114] Moss, L., Corsar, D., Shaw, M., Piper, I., & Hawthorne, C. (2022). Demystifying the black box: the importance of interpretability of predictive models in neurocritical care. *Neurocritical care*, *37*(Suppl 2), 185-191.

[115] Li, L., Wen, G., Cao, P., Liu, X., R. Zaiane, O., & Yang, J. (2023). Exploring interpretable graph convolutional networks for autism spectrum disorder diagnosis. *International Journal of Computer Assisted Radiology and Surgery*, *18*(4), 663-673.

[116] Mirakhori, F., & Niazi, S. K. (2025). Harnessing the AI/ML in Drug and Biological Products Discovery and Development: The Regulatory Perspective. *Pharmaceuticals*, *18*(1), 47.

[117] Horga, G., Kaur, T., & Peterson, B. S. (2014). Annual research review: Current limitations and future directions in MRI studies of child-and adult-onset developmental psychopathologies. *Journal of Child Psychology and Psychiatry*, *55*(6), 659-680.

[118] Sterling, E., Pearl, H., Liu, Z., Allen, J. W., & Fleischer, C. C. (2022). Demographic reporting across a decade of neuroimaging: a systematic review. *Brain Imaging and Behavior*, *16*(6), 2785-2796.

[119] Lashgari, E., Liang, D., & Maoz, U. (2020). Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, *346*, 108885.

[120] Gupta, S., Kumar, S., Chang, K., Lu, C., Singh, P., & Kalpathy-Cramer, J. (2023). Collaborative privacy-preserving approaches for distributed deep learning using multi-institutional data. *RadioGraphics*, *43*(4), e220107.

[121] Patel, A. U., Gu, Q., Esper, R., Maeser, D., & Maeser, N. (2024). The crucial role of interdisciplinary conferences in advancing explainable AI in healthcare. *BioMedInformatics*, *4*(2), 1363-1383.

[122] Nasarian, E., Alizadehsani, R., Acharya, U. R., & Tsui, K. L. (2024). Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. *Information Fusion*, 102412.

[123] Pellano, K. N., Strümke, I., Groos, D., Adde, L., & Ihlen, E. A. F. (2025). Evaluating explainable ai methods in deep learning models for early detection of cerebral palsy. *IEEE Access*.

[124] Khare, S. K., & Acharya, U. R. (2023). An explainable and interpretable model for attention deficit hyperactivity disorder in children using EEG signals. *Computers in biology and medicine*, *155*, 106676.

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

[126] Li, H., Xu, J., Gan, K., Wang, F., & Zang, C. (2025). Federated Causal Inference in Healthcare: Methods, Challenges, and Applications. *arXiv preprint arXiv:2505.02238*.

[127] Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., ... & Z-Inspection Initiative. (2022). To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, *1*(2), e0000016.

[128] Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., ... & Dou, D. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, *64*(12), 3197-3234.

[129] Lage, I., & Doshi-Velez, F. (2020). Learning interpretable concept-based models with human feedback. *arXiv preprint arXiv:2012.02898*.

[130] Ganatra, H. A. (2025). Machine Learning in Pediatric Healthcare: Current Trends, Challenges, and Future Directions. *Journal of Clinical Medicine*, *14*(3), 807.