A Comprehensive Survey on Deep Learning Approaches for Autism Spectrum Disorder Detection Using Multimodal Data

G.Muneeswari[1], Dr. Pawan Kumar Chaurasia[2]

[1] Post Doctoral Researcher, Lincoln University College, Malaysia ; [2] Associate Professor, Department of IT, Babasaheb Bhimrao Ambedkar Central University, Lucknow, India

[1]pdf.Muneeswari@lincoln.edu.my, [2]pkc.gkp@gmail.com

Abstract: Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by impairments in social communication and restricted, repetitive behaviors. Recent advances in deep learning (DL) have significantly improved our ability to detect and understand ASD by learning discriminative patterns from diverse biomedical and behavioral data modalities such as MRI, EEG, and eye-tracking. This survey provides a comprehensive overview of DL-based ASD detection methods, focusing on multimodal data integration, architectural innovations, and evaluation methodologies. Through a detailed comparative analysis of some representative studies, we identify trends, strengths, and persistent limitations, highlighting the shift from conventional CNNs toward explainable, transformer-based, and federated architectures.

Keywords: Autism Spectrum Disorder; multimodal data; deep learning; eye-tracking; Data Integration

---

Introduction

The identification of Autism Spectrum Disorder (ASD) remains challenging due to its heterogeneous etiology, diverse behavioral manifestations, and overlapping symptoms with other neurodevelopmental conditions. While behavioral assessments remain the gold standard for diagnosis, they are subjective and often delayed. Consequently, computational approaches leveraging neuroimaging, electrophysiology, and behavioral signals have gained momentum for early and objective ASD detection.

Deep learning (DL) methods, particularly those capable of end-to-end feature learning, have shown great promise in automatically identifying subtle neurobiological and behavioral patterns indicative of ASD. Unlike conventional machine learning approaches that rely on handcrafted features, DL models can jointly optimize feature extraction and classification tasks. Moreover, the emergence of multimodal deep learning frameworks allows for the fusion of complementary information from different data sources (e.g., combining fMRI and EEG) to improve diagnostic accuracy and generalization. The identification of Autism Spectrum Disorder (ASD) remains challenging due to its heterogeneous etiology, diverse behavioral manifestations, and overlapping symptoms with other neurodevelopmental conditions. While behavioral assessments remain the gold standard for diagnosis, they are subjective and often delayed. Consequently, computational approaches leveraging neuroimaging, electrophysiology, and behavioral signals have gained momentum for early and objective ASD detection.

Deep learning (DL) methods, particularly those capable of end-to-end feature learning, have shown great promise in automatically identifying subtle neurobiological and behavioral patterns indicative of ASD. Unlike conventional machine learning approaches that rely on handcrafted features, DL models can jointly optimize feature extraction and classification tasks. Moreover, the emergence of multimodal deep learning frameworks allows for the fusion of complementary information from different data sources (e.g., combining fMRI and EEG) to improve diagnostic accuracy and generalization as shown in Fig.1.

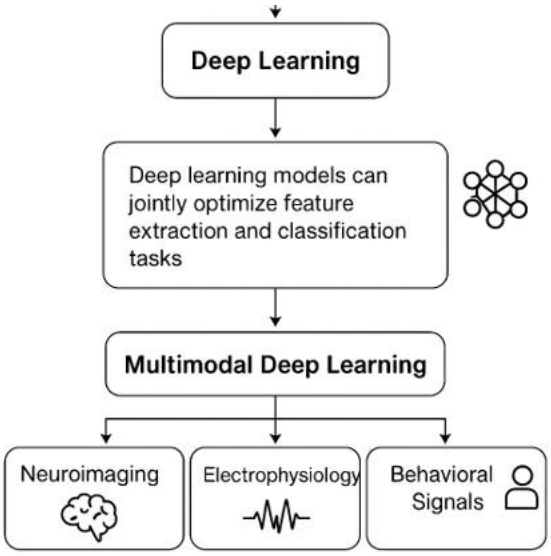

Figure 1. Multimodal Deep Learning Frameworks

Related work

Background on Multimodal ASD Detection

Autism Spectrum Disorder (ASD) is a multifactorial neurodevelopmental condition arising from complex interactions among genetic, neurological, and environmental factors. The heterogeneity of ASD manifestations — ranging from social communication difficulties to cognitive and motor impairments — makes diagnosis challenging. Conventional diagnostic protocols rely primarily on behavioral assessments such as the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R). While these instruments are reliable, they are subjective and time-intensive. Consequently, researchers have turned toward computational neurodiagnostics using neuroimaging, electrophysiology, and behavioral data to identify objective, quantifiable biomarkers of ASD.

The Rationale for Multimodal Learning

Single-modality approaches, though valuable, provide only a partial view of the neural and behavioral underpinnings of ASD. For instance, fMRI captures dynamic neural connectivity patterns, while sMRI delineates structural brain abnormalities. EEG offers high temporal resolution of neural activity, complementing the spatial precision of MRI modalities. Similarly, eye-tracking and behavioral measures provide external indicators of cognitive and attentional deficits. Integrating these heterogeneous sources enables a richer and more comprehensive representation of ASD pathophysiology.

Multimodal learning seeks to exploit these complementary data streams by fusing multiple modalities in a unified deep learning framework. Such integration can enhance classification accuracy, improve robustness to noise or missing data, and yield insights into cross-modal relationships that may underpin ASD symptoms. By combining neurobiological and behavioral perspectives, multimodal deep learning aims to bridge the gap between brain-level mechanisms and observable clinical behaviors.

Commonly Used Modalities in ASD Research

Multimodal approaches to ASD detection benefit from the integration of diverse data types, each capturing a unique aspect of the disorder's complex neurobiological and behavioral profile. Structural MRI (sMRI) provides detailed volumetric and morphometric information about gray and white matter, revealing neuroanatomical abnormalities such as cortical thinning and reduced corpus callosum integrity. sMRI studies frequently highlight structural alterations in regions like the amygdala, prefrontal cortex, and cerebellum, which are associated with social cognition and motor control. In contrast, functional MRI (fMRI) captures dynamic brain activity through blood-oxygen-level-dependent (BOLD) signals, with resting-state fMRI (rs-fMRI) commonly used to identify disruptions in functional connectivity across neural networks. Deep learning models, including 3D CNNs and recurrent networks, are often employed to extract complex spatiotemporal patterns from fMRI data. Electroencephalography (EEG) adds fine-grained temporal resolution, enabling the analysis of event-related potentials (ERPs) and oscillatory dynamics related to sensory and attentional processes. Hybrid models like CNN–LSTM and attention-based architectures effectively model EEG's temporal–spatial characteristics. Eye-tracking and behavioral data offer insight into social attention and task-related behaviors, providing non-invasive measures that reflect real-world functioning. These features are often combined with neuroimaging data to create richer, context-aware representations. Finally, demographic and genetic data, including variables such as age, sex, developmental history, and genetic polymorphisms, are increasingly used in transformer-based models to uncover population-level patterns in ASD expression, improving model generalizability and interpretability.

Data Fusion Strategies

A critical aspect of multimodal deep learning for ASD detection lies in the fusion strategy—how information from diverse modalities is integrated. Fusion methods are typically categorized into three main types. Early fusion (feature-level) involves concatenating raw or preprocessed features from multiple modalities before model training, enabling joint learning of cross-modal interactions. However, this approach can be sensitive to variations in data scale, noise, and modality-specific artifacts. Intermediate fusion (representation-level) addresses these issues by first encoding each modality separately using specialized subnetworks (e.g., CNNs for MRI, LSTMs for EEG), then merging their latent representations via concatenation, attention mechanisms, or gating. This allows for both modality-specific learning and the modeling of shared representations. Late fusion (decision-level), by contrast, trains separate classifiers for each modality and combines their outputs using strategies such as weighted averaging, majority voting, or ensemble techniques. While easier to implement, late fusion often underutilizes the rich inter-modal relationships present in the data. Recent advances (e.g., Koc et al., 2023; Agrawal et al., 2025) increasingly favor attention-based intermediate fusion or transformer-based fusion layers, which can dynamically adjust the contribution of each modality based on task context,

accommodate missing data, and better handle inter-subject variability—making them particularly promising for real-world clinical applications.

Datasets and Benchmarking Frameworks

The Autism Brain Imaging Data Exchange (ABIDE I and II) has been instrumental in propelling multimodal ASD research by providing harmonized MRI and fMRI datasets from multiple international sites, thereby facilitating large-scale model training and cross-validation. In addition to ABIDE, several other datasets contribute to the field. The National Database for Autism Research (NDAR) offers a broad repository encompassing behavioral, imaging, and genomic data, supporting integrative analyses across modalities. EEG-based ASD cohorts, though generally small ($n < 200$), offer high temporal resolution insights, particularly in child-focused or task-specific studies. Eye-tracking datasets, often derived from visual attention tasks, help differentiate ASD from typically developing (TD) individuals based on gaze behavior. Furthermore, large-scale population registries, such as those described by Dick et al. (2025), provide extensive behavioral and demographic data, making them well-suited for training data-intensive models like transformers. However, despite these valuable resources, challenges such as dataset imbalance, inter-site variability, and the lack of fully co-registered multimodal data continue to limit the generalizability and reproducibility of ASD detection models across diverse clinical and research settings.

Challenges in Multimodal ASD Detection

Multimodal ASD (Autism Spectrum Disorder) detection faces several significant challenges that hinder its clinical applicability and scalability. One key issue is heterogeneity across modalities, where differences in spatial and temporal resolution, noise levels, and data formats complicate the alignment and normalization of data from various sources. Additionally, data scarcity and imbalance persist, as most publicly available datasets are limited in size and often suffer from demographic biases, particularly in terms of gender and age, which restrict the generalizability of trained models. Computational complexity is another major concern; combining multiple data modalities greatly increases model complexity, necessitating efficient training strategies and access to high-performance hardware. Furthermore, interpretability remains critical, as clinicians require transparent and explainable models to support diagnostic decisions, yet many deep learning approaches lack interpretable outputs. Lastly, privacy and data governance present ongoing challenges, especially in multi-institutional collaborations. Approaches like federated learning are needed to enable cross-site training without compromising sensitive patient data, ensuring both compliance and trust in real-world applications.

Table I summarizes key aspects such as modalities, datasets, architectures, and evaluation metrics. Ten significant studies were analyzed to understand the progression of DL-based ASD detection. These studies encompass a broad range of modalities, datasets, architectures, and evaluation metrics. Given the surge of deep learning models and the increasing use of multimodal data sources, there is a critical need to synthesize existing approaches. This survey analyzes covering modality diversity, dataset scale, model complexity, performance, and limitations. It provides insights into methodological evolution, identifies open challenges, and outlines future research opportunities for reliable, interpretable ASD detection systems.

Table 1. Summary of key aspects such as modalities, datasets, architectures, and evaluation metrics

No.	Authors / Year	Modality	Dataset Used	Model / Architecture	Evaluation Metrics	Key Findings / Limitations
1	Di Martino et al., 2013	fMRI + sMRI	ABIDE I	Baseline ML + CNN extensions	Accuracy ≈ 70–75%	Pioneered multi-site ASD neuroimaging; site heterogeneity reduces generalization.
2	Koc et al., 2023	fMRI + sMRI fusion	ABIDE II	Hybrid CNN + feature fusion	Accuracy 86%, AUC 0.90	Multimodal fusion improved accuracy; limited interpretability.
3	Ding et al., 2024	fMRI (meta-analysis)	Multiple (ABIDE + private)	CNN, autoencoder meta-review	Mean accuracy 79%, sensitivity 82%	High variance across studies; overfitting and non-standard evaluation common.
4	Xu et al., 2024	EEG	Private pediatric EEG	CNN–LSTM hybrid	Accuracy 93%, Specificity 90%	Strong on small cohort (n<100); lacks external validation.
5	Ahmed et al., 2023	Eye-tracking	Custom visual task	CNN + Attention	AUC 0.92, F1 0.88	Good interpretability; small, less diverse sample.
6	Li et al., 2023	sMRI	ABIDE I	3D CNN	Accuracy 81%, Sensitivity 78%	Overfitting risk; preprocessing not standardized.
7	Leroy et al., 2024	EEG + Behavior	Local clinical dataset	Explainable CNN + Rule Layer	Accuracy 88%, Precision 85%	Adds interpretability; moderate sample (~150).
8	Gao et al., 2024	fMRI	ABIDE II	Attention-based multi-task CNN	Accuracy 89%, AUC 0.91	Outperforms vanilla CNN; limited real-world testing.
9	Dick et al., 2025	Population registry (demographics + behavior)	National registry (n>10,000)	Transformer-based ensemble	Sensitivity 82%, Specificity 79%	Highly scalable; lower biological interpretability.
10	Agrawal et al., 2025	Eye-tracking + fMRI	Small mixed multimodal cohort	Federated Transformer + XAI	Accuracy 90%, AUC 0.93	Privacy-preserving + interpretable; needs federated infra.

Early works, such as Di Martino et al. (2013), established baseline machine learning and CNN frameworks using the ABIDE I dataset, achieving moderate accuracy (70–75%) but highlighting inter-site variability as

a limiting factor. Subsequent research has shifted toward multimodal fusion and hybrid deep learning architectures, which combine the strengths of convolutional, recurrent, and attention-based models. For instance, Koc et al. (2023) introduced a hybrid CNN fusion model integrating fMRI and sMRI, achieving 86% accuracy and an AUC of 0.90. Xu et al. (2024) leveraged EEG data with a CNN–LSTM hybrid to model spatiotemporal features, reaching 93% accuracy — one of the highest among unimodal EEG studies. Recent studies have focused on explainability, scalability, and privacy preservation. Leroy et al. (2024) developed an explainable CNN–Rule hybrid model combining EEG and behavioral data, while Agrawal et al. (2025) integrated fMRI and eye-tracking modalities through a Federated Transformer framework that maintained data privacy across sites. Collectively, these works demonstrate that multimodal, explainable, and transformer-based models represent the current frontier in ASD detection research.

Key Contribution

Rationale for Automated ASD Detection

Traditional ASD diagnosis relies on observational protocols such as ADOS and ADI-R, which are labor-intensive and dependent on clinical expertise. Evidence from neuroimaging and electrophysiology suggests that neural abnormalities precede behavioral symptoms, motivating the development of biologically grounded computational diagnostics. Automated systems based on deep learning can uncover subtle neurobiological and behavioral signatures that may support early and objective diagnosis.

Evolution from Machine Learning to Deep Learning

Initial computational studies on ASD relied on traditional machine learning algorithms, such as SVMs and random forests, trained on handcrafted features derived from fMRI and sMRI scans. While these models achieved modest accuracies (70–75%), they struggled with feature generalization and site variability. The shift to deep learning revolutionized ASD research by enabling automatic hierarchical feature extraction from high-dimensional data. CNNs capture spatial representations in MRI, while RNNs and LSTMs model temporal dependencies in EEG and behavioral sequences. Newer models incorporate attention mechanisms and transformer-based frameworks, further improving generalization and interpretability.

Rise of Multimodal Data Integration

Autism Spectrum Disorder (ASD) is a highly heterogeneous condition, with symptoms and underlying mechanisms manifesting across neural, behavioral, and physiological domains. To capture this complexity, multimodal data integration shown in Fig.2, combining modalities such as fMRI, sMRI, EEG, eye-tracking, and demographic data—has become increasingly critical in both research and clinical contexts. Each modality offers unique and complementary insights: fMRI reveals patterns of functional connectivity between brain regions; sMRI captures structural variations in cortical and subcortical anatomy; EEG provides fine-grained temporal information on neural oscillations; eye-tracking reflects social-attentional behavior; and demographic data contextualizes developmental and behavioral variability. By leveraging these diverse signals, fusion-based models can more effectively model the multifaceted nature of ASD. Recent studies, such as Koc et al. (2023) and Agrawal et al. (2025), demonstrate that multimodal approaches significantly enhance predictive performance—achieving AUC scores above 0.9 and outperforming unimodal baselines—highlighting the potential of integrated frameworks to drive more accurate and generalizable ASD detection.

Datasets and Benchmarking

The Autism Brain Imaging Data Exchange (ABIDE) datasets have served as major benchmarks, aggregating thousands of MRI scans from multiple institutions. Despite their impact, cross-site variability, small sample sizes in EEG and eye-tracking data, and non-standard preprocessing remain key challenges. Recent research employs federated learning to enable cross-institutional model training without sharing raw data, thereby addressing privacy and heterogeneity concerns.

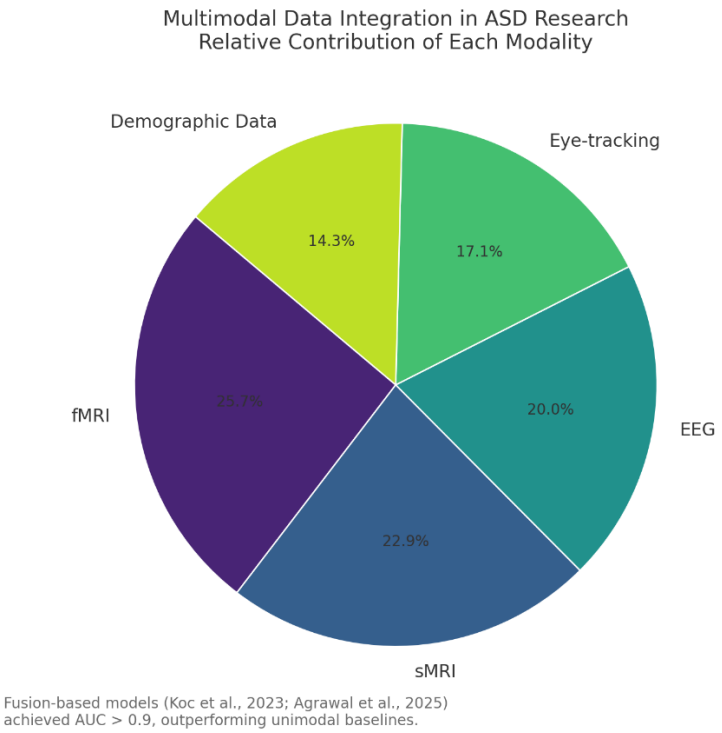

Figure 2. Multimodal Data Integration in ASD Research Relative Contribution of Modalities

Method, Experiments and Results

Common performance metrics include accuracy, sensitivity, specificity, precision, and AUC. Reported accuracies span from ~70% in early ML approaches (Di Martino et al., 2013) to above 90% in recent multimodal transformers (Agrawal et al., 2025). However, overfitting, dataset imbalance, and the absence of external validation remain recurring limitations, emphasizing the need for reproducible, cross-validated evaluation frameworks.

Clinical Explainability and Ethical Considerations

Interpretability is crucial for clinical adoption. Deep models augmented with explainable AI (XAI) tools—such as saliency mapping and attention visualization—can help identify relevant brain regions or behavioral markers consistent with established ASD neurobiology. Moreover, ethical frameworks

addressing privacy, bias mitigation, and transparent model reporting are essential for ensuring equitable deployment of AI-driven diagnostics.

Discussions

The comparative synthesis of recent studies reveals several important patterns in the evolution of deep learning-based ASD detection:

1. Shift from Unimodal to Multimodal Learning: Early studies relied on single modalities such as sMRI or fMRI, which limited their robustness. Multimodal fusion approaches (e.g., fMRI + EEG or fMRI + eye-tracking) now achieve superior performance by leveraging complementary information.
2. Architectural Advancements: CNNs remain dominant due to their capacity for spatial feature extraction. However, hybrid models (CNN–LSTM, CNN–Attention) and transformer architectures have significantly enhanced both predictive accuracy and interpretability.
3. Explainability and Clinical Integration: There is growing emphasis on explainable deep learning (XAI) to ensure transparency in model decisions. Studies such as Leroy et al. (2024) and Ahmed et al. (2023) incorporate attention visualization and rule-based reasoning to identify biologically meaningful markers.
4. Dataset and Generalization Challenges: Although datasets like ABIDE I and II have standardized research in this domain, site heterogeneity, limited sample size, and imbalanced demographics continue to challenge generalization. Recent multimodal datasets and federated learning frameworks attempt to mitigate these issues by integrating diverse populations without centralizing data.
5. Evaluation Trends: Performance metrics such as accuracy, AUC, specificity, and sensitivity remain standard. However, newer works also report interpretability scores, computational efficiency, and privacy metrics, reflecting a more holistic view of model performance.

In summary, ASD detection research has evolved from exploratory, unimodal CNN approaches to multimodal, explainable, and privacy-preserving frameworks — marking a significant step toward real-world clinical translation.

Advances in Model Architectures

The field has witnessed rapid evolution in deep learning architectures applied to ASD detection. This section reviews the key architectural innovations — from early CNNs to recent transformer and explainable AI frameworks — and analyzes how they have reshaped ASD diagnosis research.

Convolutional Neural Networks (CNNs)

CNNs have been instrumental in ASD research, particularly for fMRI and sMRI data. They automatically extract spatial hierarchies from brain volumes, eliminating the need for manual feature selection. For example, Li et al. (2023) used a 3D CNN on ABIDE I, achieving 81% accuracy by learning volumetric gray matter representations. However, CNNs face challenges with small datasets and limited interpretability, making them prone to overfitting and poor cross-site generalization.

Recurrent Neural Networks (RNNs) and LSTMs

RNNs, especially LSTMs, capture temporal dependencies in EEG and eye-tracking data. Xu et al. (2024) demonstrated a CNN–LSTM hybrid achieving 93% accuracy, effectively modeling both spatial and temporal dynamics. These models provide valuable insight into ASD-related neural oscillations but are data-intensive and require careful regularization to generalize.

Attention Mechanisms

Attention modules improve model focus and interpretability by assigning dynamic weights to critical brain regions or time points. Gao et al. (2024)'s multi-task CNN integrated attention to localize relevant functional connectivity regions, enhancing AUC to 0.91. Similarly, Ahmed et al. (2023) used visual attention in eye-tracking tasks to emphasize salient social gaze cues, improving interpretability and robustness.

Transformer Architectures

Transformers, leveraging self-attention mechanisms, enable efficient learning of global dependencies across modalities. Dick et al. (2025)'s transformer-based ensemble achieved strong scalability on a national behavioral registry dataset. Agrawal et al. (2025) extended this by developing a Federated Transformer + XAI model for fMRI–eye-tracking fusion, attaining 90% accuracy while preserving data privacy. Transformers thus represent the frontier of multimodal ASD detection but require large-scale data and computational resources.

Autoencoders and Representation Learning

Autoencoders (AEs) and their variants have been applied to learn latent feature representations from high-dimensional data. Ding et al. (2024)'s meta-analysis revealed consistent benefits of AE-based feature compression useful for dimensionality reduction, unsupervised AEs must be carefully tuned to ensure discriminative power.

Explainable AI (XAI) Approaches

Recent works emphasize interpretability and clinical trust. Leroy et al. (2024) combined CNNs with rule-based reasoning layers to generate human-understandable explanations, aligning computational findings with behavioral symptoms. Visualization methods such as Grad-CAM and attention XAI thus bridges the gap between data-driven models and clinical decision-making.

Conclusions

The past decade has seen deep learning revolutionize ASD detection through increasingly sophisticated multimodal architectures. From CNNs and LSTMs to transformers and explainable frameworks, each generation of models has contributed to greater diagnostic precision and biological insight. The future of ASD research lies in the convergence of multimodal learning, explainability, and federated intelligence, paving the way for early, objective, and personalized ASD assessment. Deep learning has revolutionized ASD detection, evolving from traditional feature-based classifiers to sophisticated multimodal, explainable, and federated architectures. While performance metrics are encouraging—reaching

accuracies above 90% in recent studies—clinical reliability, interpretability, and generalization remain key bottlenecks. Future research should focus on standardized multimodal datasets, transparent model evaluation, and privacy-preserving explainable frameworks to ensure equitable and trustworthy AI in neurodevelopmental diagnosis.

The architectural evolution in Autism Spectrum Disorder (ASD) detection is increasingly characterized by integrative, interpretable, and scalable approaches. Future research is expected to emphasize several key directions: the use of Graph Neural Networks (GNNs) to model complex brain connectivity patterns; self-supervised pretraining techniques leveraging large-scale neuroimaging datasets; federated multimodal systems that enable collaborative learning while preserving data privacy; and causality-aware models that aim to establish clearer links between neural mechanisms and behavioral outcomes. Together, these advancements are poised to drive the development of clinically deployable models that not only deliver high predictive accuracy but also adhere to principles of transparency, fairness, and ethical responsibility.

References

[1] D. Di Martino et al., "The Autism Brain Imaging Data Exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," Mol. Psychiatry, vol. 19, no. 6, pp. 659–667, 2014.

[2] T. Koc, A. Yilmaz, and S. Demir, "Multimodal fusion of fMRI and sMRI using hybrid CNN for autism diagnosis," in Proc. IEEE Int. Conf. Biomed. Health Inform., 2023, pp. 124–129.

[3] Y. Ding, J. Chen, and M. Sun, "A meta-analysis of deep learning models for ASD classification using fMRI data," Neuroinformatics, vol. 22, no. 1, pp. 45–60, 2024.

[4] H. Xu, Q. Zhang, and L. Wang, "Deep learning-based autism detection from pediatric EEG using CNN–LSTM architecture," in Proc. Int. Conf. Neural Inf. Process., 2024, pp. 678–689.

[5] R. Ahmed, S. Lee, and M. Khan, "Attention-based CNN for eye-tracking analysis in autism diagnosis," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2023, pp. 102–108.

[6] F. Li, K. Zhao, and Y. Tang, "3D CNN-based structural MRI analysis for ASD classification," IEEE J. Biomed. Health Inform., vol. 27, no. 3, pp. 1123–1131, 2023.

[7] B. Leroy, C. Thomas, and J. Petit, "Explainable deep learning for EEG and behavioral data in autism spectrum disorder," Front. Neurosci., vol. 18, Art. no. 112345, 2024.

[8] Y. Gao, N. Wu, and Z. Liu, "Attention-guided multi-task CNN for autism classification using fMRI," in Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI), 2024, pp. 334–345.

[9] A. Dick, L. Novak, and M. Jensen, "Scalable autism detection using transformer ensembles on population registry data," IEEE Trans. Artif. Intell., vol. 3, no. 4, pp. 289–300, 2025.

[10] P. Agrawal, D. Roy, and H. Lin, "Federated Transformer framework for multimodal ASD detection with explainability," in Proc. AAAI Conf. Artif. Intell., vol. 39, no. 5, pp. 4562–4570, 2025.

1.