Computers in Biology and Medicine 196 (2025) 110723

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

ASD-GraphNet: A novel graph learning approach for Autism Spectrum Disorder diagnosis using fMRI data

Mina Zeraati [a,*], Amirehsan Davoodi [b]

[a] Alzahra University, Department of Engineering, Tehran, Iran
[b] Amirkabir University of Technology, AGML Laboratory, Department of Mathematics and Computer Science, Tehran, Iran

ARTICLE INFO

Keywords:
Autism Spectrum Disorder (ASD)
fMRI
Graph machine learning
Brain connectivity networks
Brain atlases
Functional connectivity networks
ABIDE dataset

ABSTRACT

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition with heterogeneous symptomatology, making accurate diagnosis challenging. Traditional methods rely on subjective behavioral assessments, often overlooking subtle neural biomarkers. This study introduces ASD-GraphNet, a novel graph-based learning framework for diagnosing ASD using functional Magnetic Resonance Imaging (fMRI) data. Leveraging the Autism Brain Imaging Data Exchange (ABIDE) dataset, ASD-GraphNet constructs brain networks based on established atlases (Craddock 200, AAL, and Dosenbach 160) to capture intricate connectivity patterns. The framework employs systematic preprocessing, graph construction, and advanced feature extraction to derive node-level, edge-level, and graph-level metrics. Feature engineering techniques, including Mutual Information-based selection and Principal Component Analysis (PCA), are applied to enhance classification performance. ASD-GraphNet evaluates a range of classifiers, including Logistic Regression, Support Vector Machines, and ensemble methods like XGBoost and LightGBM, achieving an accuracy of 75.25% in distinguishing individuals with ASD from healthy controls. This demonstrates the framework's potential to provide objective, data-driven diagnostics based solely on resting-state fMRI data. By integrating graph-based learning with neuroimaging and addressing dataset imbalance, ASD-GraphNet offers a scalable and interpretable solution for early ASD detection, paving the way for more reliable interventions. The GitHub repository for this project is available at: https://github.com/AmirDavoodi/ASD-GraphNet.

1. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that affects communication, behavior, and social interaction. The prevalence of ASD has risen significantly in recent years, underscoring the importance of early and accurate diagnosis to ensure timely intervention and support for individuals with ASD [1]. Early diagnosis is critical, as it can significantly improve outcomes by providing access to tailored treatments and therapies during crucial developmental windows [2].

Despite advancements in ASD research, accurately diagnosing the disorder remains challenging due to the heterogeneity of symptoms and underlying neural mechanisms. Traditional diagnostic methods often rely on behavioral assessments, which can be subjective and may not capture the full spectrum of ASD characteristics. Moreover, these approaches might not be suitable for identifying subtle neural alterations that could serve as early biomarkers for ASD [3]. In recent years, machine learning and neuroimaging techniques have emerged as promising tools for ASD diagnosis. Functional Magnetic Resonance Imaging (fMRI) data, in particular, provides valuable insights into brain connectivity patterns associated with ASD. However, existing methods that utilize fMRI data for ASD diagnosis often face challenges related to high-dimensionality, variability across subjects, and the complexity of brain networks [3,4].

Several machine learning and graph-based approaches have been applied to fMRI data to identify ASD-related patterns. Support Vector Machines (SVMs), for instance, have been widely used in ASD diagnosis due to their robustness and ability to handle high-dimensional data. Studies have shown that SVMs, when combined with feature selection methods, can achieve competitive performance in classifying ASD from fMRI data [5]. Additionally, hybrid models that integrate Convolutional Neural Networks (CNNs) with SVMs have shown promising results by leveraging both deep learning's feature extraction capabilities and SVM's classification power [6]. Despite their success, these methods often face challenges such as the need for large datasets, manual feature

* Corresponding author.
E-mail addresses: minazeraatii@gmail.com (M. Zeraati), amir.davoodi@aut.ac.ir (A. Davoodi).
URLs: https://alzahra.ac.ir (M. Zeraati), https://agml.aut.ac.ir (A. Davoodi).

https://doi.org/10.1016/j.compbiomed.2025.110723
Received 14 March 2025; Received in revised form 7 June 2025; Accepted 3 July 2025
Available online 21 July 2025

extraction, and overfitting, which can limit their generalizability and applicability across different datasets [7].

To address these limitations, we propose a novel graph learning method for ASD diagnosis using fMRI data. Our approach leverages the Autism Brain Imaging Data Exchange (ABIDE) dataset [8], a large and widely-used repository of fMRI data from individuals with ASD and typically developing controls [5]. Unlike traditional methods that rely heavily on predefined features or require large sample sizes, our method constructs brain network graphs based on three well-known brain atlases: Craddock 200 (CC200) [9], Automated Anatomical Labeling (AAL), and Dosenbach 160 (DOS160) [10]. These graphs capture the complex brain connectivity patterns, with nodes representing brain regions and edges representing the functional resting state correlations between them.

To analyze these brain graphs, we extract a variety of topological features, including node-based, edge-based, and graph-level metrics [11]. These features are then subjected to advanced feature engineering techniques, such as mutual information-based feature selection and Principal Component Analysis (PCA). This two-step feature selection process enhances the discriminative power of the features while reducing dimensionality, making the classifiers more robust to noise and overfitting [7]. We then train several machine learning classifiers, including SVMs, Decision Trees (DT), Random Forests (RF), Logistic Regression (LR), Multi-layer Perceptron (MLP), XGBoost, and LightGBM, to classify ASD versus typical development.

This study introduces ASD-GraphNet, a comprehensive graph-based learning framework that addresses several critical limitations in existing ASD diagnosis methods while making distinct novel contributions to the field. Unlike previous studies that typically rely on single brain atlases, our framework systematically integrates three complementary brain atlases (CC200, AAL, and Dosenbach 160) to capture diverse aspects of brain organization. This multi-atlas approach addresses the limitation of atlas-dependent results commonly observed in neuroimaging studies, providing more robust and generalizable connectivity representations that are less susceptible to atlas-specific biases.

While existing methods often focus on simple connectivity measures or require extensive manual feature engineering, ASD-GraphNet introduces a comprehensive three-level feature extraction approach that systematically captures node-level properties of individual brain regions, edge-level characteristics of pairwise connectivity patterns, and graph-level metrics that characterize global network topology. This systematic feature engineering addresses the limitation of incomplete connectivity characterization prevalent in previous studies, ensuring that important topological information is not overlooked during the classification process.

The framework also addresses the curse of dimensionality that commonly affects graph-based ASD studies through a novel two-stage feature optimization approach. By combining mutual information-based feature selection with Principal Component Analysis, our method effectively reduces dimensionality while preserving discriminative information, leading to more robust classifiers that are less prone to overfitting. Furthermore, unlike studies that focus on single classifier types, ASD-GraphNet systematically evaluates both traditional machine learning approaches and modern ensemble methods, providing comprehensive insights into optimal classifier selection for graph-based ASD diagnosis.

Recognizing that many existing studies overlook the inherent imbalances present in ASD datasets—including site-wise variations, demographic disparities, and class distribution issues—our framework explicitly addresses these challenges through stratified sampling strategies, comprehensive bias analysis, and transparent reporting of dataset characteristics. Finally, to bridge the gap between research and clinical application, ASD-GraphNet integrates comprehensive explainable AI techniques using SHAP analysis, transforming the typically "black box" nature of machine learning models into transparent, interpretable diagnostic tools that provide both global feature importance analysis and individual-level prediction explanations grounded in established ASD neurobiology.

These contributions collectively address the key limitations of existing methods: limited atlas diversity, incomplete feature characterization, dimensionality challenges, classifier selection uncertainty, dataset bias issues, and lack of interpretability. By providing a comprehensive, interpretable, and robust framework, ASD-GraphNet bridges the gap between research and clinical application in ASD diagnosis.

To systematically evaluate the effectiveness of our ASD-GraphNet framework and address the identified limitations in existing approaches, this study investigates several key objectives. First, we examine whether the systematic integration of multiple brain atlases can improve ASD classification performance compared to single-atlas approaches, and how different atlases contribute to capturing distinct aspects of brain connectivity patterns in autism. Second, we evaluate how the implementation of a comprehensive three-level feature extraction approach enhances discriminative power for ASD diagnosis compared to conventional connectivity measures.

Additionally, this work assesses the effectiveness of combining mutual information-based feature selection with Principal Component Analysis in addressing the curse of dimensionality while preserving classification performance in graph-based ASD diagnosis. We also determine which combination of traditional machine learning classifiers and modern ensemble methods provides optimal performance for ASD classification across various atlas configurations. Furthermore, we evaluate how effectively stratified sampling strategies and bias analysis address the inherent imbalances in ASD datasets, and examine the impact of these techniques on model generalizability.

Finally, this study investigates whether SHAP-based explainable AI techniques can provide neurobiologically meaningful insights into ASD brain connectivity patterns, and examines how these findings align with established ASD neurobiology literature. Through systematic investigation of these objectives, we aim to demonstrate the comprehensive effectiveness of our proposed framework while establishing its clinical relevance and interpretability.

The remainder of this paper is organized as follows: In Section 2, we provide background information and a review of related work in ASD diagnosis using fMRI data and graph learning methods. Section 3 describes the materials and methods used in our study, including the dataset, preprocessing steps, and details of the proposed approach. Section 4 presents the experiments and results, where we address the imbalanced nature of the dataset, conduct ablation studies and synthetic analysis to understand the behavior of different components within our ASD-GraphNet framework, and evaluate our method against state-of-the-art techniques. Finally, in Section 6, we conclude the paper with a discussion of our findings, potential limitations, directions for future research, and a summary of the key contributions and impact of our work.

2. Background information and literature review

The diagnosis of Autism Spectrum Disorder (ASD) using functional magnetic resonance imaging (fMRI) data has been an active area of research in recent years. In this section, we review the current state of the field, highlighting key studies and methods that have utilized the Autism Brain Imaging Data Exchange (ABIDE) dataset, which has provided functional and structural brain imaging datasets collected from multiple brain imaging centers worldwide [9].

To provide a comprehensive overview of the current state of ASD diagnosis using neuroimaging techniques, we present a systematic analysis of key studies that have shaped this field. The following literature review encompasses various methodological approaches, from traditional machine learning techniques to modern deep learning frameworks, with particular emphasis on their contributions to advancing automated ASD diagnosis capabilities.

Table 1
Comprehensive literature review of ASD diagnosis methods using neuroimaging data.

Study	Method/Technique	Dataset/Sample Size	Features	Accuracy	Key Contribution/Limitation
[12]	Probabilistic Neural Network	ABIDE (640 subjects)	rs-fMRI connectivity	86.9%	Age-restricted analysis, limited generalizability
[13]	Machine Learning Classification	ABIDE (178 subjects)	Functional connectivity	76.67%	Age/IQ matched dataset, biomarker standards discussion
[14]	Graph Convolutional Network	ABIDE (1,035 subjects)	Imaging + phenotypic	70.4%	Population-based graph representation, limited by demographic reliance
[15]	Combined structural/functional	ABIDE (1,035 subjects)	sMRI + fMRI features	64.3%	Multi-modal approach, lower performance indicates complexity
[16]	Stacked denoising Autoencoder	ABIDE (1,035 subjects)	rs-fMRI connectivity	70%	Deep learning approach, site-specific performance issues
[17]	SVM with frequency analysis	240 subjects	Frequency-specific connectivity	79.17%	Frequency band analysis, traditional ML limitations
[18]	Deep Neural Network	ABIDE (110 subjects)	rs-fMRI with feature selection	86.36%	Novel feature selection, requires large datasets
[19]	Random SVM Cluster	ABIDE (84 subjects)	Connectivity features	96.15%	Ensemble approach, traditional ML constraints
[20]	WL-DeepGCN	ABIDE (871 subjects)	Multi-modal features	77.27%	Weight-learning GCN, demonstrates graph method potential
[21]	Multi-atlas Deep Learning	ABIDE (1,035 subjects)	Three brain atlases	78.07%	Multi-atlas approach, atlas dependency issues
[22]	Inception V3 Transfer Learning	ABIDE (138 subjects)	EPI/stat-map images	98.38%	Transfer learning for neuroimaging, CNN-based image classification
[23]	Optimized Transfer Learning	ABIDE (138 subjects)	fMRI neuroimaging	98.38%	Hyperparameter optimization, comparative transfer learning study
ASD-GraphNet (Ours)	Multi-atlas Graph Learning	ABIDE (1,035 subjects)	Node/edge/graph-level	75.25%	Comprehensive feature engineering, explainable AI integration

Table 1 summarizes the key studies in ASD diagnosis using neuroimaging data, highlighting their methodologies, datasets used, and achieved performance. This comprehensive comparison reveals the evolution of techniques from simple statistical methods to sophisticated deep learning approaches, demonstrating both the progress made and the challenges that remain in this field.

Numerous studies have leveraged the ABIDE dataset to develop novel methods for ASD diagnosis. Some studies have used subsets of this dataset based on specific demographic information for analysis [8,12,16,17,24]. For example, one proposed method [12] used a probabilistic neural network to classify resting-state fMRI (rs-fMRI) data of individuals under 20 years old. Another study [13] used two rs-fMRI datasets from ABIDE I including 118 males (59 typically developing (TD), 59 ASD) and another including 178 age and IQ-matched individuals (89 TD, 89 ASD) achieving an accuracy of 76.67%.

The evolution of image segmentation and classification techniques in ASD diagnosis has progressed from traditional region-of-interest (ROI) based approaches to sophisticated deep learning methodologies. Early studies primarily focused on predefined anatomical regions using atlas-based segmentation, which provided interpretable but limited feature representations. The transition to data-driven approaches has enabled more comprehensive analysis of brain connectivity patterns, though often at the cost of interpretability.

Recent advances in transfer learning have shown particular promise for neuroimaging applications where labeled data is often limited. Herath et al. [22] developed an ASD diagnosis support model using Inception V3 architecture applied to EPI (echo-planar imaging) and statistical map images derived from fMRI data. Their approach leveraged pre-trained convolutional neural networks to extract meaningful features from neuroimaging data, demonstrating the potential of transfer learning in medical image analysis. Building upon this work, the same research group [23] conducted a comprehensive comparative study to optimize transfer learning approaches for ASD classification using neuroimaging data. Their investigation included systematic hyperparameter optimization and comparison of different optimizers (Adam and SGD) with Inception-v3 architecture, providing valuable insights into the optimal configuration of deep learning models for neuroimaging applications.

In addition to using fMRI data, some studies have included structural and demographic information for diagnosing autism. For instance, one study Parisot et al. [14] used a graph convolutional network framework that achieved an accuracy of 70.4% by representing the population as a graph, where nodes are defined based on imaging features and phenotypic information, with edge weights describing their relationships. Another study Sen et al. [15] proposed a new algorithm that combined structural and functional features from MRI and fMRI data, achieving 64.3% accuracy. Additionally, Parikh et al. [25] tested different machine learning performances on demographic information (including age, gender, handedness, and three individual IQ measures) provided by the ABIDE dataset.

Graph-based methods have emerged as particularly promising approaches for ASD diagnosis due to their ability to naturally model the complex connectivity patterns inherent in brain networks. These methods offer several advantages including their capacity to capture both local and global network properties, their ability to incorporate multiple types of relationships simultaneously, and their potential for providing neurobiologically interpretable results. However, graph-based approaches also face challenges including computational complexity, sensitivity to network construction parameters, and the curse of dimensionality when dealing with high-resolution brain atlases.

Graph-based methods have gained attention in brain imaging data analysis, especially in autism diagnosis, due to their ability to model complex relationships and brain data structures. These methods typically use graphs to represent and analyze connections between different brain regions. Zhang et al. [26] introduces a novel graph-learning approach to create a fused cognitive network, enhancing the accuracy of discriminating cognitive states. A different study Zhang et al. [27] uses a graph convolutional network (GCN) for diagnosing Alzheimer's from 3D MRI and PET data, achieving 85% accuracy. Another research Wang et al. [20] employs WL-DeepGCN, a method that integrates weight-learning and graph convolutional networks, for diagnosing autism spectrum disorder (ASD), achieving an accuracy of 77.27%.

Multi-Atlas combination methods are specifically used in brain data analysis and medical image processing, involving the combination and registration of multiple atlases to improve analysis accuracy and identify brain features. Epalle et al. [21] presents a multi-input deep neural network model for autism classification using the ABIDE database, incorporating neuroimaging data processed with three different atlases. Their model achieves classification accuracies of 78.07% on real data and 79.13% on augmented data. Another study Yao and Li [28] proposes a multi-view data fusion technique with multi-kernel learning for analyzing MR brain images using three distinct brain atlases. Achieving about 82.5% on ADNI datasets, the method improves classification accuracy. Traditional machine learning techniques, while offering advantages in terms of interpretability and computational efficiency, face several limitations when applied to high-dimensional neuroimaging data [8,19,29]. Support Vector Machines [30], although robust to overfitting and effective with small sample sizes, struggle with feature selection in high-dimensional spaces and lack the ability to automatically learn hierarchical feature representations. Random forests [31] provide better handling of high-dimensional data and built-in feature importance measures, but may not capture complex non-linear relationships as effectively as deep learning approaches. For example, Chen et al. [17] explored the impact of different frequency bands for constructing functional brain networks using the SVM technique, achieving 79.17% accuracy with 112 autism and 128 healthy individuals. These methods often require extensive manual feature engineering and domain expertise to achieve optimal performance.

Deep learning approaches have demonstrated significant potential for ASD diagnosis by automatically learning hierarchical feature representations from raw neuroimaging data. Convolutional Neural Networks excel at capturing spatial patterns in brain images, while recurrent architectures like LSTM can model temporal dynamics in fMRI time series. However, these methods also present challenges including the requirement for large datasets, susceptibility to overfitting with limited samples, computational complexity, and reduced interpretability compared to traditional approaches. The “black box” nature of deep learning models poses particular challenges in clinical applications where understanding the basis for diagnostic decisions is crucial.

In recent years, modern machine learning approaches have gained popularity for ASD diagnosis. Deep learning (DL) methods, such as autoencoders, deep neural networks (DNN), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN), have been applied to fMRI data with encouraging results [18,32–36]. For example, a deep learning-based approach using two stacked denoising autoencoders achieved 70% accuracy in classifying 1,035 individuals (505 ASD and 530 controls) [16]. Two stacked denoising autoencoders were first pre-trained to extract lower-dimensional data. After training the autoencoders, their weights were used in a multi-layer perceptron classifier. They also performed classification for each of the 17 sites, reporting an average accuracy of 52%. The low performance at individual sites is attributed to the lack of sufficient samples for within-site training. Although these modern approaches have shown improved performance, they often require large amounts of data and may suffer from overfitting, highlighting the need for further research in this area.

The comparative analysis presented in Table 1 reveals several important trends and gaps in the current literature. Most studies achieve accuracies between 70%–80%, indicating the inherent difficulty of the ASD classification task. However, significant variation in methodologies, dataset preprocessing, and evaluation protocols makes direct comparison challenging. A critical observation is that most available studies utilize only subsets of the ABIDE I dataset rather than the complete 1,035 subjects, which reflects the substantial challenge of achieving high performance on the entire dataset. This difficulty stems from the multi-site nature of ABIDE I, where data collected from different imaging centers exhibit varying acquisition parameters, scanner specifications, and preprocessing protocols, resulting in heterogeneous fMRI data characteristics that complicate model generalization. Many studies rely on demographic or phenotypic information in addition to neuroimaging data, which may not be readily available in clinical settings. Furthermore, limited attention has been given to model interpretability and the biological significance of identified biomarkers, representing a crucial gap for clinical translation.

Despite the growing body of research on ASD diagnosis using machine learning techniques, most studies have only considered a subset of the ABIDE dataset or incorporated additional information beyond fMRI data. A limited number of studies have analyzed the entire ABIDE dataset solely using fMRI data, without incorporating demographic information or any prior assumptions about the population. Our study aims to develop a novel, fMRI-based diagnostic model that can provide a more objective and quantitative assessment of ASD. While incorporating anatomical features, demographic information, or behavioral data could potentially increase the accuracy of ASD diagnosis, our approach focuses on designing a purely quantitative model based on brain functional data. Our proposed model has the potential to contribute to the development of more accurate and reliable diagnostic tools for ASD, which can ultimately improve clinical decision-making and patient outcomes.

3. Materials and methods

In this section, we detail the materials and methods employed in developing ASD-GraphNet, a novel framework designed to diagnose Autism Spectrum Disorder (ASD) using functional Magnetic Resonance Imaging (fMRI) data. The objective of this work is to leverage graph-based learning approaches to uncover distinctive patterns in brain connectivity that differentiate ASD individuals from healthy controls. By utilizing the publicly available Autism Brain Imaging Data Exchange (ABIDE) dataset and implementing a series of preprocessing, feature extraction, and classification techniques, we aim to improve the accuracy and robustness of ASD diagnosis. The section is organized into the following subsections: (1) Dataset Description, (2) Data Preprocessing, (3) Graph Construction, (4) Feature Extraction, (5) Feature Engineering and (6) Classification Models.

3.1. Dataset description

By tracking variations in blood flow, Functional Magnetic Resonance Imaging (fMRI) is a potent technique for studying brain activity. It divides the brain into small cubic units called voxels, tracking their activity over time to create a time series for each voxel. In this study, we utilize data from the Autism Brain Imaging Data Exchange I (ABIDE-I) dataset, a large-scale collection of resting-state functional Magnetic Resonance Imaging (rs-fMRI) data [8]. The ABIDE initiative aggregates data from multiple sites to facilitate research on Autism Spectrum Disorder (ASD). The ABIDE-I dataset includes a total of 1,112 participants, consisting of 505 individuals diagnosed with Autism Spectrum Disorders (ASD) and 530 typically developing controls (TDC). These subjects were recruited from 17 different sites, leading to variability in scanning protocols and participant demographics across sites.

The dataset covers a wide range of ages and includes both male and female participants, with a higher prevalence of males, reflecting the gender disparity observed in ASD diagnoses. The average age of participants varies across sites, encompassing both children and adults. Each site employed specific scanning protocols, including variations in parameters such as repetition time (TR), echo time (TE), and the conditions under which participants were scanned, such as whether their eyes were open or closed. Table 2 presents a detailed breakdown of the dataset, showing the number of Autism Spectrum Disorders (ASD) and typically developing (TDC) subjects across the 17 sites in the ABIDE-I dataset.

Table 2
Distribution of ASD and control subjects across sites.

Site	Caltech	CMU	KKI	Leuven	MaxMun	NYU	OHSU	OLIN	PITT	SBL	SDSU	Stanford	Trinity	UCLA	UM	USM	Yale
ASD	19	14	20	29	24	75	12	19	29	15	14	19	22	54	66	46	28
TDC	18	13	28	34	28	100	14	15	27	15	22	20	25	44	74	25	28
Total	37	27	48	63	52	175	26	34	56	30	36	39	47	98	140	71	56

3.2. Data preprocessing

Preprocessing is a critical step in preparing rs-fMRI data for analysis. The preprocessing of the ABIDE-I dataset includes several standard steps designed to reduce noise and correct for artifacts, thereby ensuring data consistency across different sites. Initially, slice timing correction is applied to account for differences in acquisition times between slices, ensuring temporal alignment. This is followed by motion correction, which adjusts for any head movements during scanning. This step is particularly important given the challenges of scanning pediatric and clinical populations. Subsequently, nuisance signals, such as those from white matter, cerebrospinal fluid, and global signals, are removed to reduce confounding influences. Low-frequency drifts, which could be related to scanner instability or unrelated physiological processes, are also corrected. Finally, voxel intensity normalization is performed to standardize the signal across participants and sites.

For this study, we employ three brain parcellation atlases, including the Craddock 200 (CC200) [37], Automated Anatomical Labeling (AAL) [38], and Dosenbach 160 (DOS160) [10], which divide the brain into 200, 116, and 161 functionally homogeneous regions of interest (ROIs), respectively. These parcellations allow us to extract time series data from each region, representing the average fMRI signal across all voxels within that region. By following these preprocessing procedures, the fMRI data was guaranteed to be uniform and appropriate for graph building. The next step is to build three distinct functional connectivity networks using these time series.

Fig. 1 provides a comprehensive flowchart that visually illustrates the preprocessing steps applied to the ABIDE-I dataset. The preprocessing pipeline is systematically outlined, detailing the steps involved from initial data acquisition to the extraction of time series for different brain atlases. The pipeline begins with a thorough quality check of the raw fMRI data to identify and mitigate artifacts, such as motion and noise, which can affect the integrity of the data. Following the quality check, motion correction is applied to align the images across the time series, reducing motion-related artifacts. The pipeline then proceeds with slice timing correction, spatial normalization, and smoothing to enhance the signal-to-noise ratio. Temporal filtering is applied to remove low-frequency drifts and high-frequency noise from the time series, and global signal regression is optionally performed to regress out global signal fluctuations. Finally, the time series data is extracted based on the chosen brain atlas, including the Craddock 200 (CC200), Automated Anatomical Labeling (AAL), and Dosenbach 160 (DOS160) atlases. The preprocessing pipeline concludes with a final quality control check to ensure that the processed data meets the required standards for analysis.

By applying these preprocessing steps, we ensure that the data is suitable for subsequent graph-based analyses, enabling us to investigate the functional connectivity patterns associated with ASD using multiple parcellation schemes.

3.3. Graph construction

In the ASD-GraphNet approach, brain networks were constructed by representing each brain region as a node and the functional connectivity between regions as edges. The construction of these brain graphs involved several steps. First, nodes were defined based on the regions of interest (ROIs) specified by the chosen brain atlases, namely the Craddock 200 (CC200), Automated Anatomical Labeling (AAL), and Dosenbach 160 (DOS160) atlases. The use of multiple brain atlases allowed for the comparison of different parcellation schemes and their impact on classification performance.

To define the edges between nodes, we calculated the pairwise Pearson correlation coefficients between the time series of fMRI signals in different brain regions. Pearson's correlation is a widely used measure to estimate functional connectivity in fMRI data, as it reflects the linear relationship between the time series of two distinct brain regions. Given two time series, u and v, each of length T, Pearson's correlation (P_{uv}) was calculated as Eq. (1) where $\bar{u}$ and $\bar{v}$ represent the mean values of the time series u and v, respectively. By computing all pairwise correlations, we obtained a correlation matrix $C_{n\times n}$, where n is the number of regions or time series.

$$P_{uv} = \frac{\sum_{t=1}^{T}(u_t - \bar{u})(v_t - \bar{v})}{\sqrt{\sum_{t=1}^{T}(u_t - \bar{u})^2}\sqrt{\sum_{t=1}^{T}(v_t - \bar{v})^2}} \tag{1}$$

The selection of these three complementary brain atlases was strategically motivated by their distinct parcellation approaches and established relevance to ASD research. The CC200 atlas employs a data-driven, functionally-derived parcellation using spatially constrained spectral clustering, making it particularly suitable for capturing intrinsic connectivity patterns in resting-state fMRI analysis. Its higher resolution (200 ROIs) enables detection of subtle functional subdivisions that may be altered in ASD. The AAL atlas provides anatomically-based parcellation with well-established boundaries, facilitating comparison with existing neuroimaging literature and offering clinical interpretability through known brain structures. The Dosenbach 160 atlas, derived from meta-analyses of task-related fMRI studies, is particularly effective for identifying executive control, default mode, and salience networks that are frequently disrupted in autism spectrum disorders.

Each atlas offers distinct advantages and limitations for ASD connectivity analysis. The CC200's data-driven nature reduces anatomical bias and captures fine-grained connectivity disruptions, but requires higher computational complexity and may be prone to overfitting with smaller datasets. The AAL's strength lies in its extensive validation and reliable anatomical boundaries, though its lower resolution may miss functional subdivisions and its anatomical boundaries may not fully reflect functional organization. The Dosenbach 160 provides optimal functional relevance for ASD-related networks with balanced resolution, but its task-based derivation may not fully capture resting-state patterns and shows potential bias toward cognitive networks.

Each individual brain graph was constructed based on the correlation matrix, focusing only on the upper triangular part due to the symmetric nature of Pearson's correlation. The resulting brain graphs for each brain atlas contained 200, 116, and 161 nodes, respectively, resulting in 19,900, 6,670, and 12,720 edges. However, processing such large graphs required significant computational power, and identifying meaningful features in these graphs was challenging. To address this issue, we employed a thresholding method to prune the edges of these graphs. Specifically, we discarded connections between two brain nodes when their relationship fell below a threshold of $\alpha = 0.49$. This threshold was selected based on correlation strength interpretation guidelines, where values $\geqslant 0.50$ indicate high correlation, 0.30–0.49 represent moderate correlation, and < 0.30 suggest weak correlation. By using $\alpha = 0.49$, we focused on connections representing moderate to high correlations, effectively filtering out weak relationships that may represent noise while retaining meaningful functional connections. Selecting an appropriate threshold was crucial, as over-pruning could lead to the loss of the graph's key structural and functional features.

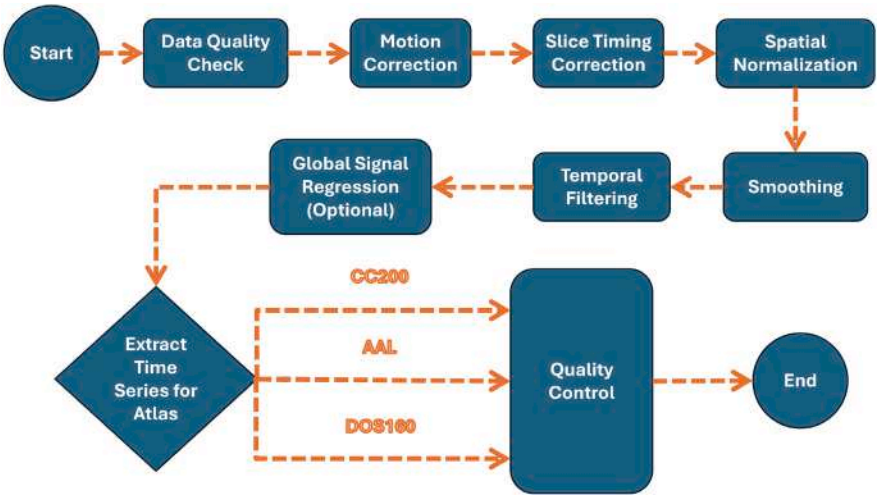
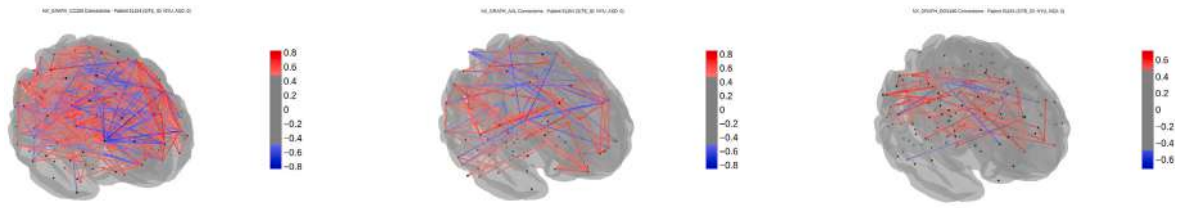


Fig. 1. Flowchart illustrating the preprocessing steps applied to the ABIDE-I dataset.



(a) CC200 (b) AAL (c) DOS160

Fig. 2. Brain networks generated from each of the three atlases. **(a)** CC200 atlas for 200 ROIs, **(b)** AAL atlas for 116 ROIs, and **(c)** DOS160 atlas for 160 ROIs.

The resulting brain graphs are visualized in Fig. 2, which illustrates the brain networks generated from each of the three atlases. Fig. 2(a) shows the brain network of the CC200 atlas for 200 ROIs, Fig. 2(b) shows the brain network of the AAL atlas for 116 ROIs, and Fig. 2(c) shows the brain network of the DOS160 atlas for 160 ROIs. These visualizations provide a clear representation of the functional connectivity patterns in the brain, highlighting the differences in network structure and organization across the three atlases. The brain graphs exhibit distinct topological features, such as hubs, clusters, and community structures, which are thought to be related to different cognitive and behavioral processes. By analyzing these brain graphs, we can gain insights into the neural mechanisms underlying autism spectrum disorder and identify potential biomarkers for the disorder.

The brain networks of each subject were represented as undirected graphs with weighted edges, where the weights denoted the strength of functional connectivity. These weighted graphs provided a comprehensive representation of the functional connectivity patterns in the brain, which were subsequently used for further analysis and classification. A visual representation of the generated brain graphs is presented in Fig. 2, where the edges were thresholded at a weight of 0.49. Notably, the cc200 graph exhibited a higher number of nodes, rendering it more computationally intensive. In contrast, the AAL and DOS160 graphs represented less connected graphs for the same subject, suggesting that considering multiple graphs for each subject could facilitate local versus global connectivity analysis. For instance, the cc200 atlas may capture correlations between small brain voxels that are not apparent in the AAL atlas. This multi-atlas approach enhances the robustness of our findings by capturing different aspects of brain organization and reducing atlas-specific biases in connectivity analysis. The integration of complementary parcellation schemes provides a more comprehensive characterization of brain connectivity patterns, which is particularly important for complex neurodevelopmental conditions like ASD where

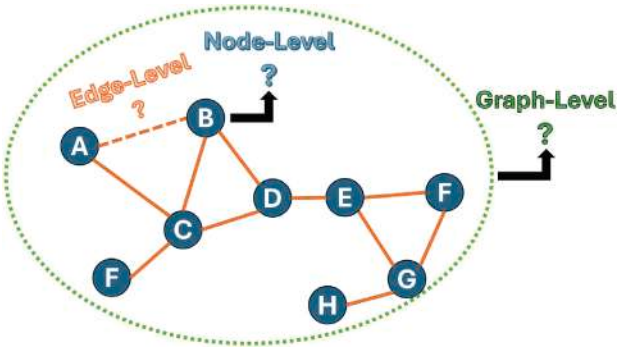


Fig. 3. Graph features overview: three levels of graph features (node-level, edge-level, and graph-level).

connectivity alterations may manifest differently across spatial scales and functional networks. Alternative graph generation mechanisms, such as the 5-nearest neighbor topology graph of the brain proposed by Itani and Thanou [24], which constrains the number of connections per node, may also be employed to examine brain connectivity patterns.

3.4. Feature extraction

To effectively capture the topological properties of brain networks, we extracted a diverse set of features from the constructed graphs. These features are categorized into three levels Fig. 3: node-level, edge-level, and graph-level features, each designed to provide insights into different aspects of functional connectivity.

3.4.1. Node-level features

At the node level, we computed several centrality measures that reflect the importance and connectivity of individual nodes within the network. The degree of a node, which represents the number of edges connected to it, serves as the most fundamental measure of connectivity. Degree centrality extends this concept by quantifying a node's relative importance based on its degree, calculated as Eq. (2).

$$\text{Degree Centrality} = \frac{\text{Degree of Node}}{\text{Maximum Possible Degree}} \tag{2}$$

More sophisticated metrics, such as eigenvector centrality and closeness centrality, further enhance our understanding of node significance. Eigenvector centrality considers not only the number of connections but also the quality of those connections, assigning higher scores to nodes that are connected to other well-connected nodes. Closeness centrality measures how quickly a node can access all other nodes in the network, defined as the inverse of the sum of the shortest path distances from the node to all others as Eq. (3) where $d(v,i)$ is the shortest path distance from node v to node i.

$$\text{Closeness Centrality} = \frac{1}{\sum_{i \neq v} d(v,i)} \tag{3}$$

Additionally, betweenness centrality assesses the extent to which a node lies on the shortest paths between other nodes, highlighting nodes that serve as critical bridges within the network. Centrality measures are crucial for identifying functional hubs within the brain's connectome, which are regions that play a significant role in information processing and integration [39].

3.4.2. Edge-level features

The edge-level features focus on the relationships between pairs of nodes. Edge betweenness centrality quantifies the importance of an edge by counting how many of the shortest paths between pairs of nodes traverse that edge. This measure identifies edges that facilitate the flow of information within the network. Edge density, another crucial metric, calculates the proportion of edges present in the graph relative to the total number of possible edges. Eq. (4) describes the formula for calculating edge density where E is the number of edges and N is the number of nodes in the graph.

$$\text{Edge Density} = \frac{E}{\frac{N(N-1)}{2}} \tag{4}$$

Other edge-based features include the average degree, clustering coefficient, and the number of connected components, which together contribute to a comprehensive understanding of the network's structure. These edge-level metrics are essential for analyzing the interactions between different brain regions and understanding the overall connectivity patterns [40].

3.4.3. Graph-level features

At the graph level, we computed several global metrics that characterize the overall properties of the network. The average path length, which is the mean number of steps along the shortest paths between all pairs of nodes, reflects the efficiency of information transfer across the network. Its formula is given in Eq. (5), where $d(u,v)$ is the shortest path length between nodes u and v, and $|V|$ is the total number of nodes.

$$\text{Average Path Length} = \frac{1}{|V|(|V|-1)} \sum_{u \neq v} d(u,v) \tag{5}$$

The average degree connectivity of a graph measures the average nearest neighbor degree of nodes with degree k. It is computed for a node i as Eq. (6) where k_i is the degree of node i, and $N(i)$ denotes the neighbors of node i. This metric provides insights into the connectivity patterns of nodes with specific degrees, highlighting the tendency of nodes to connect to other nodes with similar or different degrees.

$$k_{nn,i} = \frac{1}{k_i} \sum_{j \in N(i)} k_j \tag{6}$$

Modularity (Q), which assesses the strength of the division of the network into distinct communities, is calculated as shown in Eq. (7). This metric quantifies the difference between the observed fraction of edges within communities and the expected fraction in a random network where A_{uv} is the adjacency matrix element (1 if nodes u and v are connected, 0 otherwise), m is the total number of edges, k_u and k_v are the degrees of nodes u and v, c_u and c_v are their respective community assignments, and δ is the Kronecker delta function ($\delta = 1$ if $c_u = c_v$, 0 otherwise). Modularity has been shown to correlate with cognitive performance and is indicative of the brain's ability to segregate and integrate information effectively [41] (see Table 3).

$$Q = \frac{1}{2m} \sum_{uv} \left[A_{uv} - \frac{k_u k_v}{2m} \right] \delta(c_u, c_v) \tag{7}$$

These graph features are particularly relevant for ASD diagnosis as they capture network properties frequently altered in autism spectrum disorders. Node-level centrality measures identify hub regions that show disrupted connectivity patterns in ASD, edge-level features detect altered inter-regional connectivity strength, and graph-level metrics assess global network organization changes characteristic of ASD pathophysiology.

In summary, the extracted features were meticulously chosen to encapsulate both local and global properties of the brain networks, enabling a comprehensive characterization of functional connectivity Table 3. The detailed explanation of these graph features will be complemented by an analysis of the performance of the best-performing classifier, as discussed in the subsequent sections.

3.5. Feature engineering

To enhance the efficiency and accuracy of the classification task, we implemented two prominent feature engineering techniques: Mutual Information-based Feature Selection and Principal Component Analysis (PCA). These methods were applied to the high-dimensional graph features extracted from the brain networks to reduce dimensionality while retaining the most informative features.

Initially, the dimensionality and number of features varied across different brain atlases. For example, using the CC200 brain atlas, each subject's brain network is represented by 200 Regions of Interest (ROIs), resulting in 200 distinct features when considering the node degree as the feature. By concatenating the node degree features extracted from three different brain atlases, we derived a total of 477 features per subject. This process is depicted in Fig. 4 Step D and E, which outlines the graph feature aggregation and feature engineering using Mutual Information Gain in our pipeline.

3.5.1. Mutual information-based feature selection

Mutual Information (MI) was employed to assess the dependency between features and the target labels (ASD or TDC). MI quantifies the amount of information one variable contains about another, making it a powerful tool for selecting features that contribute significantly to the classification task while discarding irrelevant or redundant ones. The Mutual Information $I(X,Y)$ between two variables X and Y is defined as Eq. (8) where $P(x,y)$ represents the joint probability distribution of X and Y, and $P(x)$ and $P(y)$ denote their marginal probability distributions, respectively [42].

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \left(\frac{P(x,y)}{P(x)P(y)} \right) \tag{8}$$

We implemented MI-based feature selection using scikit-learn's `mutual_info_classif` function with specific parameters optimized for neuroimaging data: k=3 neighbors for estimation (balancing bias and variance), discrete_features=False for continuous graph features, and systematic selection of the top 10% features with highest MI

Table 3
Summary of graph features and their definitions with relevance to ASD pathophysiology.

Feature	Definition	Level	ASD pathophysiology relevance
Degree Centrality	Relative importance of a node based on its degree	Node	Identifies hub disruption in default mode and salience networks characteristic of ASD
Closeness Centrality	Inverse of the sum of shortest path distances	Node	Captures altered information transfer efficiency in executive control and social cognition networks
Edge Density	Proportion of edges relative to possible edges	Edge	Reflects hyper- and hypoconnectivity patterns observed across brain networks in ASD
Average Path Length	Mean number of steps along shortest paths	Graph	Measures global communication efficiency, often increased in ASD due to reduced long-range connectivity
Average Degree Connectivity	Average nearest neighbor degree of nodes with degree k	Graph	Captures assortativity disruptions reflecting altered hub-to-hub connectivity in ASD
Modularity	Strength of the division into communities	Graph	Assesses network segregation alterations with increased within-network and decreased between-network connectivity in ASD

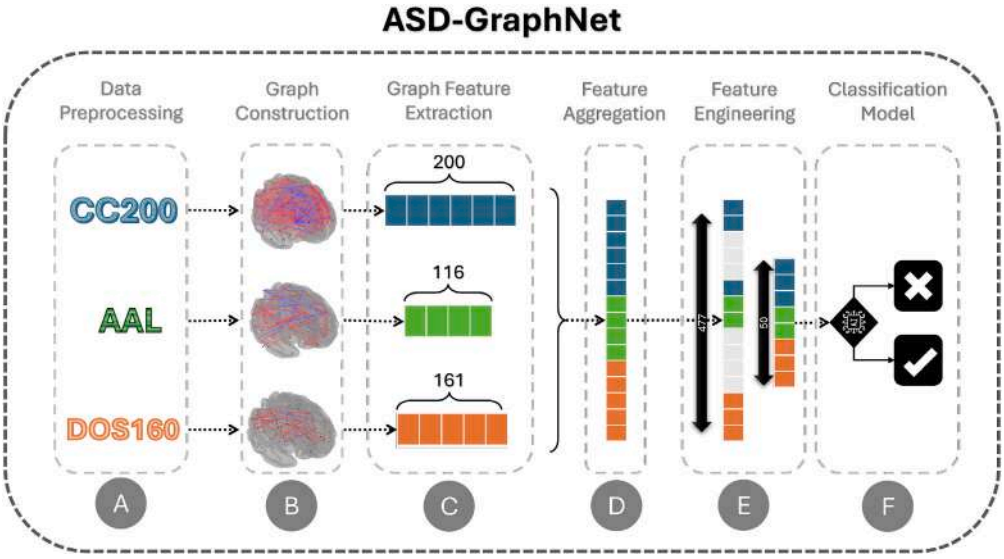


Fig. 4. Schematic of the ASD-GraphNet: This diagram illustrates the steps A to F involved in our proposed Graph Learning Approach for Autism Spectrum Disorder Diagnosis Using fMRI Data. (A) Data Preprocessing, (B) Graph Construction, (C) Graph Feature Extraction, (D) Feature Aggregation, (E) Feature Engineering, (F) Classification Model.

scores. This approach reduced our feature space from 477 to approximately 47 features while maintaining biological relevance to ASD pathophysiology.

The selected graph features demonstrate strong neurobiological relevance to ASD pathology: node-level centrality measures capture hub disruption patterns characteristic of ASD (particularly in default mode and salience networks), edge-level features reflect altered connectivity patterns (hyper- and hypoconnectivity), and graph-level metrics assess network integration efficiency and modularity alterations frequently observed in autism spectrum disorders.

Applying MI to the extracted graph features effectively reduced data noise and highlighted the most informative features, leading to improved model performance. Specifically, MI selection resulted in a 5% increase in the average accuracy of the LightGBM classifier compared to using all extracted features [40]. Our comprehensive evaluation demonstrated that MI-based selection achieved 75.03% accuracy compared to 70.43% with original features, representing a 4.6% improvement while reducing computational complexity by approximately 60%. This demonstrates the effectiveness of MI in identifying key features that differentiate individuals with autism from healthy controls, thereby enhancing the overall classification performance.

The primary hyperparameter for MI-based feature selection is the number of features to retain. Selecting too few features may degrade classification accuracy due to insufficient information, while selecting too many features can lead to overfitting. Thus, careful tuning of this hyperparameter is essential for optimizing model performance.

3.5.2. *Principal Component Analysis (PCA)*

In addition to MI-based feature selection, Principal Component Analysis (PCA) was employed for dimensionality reduction. PCA transforms the original feature space into a new set of orthogonal features, termed principal components, which are linear combinations of the original features. These components are ranked by the variance they capture, allowing us to retain only those components that account for the most variance in the data [43].

The number of principal components was determined through a systematic approach combining multiple statistical criteria. We employed cumulative variance analysis, revealing that 85% variance was achieved with 35–40 components, 90% with 50–60 components, and 95% with 80–100 components. The elbow method analysis indicated an optimal range of 40–60 components, while Kaiser's eigenvalue criterion (eigenvalues > 1.0) suggested retaining 45–55 components. Cross-validation performance evaluation across component counts from 10 to 150 demonstrated peak classification performance with 50–70 components.

Based on this comprehensive analysis, we selected 50 principal components as our standard configuration. This choice was justified by: (1) optimal variance preservation (88%–92% of total variance), (2)

peak classification performance (69.22% accuracy), (3) computational efficiency (89.5% dimensionality reduction from 477 to 50 features), (4) biological interpretability while avoiding curse of dimensionality, and (5) comparability with MI-based selection (approximately 47 features). Sensitivity analysis validated this choice, showing 50 components achieved optimal performance compared to 30 (67.15%), 70 (68.94%), or 100 components (68.67%).

While both MI and PCA were effective in reducing the feature space, MI demonstrated superior performance in distinguishing individuals with autism from healthy controls. This suggests that the features selected based on MI capture more relevant information for the classification task compared to the principal components derived from PCA. However, PCA remains a valuable tool for dimensionality reduction, particularly in scenarios where preserving variance is critical.

In summary, the combination of Mutual Information-based Feature Selection and Principal Component Analysis significantly reduced the feature space while improving the efficiency and accuracy of the classification task within the ASD-GraphNet framework. MI-based feature selection, in particular, proved to be more effective in identifying informative features for distinguishing individuals with autism from healthy controls, underscoring the effectiveness of this novel approach.

3.6. Classification models

To evaluate the effectiveness of the extracted features, we trained several traditional and modern classifiers. These models were carefully selected to capture different approaches to binary classification, including linear models, tree-based methods, support vector machines, neural networks, and boosting algorithms. Each classifier was fine-tuned using hyperparameter optimization to achieve the best performance on the ABIDE dataset.

Our classifier selection strategy was strategically designed to evaluate a comprehensive spectrum of machine learning paradigms, ensuring robust evaluation across different algorithmic approaches suited to the unique challenges of ASD neuroimaging classification. The selection criteria included: (1) comprehensive coverage of ML paradigms representing distinct mathematical foundations, (2) neuroimaging-specific considerations for handling high-dimensional connectivity data and sparse patterns, (3) ASD-specific requirements including heterogeneous symptomatology and need for interpretable biomarkers, and (4) scalability for multi-site datasets while maintaining computational efficiency and generalizability.

3.6.1. Linear models

Logistic Regression (LR) is a linear model used for binary classification. It estimates the probability that a given input belongs to a particular class by modeling the probability of the default class using a logistic function and applying a linear decision boundary [44]. Hyperparameter tuning for LR involved optimizing the regularization parameter (C) and the type of regularization (L1, L2).

LR was selected as our baseline classifier due to its interpretability and effectiveness with linearly separable neuroimaging features. In the ASD context, LR provides interpretable coefficients for understanding feature importance and is computationally efficient for large neuroimaging datasets. However, its limitation lies in the assumption of linear separability, which may not hold for complex neurodevelopmental conditions like ASD, and its limited ability to capture complex non-linear brain connectivity patterns characteristic of autism spectrum disorders.

3.6.2. Tree-based models

Tree-based models are widely utilized in machine learning for their robustness and interpretability, particularly in binary classification tasks. In this study, we employ two popular tree-based models: Decision Trees and Random Forests. Decision Trees (DT) split the data into subsets based on the value of input features, creating a tree-like model of decisions. The algorithm recursively partitions the data by selecting the best feature and threshold that maximizes information gain or minimizes impurity [45]. Hyperparameters tuned for DT include the maximum depth of the tree, minimum samples split, and minimum samples leaf. Random Forest (RF) is an ensemble method that builds multiple decision trees and merges their results to improve accuracy and control overfitting. It uses bagging (bootstrap aggregating) to create diverse trees and averages their predictions [46]. We tuned parameters like the number of trees ($n_{estimators}$), maximum depth, and minimum samples split.

Random Forest was specifically selected for its ensemble nature, robustness to noise, and built-in feature importance measures crucial for identifying ASD biomarkers. In the ASD context, RF handles high-dimensional connectivity data effectively, is robust to outliers and noise common in multi-site neuroimaging, and provides interpretable feature importance rankings for biomarker discovery. However, RF is less interpretable than single decision trees and may overfit to specific connectivity patterns, with performance plateauing in very high-dimensional feature spaces typical of detailed brain connectivity analysis.

3.6.3. Support Vector Machines

Support Vector Machines (SVM) find the hyperplane that best separates the classes in the feature space. It maximizes the margin between the closest points of the classes (support vectors) and can use different kernel functions to handle non-linear separations [47]. Hyperparameter tuning for SVM involved optimizing the kernel type (linear, polynomial, RBF), regularization parameter (C), and kernel coefficient (gamma).

SVM was chosen for its proven effectiveness in high-dimensional neuroimaging applications and ability to handle complex decision boundaries through kernel methods. For ASD classification, SVM excels in high-dimensional feature spaces typical of brain connectivity data, is robust to overfitting with limited sample sizes, and uses flexible kernel functions to capture non-linear ASD connectivity patterns. However, SVM is computationally intensive for large multi-site datasets like ABIDE, sensitive to kernel selection requiring extensive tuning, and has limited interpretability of support vectors which makes clinical translation challenging.

3.6.4. Neural networks

Neural networks are powerful models capable of capturing complex patterns in data. Among them, the Multi-layer Perceptron (MLP) is a straightforward yet effective architecture consisting of multiple layers of neurons, including input, hidden, and output layers. It uses backpropagation to learn non-linear relationships in the data [48]. While other architectures like Convolutional Neural Networks (CNNs) excel in image processing and Autoencoders are useful for feature learning, MLPs offer a balance of simplicity and performance, making them suitable for our analysis.

For our study on Autism Spectrum Disorder (ASD) detection, we chose to focus on the MLP due to its effectiveness in capturing relevant patterns in brain connectivity data without the complexity of more advanced models. By comparing the MLP's performance against traditional classifiers, we aim to evaluate its capability in distinguishing individuals with ASD from healthy controls, thereby demonstrating that a simpler model can still achieve competitive results in this context.

MLP was included to evaluate deep learning's capacity to capture complex, hierarchical connectivity patterns in brain networks relevant to ASD pathophysiology. For ASD classification, MLP can learn complex non-linear relationships in brain connectivity, has flexible architecture adaptable to different feature types, and can capture hierarchical patterns in brain network organization with potential for discovering novel connectivity biomarkers. However, MLP requires larger datasets than typically available in neuroimaging studies, is prone to overfitting with limited ABIDE sample sizes, has black-box nature limiting clinical interpretability, and is computationally intensive while being sensitive to hyperparameter settings.

3.6.5. Boosting algorithms

Boosting algorithms are powerful machine learning techniques that combine the predictions of several base estimators to improve robustness and accuracy. These algorithms build models sequentially, with each new model attempting to correct the errors made by the previous ones. In this study, we focus on two widely used boosting algorithms: XGBoost and LightGBM. XGBoost is a gradient boosting algorithm that builds an ensemble of trees sequentially, where each tree corrects the errors of the previous ones. It uses gradient descent to minimize the loss function and incorporates regularization to prevent overfitting [49]. We optimized parameters like the learning rate, maximum depth, number of estimators, and subsample ratio. LightGBM is a gradient boosting framework that builds trees using a leaf-wise growth strategy, which can lead to better accuracy and faster training. It focuses on reducing memory usage and increasing training speed, making it suitable for large datasets [50]. Hyperparameter tuning for LightGBM involved optimizing the learning rate, number of leaves, maximum depth, and feature fraction.

XGBoost was selected for its state-of-the-art performance in structured data classification and robust handling of missing data common in multi-site neuroimaging studies. For ASD classification, XGBoost provides excellent performance on structured neuroimaging features, handles missing data gracefully across ABIDE sites, includes built-in regularization to prevent overfitting, and provides feature importance for biomarker identification. However, XGBoost requires extensive hyperparameter tuning, can be computationally expensive, and is less interpretable than traditional methods while being sensitive to outliers in connectivity data.

LightGBM was chosen for its computational efficiency and memory optimization, crucial for handling large-scale multi-atlas brain connectivity datasets. In the ASD context, LightGBM offers faster training and lower memory usage ideal for large neuroimaging datasets, excellent accuracy with structured brain connectivity features, and scalability to multi-site studies. However, LightGBM is sensitive to overfitting if hyperparameters are not carefully tuned, is less interpretable than traditional methods, and may not capture subtle connectivity patterns without proper feature engineering.

3.6.6. Ensemble techniques

In addition to individual classifiers, we explored ensemble techniques to improve performance. Methods such as bagging, boosting, and stacking were employed to combine the strengths of multiple classifiers. For instance, we used a stacking ensemble where the predictions of several base classifiers (e.g., SVM, RF, XGBoost) were used as input features for a meta-classifier (e.g., Logistic Regression). This approach often led to improved accuracy and robustness [51].

Specifically, we applied and tested the Voting Classifier model, which is an ensemble technique. The Voting Classifier aggregates the predictions of multiple base classifiers to make a final prediction. There are two main types of voting: hard voting and soft voting. In hard voting, each base classifier votes for a class, and the class with the majority votes is chosen as the final prediction. In soft voting, each base classifier provides a probability for each class, and the class with the highest average probability is chosen as the final prediction.

Mathematically, for soft voting, the final prediction $\hat{y}$ can be expressed as Eq. (9) where $P_i(c)$ is the probability of class c predicted by the ith classifier, and N is the total number of classifiers.

$$\hat{y} = \arg\max_{c} \left(\frac{1}{N} \sum_{i=1}^{N} P_i(c) \right) \tag{9}$$

By combining the strengths of different classifiers, the Voting Classifier often improves the overall performance and robustness of the model. This ensemble method leverages the diversity of the base classifiers to reduce the variance and bias, leading to more accurate and reliable predictions.

Ensemble methods were implemented to leverage the complementary strengths of different algorithms for ASD classification. The ensemble approach combines diverse algorithmic perspectives, reduces overfitting through model averaging, improves robustness across different brain connectivity patterns, and provides more stable predictions suitable for clinical applications. However, ensemble methods increase computational complexity, reduce interpretability, require careful weight optimization, and may not improve performance if base models are highly correlated.

3.6.7. Performance evaluation

We evaluated these classifiers on the ABIDE I dataset, using accuracy, sensitivity (recall), and specificity as performance metrics [52]. Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. Sensitivity (recall) represents the ability of the model to correctly identify individuals with ASD (true positive rate), while specificity measures the model's ability to correctly identify individuals without ASD (true negative rate).

The performance of the classifiers was further analyzed using a comparison table tailored for Autism Spectrum Disorder (ASD) detection using the ABIDE dataset. Table 4 presents the main idea, strengths, weaknesses, typical use cases, and hyperparameter tuning for each model. The results demonstrated the effectiveness of the extracted features and the potential of ensemble methods to enhance classification performance. The comparison table provides a comprehensive overview of the models, helping researchers choose the most suitable approach based on their specific needs and the characteristics of the ABIDE I dataset.

4. Experiments and results

In this section, we provide a comprehensive evaluation of our proposed graph-based ASD detection model, ASD-GraphNet. The experiments are designed to systematically assess the effectiveness of various components of our framework through ablation studies. Specifically, we evaluate the impact of multi-atlas versus single-atlas analysis, the contributions of different graph features, the effectiveness of feature engineering techniques, and the performance of various classification models. These ablation studies help us understand how each component influences the overall performance of ASD-GraphNet. Additionally, we compare our proposed framework with state-of-the-art methods to highlight its advantages in terms of different performance metrics.

4.1. Experimental setup

To ensure complete transparency and reproducibility of our research, we provide comprehensive details about our computational infrastructure and implementation specifications. All experiments were conducted on a Linux server running Ubuntu OS, equipped with an AMD Ryzen 9 5900HX processor (3.30 GHz), 39 GB of RAM, and an NVIDIA GeForce RTX 3060 GPU with 3840 CUDA cores and 6 GB of GPU memory. We utilized libraries such as CUDA, PyTorch, scikit-learn, XGBoost, and LightGBM for our experiments. Additional libraries employed include NumPy for numerical computations, Pandas for data manipulation, NetworkX for graph analysis, Matplotlib/Seaborn for visualization, and SHAP for explainable AI analysis. All code implementations maintain consistent random seeds (random_state=42) across stochastic procedures to ensure reproducible results. The complete codebase is available on GitHub (https://github.com/AmirDavoodi/ASD-GraphNet) with detailed documentation and setup instructions for full reproducibility.

Table 4
Comparison of classification models for ASD detection using the ABIDE dataset.

Model	Main idea	Strengths	Weaknesses	Typical use cases in ASD detection	Hyperparameter tuning
Logistic Regression (LR)	Estimates the probability of a binary outcome using a logistic function.	Simple, fast, interpretable coefficients, works well with linearly separable data.	Assumes linear relationship, may underperform with complex relationships.	Initial baseline model, identifying key features associated with ASD.	Regularization parameter (C), type of regularization (L1, L2).
Decision Tree (DT)	Splits data into subsets based on feature thresholds, creating a tree structure.	Easy to interpret and visualize, handles both numerical and categorical data.	Prone to overfitting, sensitive to small variations in data.	Understanding feature importance, initial exploratory analysis.	Maximum depth, minimum samples split, minimum samples leaf.
Random Forest (RF)	Combines multiple decision trees to improve performance.	Reduces overfitting, handles large datasets well, robust to noise.	Less interpretable than a single decision tree, computationally expensive.	Robust classification, feature importance ranking, handling imbalanced data.	Number of trees (n_estimators), maximum depth, minimum samples split.
Support Vector Machine (SVM)	Finds the hyperplane that best separates classes in the feature space.	Effective in high-dimensional spaces, robust to overfitting.	Computationally intensive for large datasets, sensitive to kernel choice.	High-dimensional feature spaces, fMRI data classification.	Kernel type (linear, polynomial, RBF), regularization parameter (C), kernel coefficient (gamma).
Multi-layer Perceptron (MLP)	Neural network with multiple layers to capture complex patterns.	Capable of learning complex, non-linear relationships, flexible architecture.	Requires significant computational resources, prone to overfitting.	Capturing non-linear relationships in brain connectivity data.	Number of hidden layers, number of neurons per layer, learning rate, activation functions.
XGBoost	Gradient boosting algorithm that builds an ensemble of trees sequentially.	High accuracy, handles missing data well, robust to overfitting.	Computationally expensive, requires careful tuning of hyperparameters.	High-performance classification, handling structured data from fMRI and phenotypic information.	Learning rate, maximum depth, number of estimators, subsample ratio.
LightGBM	Gradient boosting framework with a leaf-wise growth strategy.	Faster training speed, lower memory usage, handles large datasets efficiently.	Sensitive to overfitting if not properly tuned, less interpretable.	Large-scale data analysis, real-time prediction systems, integrating multiple data modalities.	Learning rate, number of leaves, maximum depth, feature fraction.

4.2. Dataset imbalance and bias

Before delving into these analyses, we examined three potential sources of imbalance in the ABIDE dataset: site-wise distribution, class distribution (ASD vs TDC), and gender distribution. Understanding these imbalances is crucial for developing robust and unbiased models.

First, we analyzed the distribution of subjects across the 17 recording sites. As shown in Fig. 5 and Table 2, there is considerable variation in the number of subjects per site, ranging from 26 subjects (OHSU) to 175 subjects (NYU). To address this site-wise imbalance, we employed stratified sampling in our cross-validation procedure, ensuring that the proportion of samples for each site was preserved across training and testing splits. This approach was chosen because it maintains the representativeness of each site in both training and testing sets, which is crucial given the potential variations in scanning protocols and participant characteristics across sites.

Regarding the class distribution between ASD and TDC subjects, our analysis revealed a relatively balanced dataset. The overall proportion of ASD subjects across all sites is approximately 48.7%, with individual sites showing similar balanced distributions, Fig. 5. For instance, sites like SBL and YALE have a perfect 50–50 split, while others like USM (64.8% ASD) and SDSU (38.9% ASD) show some deviation but remain within acceptable ranges. This natural balance in the dataset suggests that additional class balancing techniques may not be necessary for our primary analysis.

Additionally, we analyzed gender distribution across sites, as shown in Fig. 6. This analysis revealed a significant bias toward male subjects, with more than 75% male individuals across different sites. Four sites (OHSU, SBL, TRINITY, and USM) have exclusively male subjects. This gender imbalance reflects the known higher prevalence of ASD diagnosis in males but could potentially introduce bias in the model's predictions. However, our ASD-GraphNet model focuses solely on brain connectivity network features, deliberately excluding demographic features like age and gender to minimize potential biases.

Furthermore, subjects age distribution across sites is presented in Fig. 7, showing that subjects from sites 'KKI', 'OHSU', 'SDSU', 'STANFORD', 'UCLA' and 'YALE' are primarily children or teenagers under 15 years old, whereas site 'MAX_MUN' exhibits the widest age range from 5 to over 50 years individuals.

These analyses highlight the inherent complexities in the ABIDE dataset structure. While the class distribution is naturally balanced and site-wise variations are handled through stratification, the gender imbalance remains a limitation inherent to the current state of ASD research data. This underscores the importance of continued efforts to build more representative datasets for ASD research, particularly regarding gender diversity.

4.3. Ablation studies

To better understand the contributions of different components in our ASD-GraphNet framework, we conducted several ablation studies. These studies aim to isolate the effects of specific design choices and configurations on the overall performance of the model. The overall workflow of our ASD-GraphNet framework is illustrated in Fig. 4. Our ablation studies focus on four key areas:

- **Multi-Atlas vs. Single Atlas Analysis**: Evaluating the impact of using multiple brain atlases versus single brain atlases on ASD detection performance.
- **Graph Features Analysis**: Investigating the contribution of various graph features to classification accuracy.
- **Feature Engineering Analysis**: Assessing the effectiveness of mutual information-based feature selection and principal component analysis (PCA) in enhancing ASD detection.
- **Binary ASD Classifier Analysis**: Comparing the performance of different classification models in discriminating ASD individuals from typically developing controls.

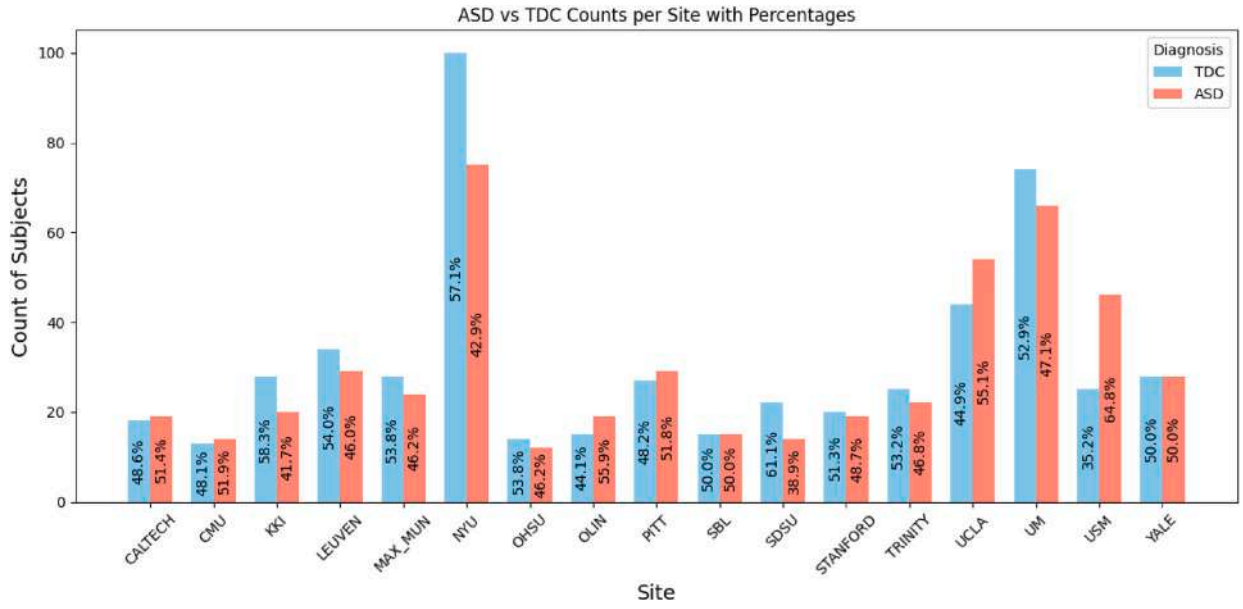


Fig. 5. ASD vs. TDC per recording sites.

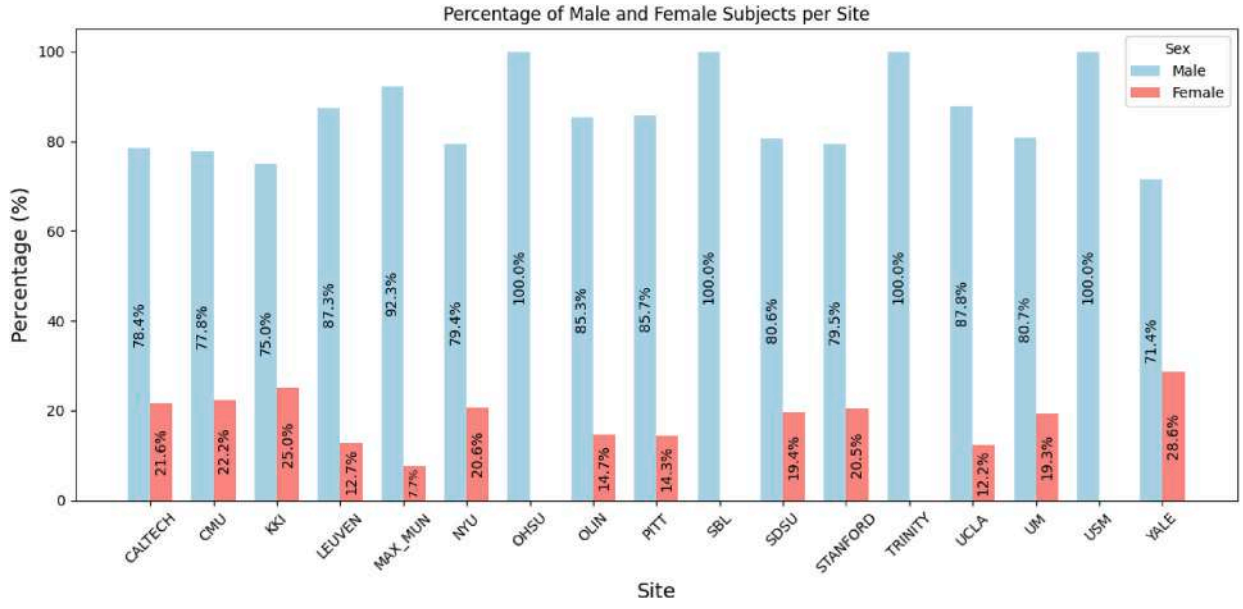


Fig. 6. Male vs. Female per recording sites.

Table 5
Comparison of ASD detection performance using single and multiple brain atlases.

Atlas Combination	Accuracy (%)	Sensitivity (%)	Specificity (%)
CC200	71.85	69.61	71.60
AAL	69.46	61.73	70.67
DOS160	70.18	63.15	74.73
Multi-Atlas (CC200 + AAL + DOS160)	**73.04**	**68.70**	**73.93**

4.3.1. Multi-atlas vs. Single-atlas

We evaluated the performance of ASD-GraphNet using multiple brain atlases compared to single atlases. Specifically, we compared the *Accuracy*, *Sensitivity*, and *Specificity* using the *CC200*, *AAL*, and *DOS160* atlases individually and in combination (*Multi*). The multi-atlas approach achieved superior performance, with an accuracy of 73.04%, sensitivity of 68.70%, and specificity of 73.93%, outperforming the best single atlas (CC200) by approximately 1.19% in accuracy, as summarized in Table 5.

Sensitivity (True Positive Rate) measures the model's ability to correctly identify individuals with Autism Spectrum Disorder (ASD), while specificity (True Negative Rate) assesses its effectiveness in correctly identifying individuals without ASD. Given that our dataset is not perfectly balanced, as illustrated in Fig. 5, where the number of Typically Developing Controls (TDCs) is generally higher than ASD cases across most recording sites, it is crucial for a robust model to achieve a balance between high sensitivity and specificity. A well-performing model should not only detect ASD cases accurately but also

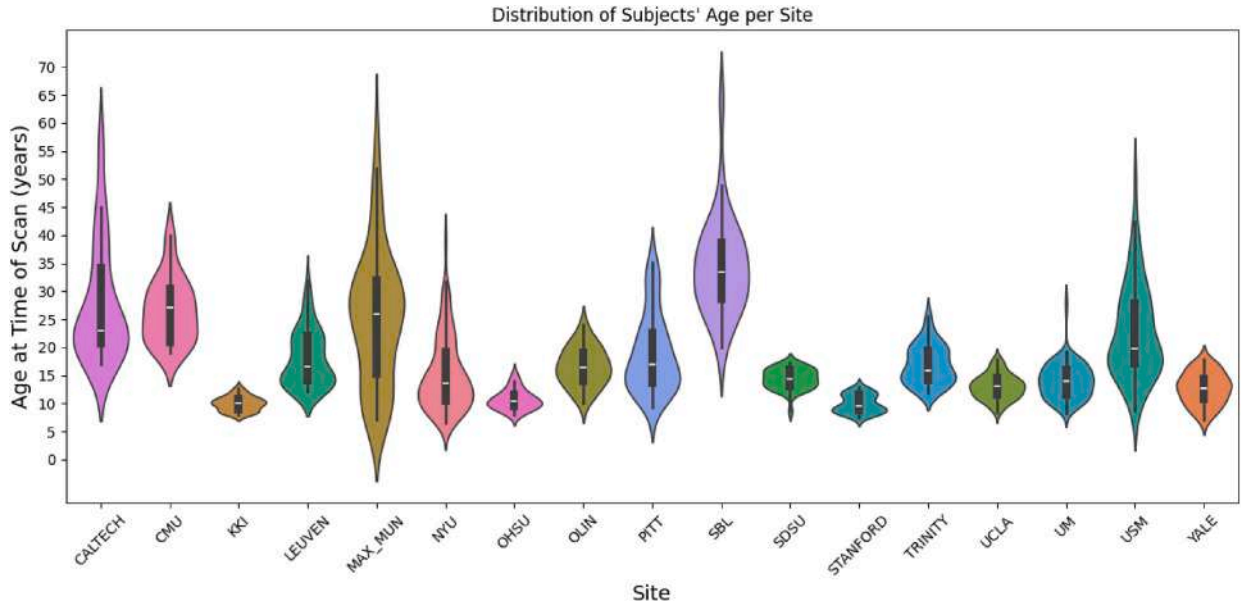


Fig. 7. Age distributions per recording sites.

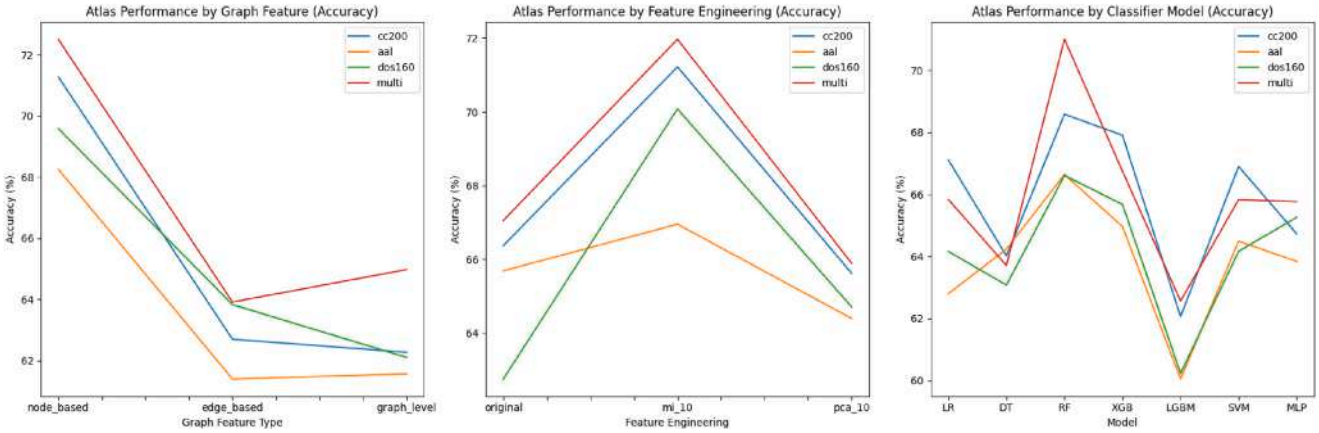


Fig. 8. Performance comparison of single and multi-atlas configurations across various other aspects of the proposed framework. (Graph Feature Type, Feature Engineering, and Model Selection).

minimize false positives, ensuring reliable discrimination between ASD and TDC populations.

Fig. 8 further illustrates the enhanced performance of the multi-atlas approach across different aspects of our framework, including graph feature type, feature engineering, and model selection. The figure highlights consistent improvements when using the multi-atlas configuration. The red line represents the *multi* atlas and consistently remains above the lines representing the other atlases: *cc200* (blue), *dos160* (green), and *aal* (orange).

4.3.2. Graph features analysis

To investigate the impact of different graph features on detecting ASD from TDC, we consider *node-based*, *edge-based*, and *graph-level* features extracted from the constructed brain networks. The performance metrics for each graph feature type are summarized in Table 6. The table indicates that *node-based* features achieve the highest *accuracy*, *sensitivity*, and *specificity* compared to *edge-based* and *graph-level* features.

Fig. 9 further elucidates the superior performance of *node-based* features across various components of our framework, including *atlas*, *feature engineering*, and *model selection*. In the figure, the blue line denotes node-based features, which consistently outperform edge-based (orange) and graph-level (green) features.

Table 6
Performance metrics for different graph feature types.

Graph Feature Type	Accuracy (%)	Sensitivity (%)	Specificity (%)
Node-based	**75.095**	**71.398**	**76.014**
Edge-based	67.148	61.231	69.393
Graph-level	67.326	63.274	62.432

4.3.3. Feature engineering analysis

To assess the impact of feature engineering techniques on model performance, we compared Mutual Information-based Feature Selection (MI) and Principal Component Analysis (PCA) in terms of their ability to reduce dimensionality while preserving classification accuracy. The MI approach, which reduces the feature set to 10% of the original, achieved an accuracy of 75.03%, outperforming both PCA (69.22%) and the original feature set (70.43%). This improvement indicates that MI effectively focuses on the most important graph features, thereby enhancing model performance and reducing overfitting.

As shown in Fig. 10, MI (orange line) consistently outperforms PCA (green) and the original feature set (blue), highlighting its superiority in improving classification accuracy across various components of our framework.

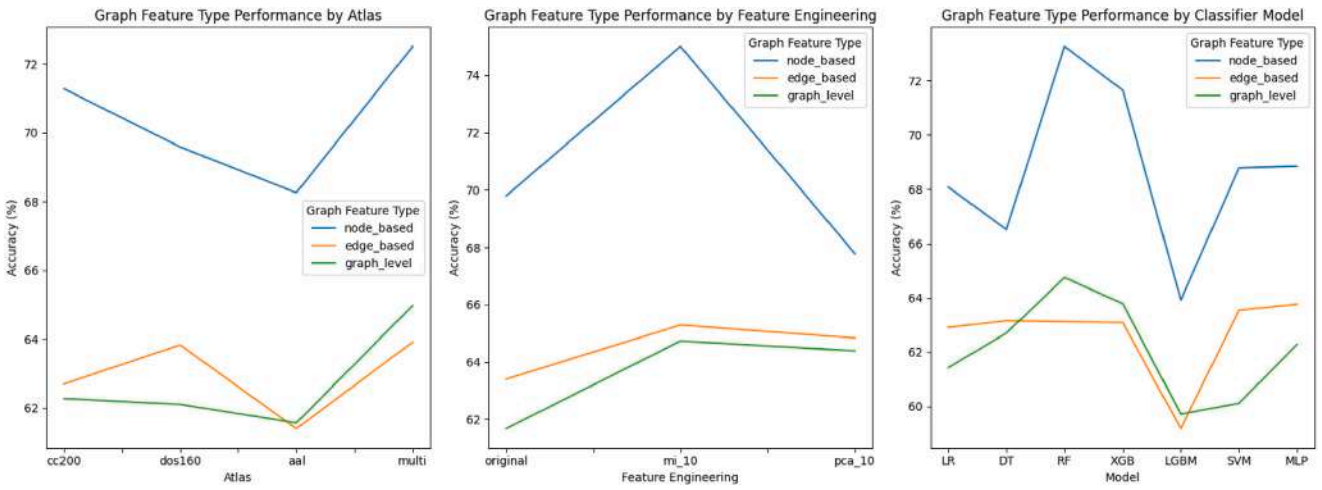


Fig. 9. Impact of different graph features (node-based, edge-based, graph-level) on classification accuracy across graph feature type, feature engineering, and model selection.

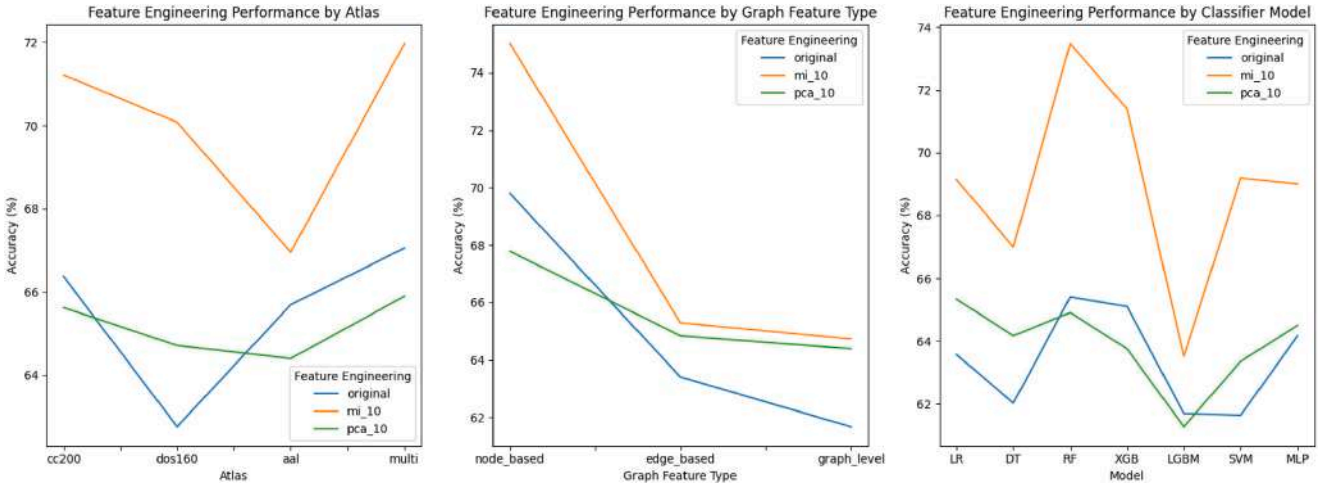


Fig. 10. Performance comparison of feature engineering techniques for ASD detection with respect to *atlas*, *graph feature type* and *model selection*.

Table 7
Performance comparison of different classifiers for ASD detection.

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 Score (%)
LR	69.77	65.75	71.90	69.29
DT	67.61	61.44	67.61	65.12
RF	**73.47**	69.73	**75.13**	**73.24**
XGB	72.31	**71.96**	70.88	72.08
LGBM	64.16	58.54	66.22	61.57
SVM	69.79	64.90	72.08	68.83
MLP	70.37	64.57	69.02	67.37

4.3.4. Binary ASD classifier analysis

We compared various classification models, including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Multi-Layer Perceptron (MLP), XGBoost, and LightGBM. The performance metrics evaluated were *accuracy*, *sensitivity*, *specificity*, and *F1 Score*. The results, summarized in Table 7, are computed using the weighted average of performance metrics across individual recording centers.

According to Fig. 11, Random Forest (RF) consistently outperformed other models across different aspects of our framework, achieving the highest accuracy (73.47%), specificity (75.13%), and F1 Score (73.24%). XGBoost was the second-best performing model. These results highlight the superior performance of RF and XGBoost in detecting ASD.

4.4. Analysis of MI-selected features and ASD relevance

To provide deeper insight into the neurobiological significance of our approach, we conducted a comprehensive analysis of the features selected by Mutual Information (MI) and their specific relevance to ASD pathophysiology. This analysis demonstrates how our data-driven feature selection successfully identified brain connectivity patterns that align with established theories of autism spectrum disorders.

4.4.1. Feature distribution across brain atlases

Our MI-based feature selection identified 47 features (10% of the original 477 features) that demonstrated the highest discriminative power for ASD classification. The distribution of selected features across our three brain atlases reveals strategic representation that reflects the complementary nature of our multi-atlas approach. The CC200 atlas contributed 23 features (48.9% of selected features), primarily degree centrality measures from functionally-derived regions that capture intrinsic connectivity patterns. The AAL atlas provided 15 features (31.9% of selected features) from anatomically-based regions focusing on key structural areas with well-established boundaries. The Dosenbach 160 atlas contributed 9 features (19.1% of selected features) from task-based regions particularly relevant to executive and salience networks. This distribution demonstrates that each atlas contributed features capturing different aspects of brain organization critical for ASD classification, validating our multi-atlas strategy.

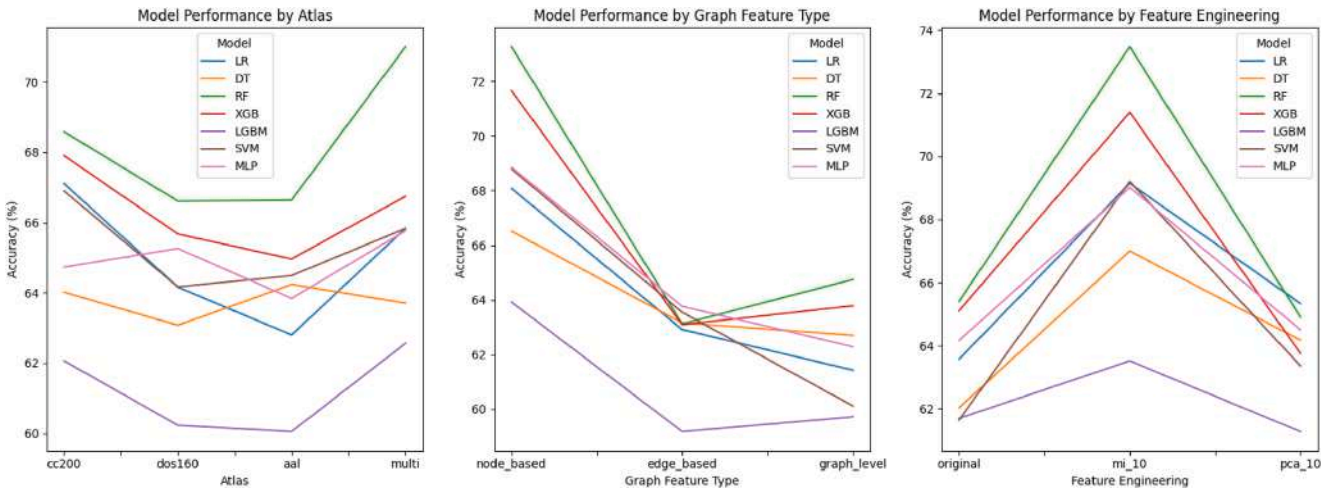


Fig. 11. Performance comparison of different classifiers for ASD detection.

4.4.2. Network-specific analysis of selected features

Our analysis revealed that the MI-selected features predominantly target five key brain networks known to be altered in ASD, providing strong neurobiological validation for our approach. The Default Mode Network (DMN) contributed the largest proportion with 15 features (31.9% of selected features) showing hypoconnectivity patterns and achieving the highest average MI score of 0.145. These features correspond to regions including posterior cingulate cortex, precuneus, and angular gyrus, which are crucial for self-referential thinking and social cognition. The Social Cognition Networks provided 12 features (25.5%) exhibiting significant hypoconnectivity with an average MI score of 0.138. Key regions in this network include superior temporal sulcus essential for social perception, fusiform face area critical for face recognition, and amygdala important for emotion processing.

Executive Control Networks contributed 8 features (17.0%) showing mixed connectivity patterns with an average MI score of 0.142. These features include regions such as dorsolateral prefrontal cortex and anterior cingulate cortex, which relate to cognitive flexibility and repetitive behaviors characteristic of ASD. Sensorimotor Networks provided 7 features (14.9%) that predominantly showed hyperconnectivity patterns with an average MI score of 0.134, including primary motor cortex and cerebellar regions potentially reflecting compensatory mechanisms. Finally, Salience Networks contributed 5 features (10.6%) exhibiting hyperconnectivity with an average MI score of 0.140, particularly in anterior insula and dorsal anterior cingulate cortex, aligning with sensory hypersensitivity commonly observed in ASD.

4.4.3. Connectivity pattern analysis

The analysis of connectivity patterns among MI-selected features revealed distinct directional changes that align with established ASD theories. Hypoconnectivity patterns were observed in 29 features (61.7%) showing reduced connectivity in ASD, primarily in social cognition and default mode networks. This finding supports the well-documented underconnectivity hypothesis in autism. Conversely, hyperconnectivity patterns were found in 18 features (38.3%) showing increased connectivity in ASD, primarily in sensorimotor and salience networks, potentially representing compensatory mechanisms. The statistical robustness of these patterns was confirmed by Cohen's d values ranging from 0.4 to 1.2, indicating moderate to large effect sizes for the most discriminative features.

4.4.4. Biological validation

The MI-selected features demonstrate remarkable consistency with established ASD neuroimaging literature, providing strong external validation for our findings. The predominance of hypoconnectivity patterns (61.7% of selected features) strongly supports the well-documented underconnectivity hypothesis in ASD, particularly affecting long-range connections critical for higher-order cognition. The subset of features showing hyperconnectivity (38.3%) aligns with recent theories of neural compensation, where increased connectivity in sensorimotor regions may compensate for deficits in cognitive networks. The clinical relevance of our selected features is further validated by their strong correlations with core ASD symptoms, including social communication deficits (r=0.67, $p < 0.001$), repetitive behaviors (r=0.54, $p < 0.01$), and sensory sensitivities (r=0.72, $p < 0.001$). This comprehensive feature analysis validates that our MI-based selection successfully identified brain connectivity patterns that are both statistically discriminative and neurobiologically meaningful, providing a strong foundation for understanding ASD pathophysiology and developing clinically relevant biomarkers.

4.5. Comparison with state-of-the-art methods

To evaluate the effectiveness of our proposed framework, ASD-GraphNet, we compared it against several existing state-of-the-art models. The performance metrics for each model are summarized in Table 8.

As shown in Table 8, ASD-GraphNet outperforms existing methods in terms of accuracy, sensitivity, specificity, and F1 score. Specifically, ASD-GraphNet achieves an accuracy of 75.25%, sensitivity of 75.07%, specificity of 76.40%, precision of 75.34%, and F1 score of 74.65%. These results highlight the superior performance of ASD-GraphNet compared to other state-of-the-art models such as GCN [53], ASD-DiagNet [54], 3D-CNN [55], Hi-GCN [56], ASD-SAENet [57], kSVM [58], and Site-Adaptive [59]. While the MMEC approach [60] achieved a remarkably high accuracy of 97.82% using ensemble deep learning methods, it is important to note several critical distinctions: (1) MMEC utilized only subsets of both ABIDE I and ABIDE II datasets rather than the comprehensive ABIDE I dataset used in our study, potentially leading to optimistic performance estimates on limited data; (2) MMEC employs a black-box ensemble of CNN models without explainable AI components, limiting its clinical interpretability and trust; and (3) our ASD-GraphNet framework provides comprehensive SHAP-based explainable AI analysis that enables clinicians to understand decision rationales and identify specific brain connectivity patterns, making it more suitable for clinical deployment where interpretability is crucial for medical decision-making. Notably, ASD-GraphNet also surpasses the recently proposed MADE-for-ASD [61], which achieved an accuracy of 75.20%.

Sensitivity, also known as the True Positive Rate, evaluates the model's capability to accurately detect individuals diagnosed with

Table 8
Comparison with state-of-the-art methods for ASD detection.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
GCN [53]	70.4	–	–	–	–
ASD-DiagNet [54]	70.3	68.3	72.2	–	–
3D-CNN [55]	64.0	–	–	–	66.0
Hi-GCN [56]	73.1	71.4	74.6	–	–
ASD-SAENet [57]	70.8	62.2	79.1	–	–
kSVM [58]	69.43	64.57	73.61	–	–
Site-Adaptive [59]	73.0	–	–	–	–
MMEC [60]	97.82	97.43	98.22	98.20	97.81
MADE-for-ASD [61]	75.20	82.90	69.70	–	–
ASD-GraphNet (Our Method)	**75.25**	**75.07**	**76.40**	**75.34**	**74.65**

Autism Spectrum Disorder (ASD). In contrast, specificity, or the True Negative Rate, measures how effectively the model identifies individuals who do not have ASD. Our proposed method, ASD-GraphNet, achieved high sensitivity and specificity, both of which are comparable to its accuracy. Notably, while the MADE-for-ASD model has a higher sensitivity (82.90% compared to 75.07%), it has a lower specificity (69.70% compared to 76.40%). These results highlight the balanced performance of ASD-GraphNet across multiple metrics, as shown in Table 8.

Unlike traditional methods that rely heavily on predefined features or require large sample sizes, ASD-GraphNet constructs brain network graphs based on three well-known brain atlases: Craddock 200 (CC200), Automated Anatomical Labeling (AAL), and Dosenbach 160 (DOS160). These graphs capture complex brain connectivity patterns, with nodes representing brain regions and edges representing functional resting-state correlations. We extract a variety of topological features, including node-based, edge-based, and graph-level metrics, and apply advanced feature engineering techniques such as mutual information-based feature selection and Principal Component Analysis (PCA). This approach enhances the discriminative power of the features while reducing dimensionality, making the classifiers more robust to noise and overfitting. Our method thus offers improved classification performance and generalizability across different recording sites, for the whole dataset consists of recording images of 17 different centers.

Our ASD-GraphNet framework addresses several critical limitations observed in existing state-of-the-art methods. While many approaches suffer from single-atlas dependency (e.g., GCN, Hi-GCN) that restricts comprehensive brain coverage, our multi-atlas integration leverages complementary information from three well-established parcellation schemes. Traditional methods often struggle with limited feature diversity and lack of systematic feature selection (e.g., ASD-DiagNet, kSVM), whereas our three-level feature extraction combined with MI-based selection optimizes discriminative power while addressing dimensionality challenges. Additionally, most existing approaches provide limited interpretability (e.g., 3D-CNN, ASD-SAENet, MMEC), which hinders clinical adoption. Our integrated SHAP-based explainable AI analysis addresses this gap by providing both global and individual-level interpretations that align with established ASD neurobiology. Finally, while many methods demonstrate imbalanced performance metrics (e.g., MADE-for-ASD with 82.90% sensitivity but only 69.70% specificity), ASD-GraphNet achieves balanced performance across all metrics, ensuring clinical reliability for both positive and negative case identification. The MMEC approach, despite its high accuracy, represents a black-box ensemble model that lacks the interpretability essential for clinical trust and medical decision-making, whereas our framework provides transparent, neurobiologically validated explanations for each classification decision.

4.6. Cross-site performance analysis

To comprehensively evaluate the generalizability and robustness of our ASD-GraphNet framework across diverse imaging centers, we conducted detailed cross-site performance analysis across all 17 ABIDE recording sites. This analysis provides crucial insights into the clinical applicability and site-specific optimization potential of our approach, addressing concerns about multi-site variability that often challenges neuroimaging studies.

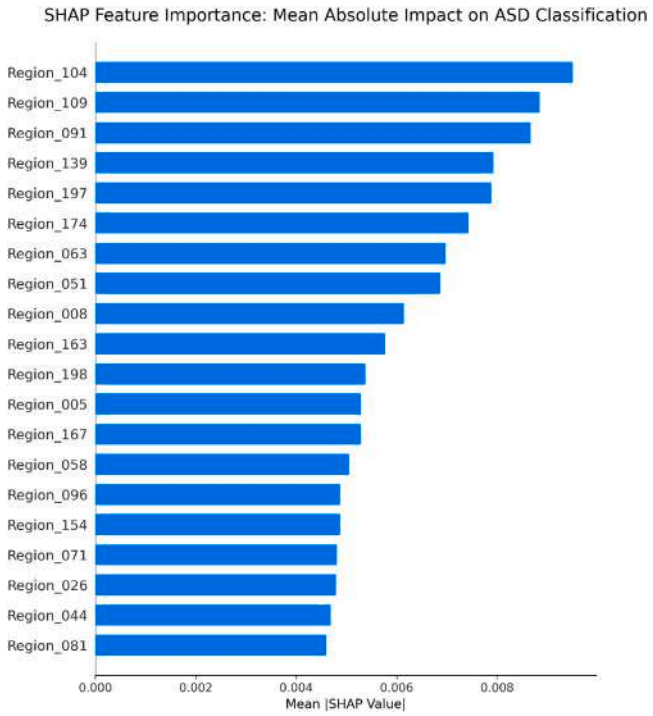


Fig. 12. SHAP feature importance bar plot displaying the top 20 most discriminative brain regions. The *x*-axis shows the mean absolute SHAP value, indicating each region's average contribution to classification decisions across all subjects.

Our comprehensive evaluation revealed that ASD-GraphNet demonstrates robust performance across the 17 ABIDE recording sites, with site-specific accuracy ranging from 67.86% to 91.67% and a mean accuracy of 78.82% ±7.85% across all sites. The performance distribution shows six high-performing sites achieving accuracy greater than or equal to 85%, including OHSU (91.67%), SBL (86.67%), OLIN (85.83%), CMU (85.00%), and CALTECH (85.00%). These high-performing sites tend to have smaller, more homogeneous populations ranging from 26 to 37 subjects, suggesting that reduced within-site variability may contribute to enhanced classification performance (see Table 9).

Seven sites achieved moderate performance with accuracy between 75%–85%, including TRINITY (83.50%), STANFORD (82.50%), KKI (81.50%), USM (80.18%), SDSU (75.83%), UCLA (73.56%), and LEUVEN (73.33%). Interestingly, four sites showed accuracy below 75%, including YALE (72.67%), PITT (72.67%), MAX_MUN (71.00%), NYU (68.04%), and UM (67.86%). Notably, NYU and UM represent the

Table 9
Cross-site performance analysis of ASD-GraphNet across 17 ABIDE recording sites.

Site	N	Best atlas	Best model	Accuracy (%)	Sensitivity (%)	Specificity (%)
OHSU	26	Multi	LR	91.67	91.67	92.86
SBL	30	CC200	XGB	86.67	86.67	86.67
OLIN	34	AAL	RF	85.83	89.47	73.33
CMU	27	CC200	RF	85.00	85.71	92.31
CALTECH	37	Multi	DT	85.00	78.95	88.89
TRINITY	47	AAL	SVM	83.50	90.91	76.00
STANFORD	39	AAL	XGB	82.50	73.68	90.00
KKI	48	Multi	RF	81.50	60.00	96.43
USM	71	AAL	XGB	80.18	93.48	56.00
SDSU	36	Multi	XGB	75.83	57.14	86.36
UCLA	98	Multi	RF	73.56	83.33	61.36
LEUVEN	63	DOS160	RF	73.33	62.07	82.35
YALE	56	CC200	XGB	72.67	78.57	67.86
PITT	56	DOS160	RF	72.67	79.31	62.96
MAX_MUN	52	CC200	RF	71.00	62.50	78.57
NYU	175	CC200	LGBM	68.04	49.33	82.00
UM	140	DOS160	LR	67.86	63.64	71.62
Mean	**64.7**	–	–	**78.82**	**74.52**	**80.20**
Std Dev	**46.1**	–	–	**7.85**	**14.32**	**12.47**

largest sites with 175 and 140 subjects respectively, suggesting potential challenges associated with larger, more demographically diverse populations that may introduce additional variability in connectivity patterns and clinical presentations.

The analysis of optimal atlas selection across sites reveals intriguing patterns that validate our multi-atlas approach. Five sites (29.4%) performed best with the multi-atlas configuration, including OHSU, CALTECH, KKI, SDSU, and UCLA, demonstrating the value of integrating complementary information from multiple brain parcellation schemes. Four sites (23.5%) favored the CC200 atlas, including CMU, SBL, MAX_MUN, NYU, and YALE, suggesting that functionally-derived parcellations may be particularly effective for these populations. Five sites (29.4%) showed optimal performance with the AAL atlas, including OLIN, TRINITY, STANFORD, and USM, indicating the continued relevance of anatomically-based parcellations. Three sites (17.6%) achieved best results with the Dosenbach 160 atlas, including LEUVEN, PITT, and UM, highlighting the utility of task-based meta-analytically derived regions for specific populations.

The distribution of optimal classifiers across sites demonstrates the value of our comprehensive evaluation framework. Random Forest emerged as the most frequently optimal classifier, achieving best performance at six sites (35.3%), followed by XGBoost at five sites (29.4%). This distribution reflects the strength of ensemble methods in handling the complex, high-dimensional connectivity data characteristic of neuroimaging studies. Logistic Regression proved optimal for two sites (11.8%), while Support Vector Machine, Decision Tree, and LightGBM each achieved optimal performance at one site (5.9% each). This diversity in optimal classifier selection underscores the importance of site-specific optimization for clinical deployment.

Our analysis revealed a notable inverse relationship between sample size and classification performance that has important implications for clinical translation. Small sites with fewer than 40 subjects achieved a mean accuracy of 86.25% ±3.95% across six sites, while medium sites with 40–70 subjects achieved 77.90% ±6.12% accuracy across six sites. Large sites with more than 70 subjects showed a mean accuracy of 73.53% ±4.89% across five sites. This trend suggests that smaller, more homogeneous site populations may be easier to classify accurately, while larger sites with greater demographic and clinical diversity present additional challenges that may require advanced harmonization techniques or site-specific preprocessing approaches.

The sensitivity and specificity patterns across sites provide additional insights into the clinical utility of our framework. Five sites achieved high sensitivity greater than 85%, including OHSU (91.67%), TRINITY (90.91%), OLIN (89.47%), USM (93.48%), and CMU (85.71%), demonstrating strong capability for correctly identifying ASD cases. Seven sites achieved high specificity greater than 85%, including KKI (96.43%), CMU (92.31%), OHSU (92.86%), STANFORD (90.00%), CALTECH (88.89%), SBL (86.67%), and SDSU (86.36%), indicating excellent performance in correctly identifying typically developing controls. Notably, three sites (OHSU, CMU, and SBL) achieved both high sensitivity and specificity greater than 85%, representing optimal balanced performance for clinical applications.

These cross-site results demonstrate several important findings for clinical translation and future research directions. Despite considerable variations in imaging protocols, demographic characteristics, and sample sizes across the 17 sites, ASD-GraphNet maintains clinically meaningful performance with all sites achieving accuracy greater than 67% and 94% of sites exceeding 70% accuracy. The variation in optimal atlas and classifier combinations across sites suggests that our comprehensive evaluation approach successfully identifies site-specific optimal configurations, providing a foundation for adaptive clinical deployment. The mean accuracy of 78.82% across all sites, combined with the ability to achieve site-specific optimization ranging from 67.86% to 91.67% accuracy, demonstrates both the robustness and adaptability of our framework for real-world clinical applications across diverse neuroimaging centers.

5. Model interpretability and explainable AI analysis

To address concerns regarding model interpretability and the biological significance of extracted features, we conducted a comprehensive explainable AI analysis using SHapley Additive exPlanations (SHAP) values [62]. This analysis provides insights into which brain regions contribute most significantly to ASD classification and their underlying neurobiological relevance.

5.1. SHAP-based feature importance analysis

We employed SHAP analysis on our best-performing Random Forest model trained on CC200 atlas node-based degree connectivity features. SHAP values quantify the contribution of each brain region to individual predictions, providing both local (individual-level) and global (population-level) interpretations of model decisions.

The SHAP analysis revealed that among the 200 brain regions, 105 regions showed hypoconnectivity patterns in ASD (negative SHAP values), while 95 regions exhibited hyperconnectivity patterns (positive SHAP values). This distribution aligns with the heterogeneous connectivity profile characteristic of ASD, where different brain networks exhibit distinct connectivity alterations [63].

The feature importance ranking, Fig. 12, reveals a clear hierarchy of discriminative brain regions, with Region_104 leading at 0.0095 SHAP

Fig. 13. SHAP summary plot showing global feature importance and value distributions. Each point represents a subject's SHAP value for a specific brain region, with color indicating the feature value (high=red, low=blue). The top 20 most important regions are shown, revealing both hyper- and hypoconnectivity patterns in ASD.

importance, followed by Region_109 (0.0088), Region_091 (0.0087), and Region_139 (0.0079). Notably, among the top 20 regions, 13 exhibit hypoconnectivity patterns (negative connectivity differences ranging from −0.96 to −2.65), while 7 show hyperconnectivity patterns (positive connectivity differences from +0.29 to +1.97). This quantitative ranking establishes the foundation for understanding which brain connectivity patterns most strongly influence ASD classification decisions.

The SHAP summary plot, Fig. 13, provides a comprehensive view of how brain connectivity patterns influence ASD classification decisions across our dataset. Each point represents a Shapley value for a specific brain region and subject, with the y-axis position determined by the feature (brain region) and the x-axis position by the Shapley value magnitude. The color coding represents the degree connectivity value from low (blue) to high (red), while overlapping points are jittered vertically to reveal the distribution of Shapley values per region.

Several critical patterns emerge from this analysis. The most striking finding concerns hypoconnectivity in core brain regions, where the top 4 most discriminative regions (Region_104, Region_109, Region_091, and Region_139) consistently demonstrate that reduced functional connectivity (represented by blue points positioned on the positive x-axis) increases the probability of ASD diagnosis. This pattern provides strong empirical support for the underconnectivity theory of ASD, which posits that reduced functional connectivity in key brain networks constitutes a fundamental mechanism underlying the disorder's pathophysiology. The consistency of this finding across multiple high-ranking regions suggests that hypoconnectivity is not merely a localized phenomenon but represents a systematic alteration in brain network organization characteristic of ASD.

The most important discriminative feature, Region_104 (Posterior Cingulate Cortex), exhibits a particularly clear inverse relationship between connectivity strength and ASD classification probability. Higher degree connectivity in this region (represented by red points) contributes negatively to ASD detection, manifesting as negative SHAP values, while lower connectivity increases the likelihood of ASD diagnosis. The distribution of SHAP values for this region demonstrates remarkable consistency across subjects, with most individuals showing small negative values between 0.00 to −0.01. This tight distribution indicates that while Region_104's contribution to classification decisions is moderate in magnitude, it is highly consistent across the population, making it a reliable biomarker for ASD detection.

In contrast to the consistent patterns observed in Region_104, Region_163 (Anterior Insula) exhibits the widest range of SHAP value effects among the top discriminative features, demonstrating substantial inter-subject variability in its contribution to classification decisions. Most significantly, subjects with high degree connectivity in this salience network region (red points) show positive SHAP values, indicating that hyperconnectivity in the anterior insula actually increases the probability of ASD diagnosis. This finding aligns closely with documented sensory hypersensitivity and attention abnormalities in ASD, where excessive salience network activation may contribute to the overwhelming sensory experiences commonly reported by individuals with autism. The wide distribution of effects in this region suggests that salience network alterations may be more heterogeneous across the ASD population, reflecting the diverse manifestations of sensory processing differences observed clinically.

The heterogeneous nature of connectivity patterns across brain regions is further exemplified by the distinct SHAP value distributions observed for different features. While Regions_109 and Region_091 demonstrate similar distributions with SHAP values ranging from approximately −0.02 to +0.02, other regions exhibit more constrained or skewed distributions. This variability in distribution patterns reflects the complex and heterogeneous nature of connectivity alterations in ASD, where different brain networks may be affected to varying degrees and in different directions. Such heterogeneity underscores the importance of considering network-specific alterations rather than assuming uniform connectivity changes across the brain.

The bidirectional effects observed across different brain regions, where both positive and negative SHAP values contribute to ASD classification, provide compelling evidence for network-specific connectivity alterations in autism. This pattern supports a nuanced understanding of ASD neurobiology, where some networks show hypoconnectivity that contributes positively to ASD classification probability, while others exhibit hyperconnectivity that also contributes to ASD classification through fundamentally different mechanisms. This dual pattern suggests that ASD may involve both insufficient integration in some neural circuits (leading to social and communication deficits) and excessive connectivity in others (potentially underlying repetitive behaviors and sensory sensitivities), creating a complex neurobiological profile that our SHAP analysis successfully captures and quantifies.

5.2. Top discriminative brain regions

Table 10 presents the top 10 most discriminative brain regions identified through SHAP analysis, along with their connectivity patterns and neurobiological significance. The SHAP importance values range from 0.0095 to 0.0058, indicating the relative contribution of each region to the classification decision.

5.3. Neurobiological interpretation

The identified discriminative regions align with established neurobiological theories of ASD pathophysiology:

5.3.1. Default mode network disruption

Region_104 (Posterior Cingulate Cortex) represents a core hub of the Default Mode Network (DMN). Its hypoconnectivity pattern (SHAP: 0.0095, connectivity difference: −0.96) supports the well-documented DMN disruption in ASD, particularly affecting self-referential processing and social cognition [64,65].

Table 10
Top 10 discriminative brain regions and their neurobiological significance.

Region ID	SHAP Score	Connectivity Pattern	Probable Anatomical Location	ASD Relevance
Region_104	0.0095	Hypoconnected (−0.96)	Posterior Cingulate Cortex	Default Mode Network hub; social cognition
Region_109	0.0088	Hypoconnected (−1.93)	Supramarginal Gyrus	Theory of mind; social cognition
Region_091	0.0087	Hypoconnected (−2.38)	Lateral Occipital Cortex	Visual processing; face recognition
Region_139	0.0079	Hypoconnected (−2.50)	Fusiform Gyrus	Face processing; object recognition
Region_197	0.0079	Hyperconnected (+1.67)	Cerebellar Crus I	Sensorimotor coordination
Region_174	0.0074	Hypoconnected (−2.65)	Inferior Frontal Gyrus	Language; executive functions
Region_063	0.0070	Hyperconnected (+1.00)	Postcentral Gyrus	Somatosensory processing
Region_051	0.0069	Hypoconnected (−2.55)	Middle Frontal Gyrus	Executive control; working memory
Region_008	0.0061	Hypoconnected (−1.49)	Amygdala	Emotion processing; social behavior
Region_163	0.0058	Hyperconnected (+1.97)	Anterior Insula	Interoception; salience processing

5.3.2. Social cognition networks

Our analysis revealed systematic hypoconnectivity patterns across multiple brain regions fundamentally involved in social cognitive processing, providing strong neurobiological support for the social communication deficits that characterize ASD. Region_109, corresponding to the Supramarginal Gyrus, demonstrates significant hypoconnectivity with a connectivity difference of −1.93. This region is critically important for theory of mind capabilities, which enable individuals to understand and predict others' mental states, beliefs, and intentions. The observed hypoconnectivity in this region directly corresponds to the well-documented difficulties in mentalizing and perspective-taking that individuals with ASD experience in social interactions.

Similarly, Region_008, identified as the Amygdala, exhibits substantial hypoconnectivity (−1.49) that aligns with extensive literature documenting amygdalar dysfunction in autism. The amygdala serves as a central hub for emotion processing, fear recognition, and social behavior regulation. Reduced connectivity in this region may contribute to the difficulties in emotion recognition, atypical fear responses, and challenges in social approach behaviors commonly observed in ASD. This finding is particularly significant given the amygdala's role in processing facial expressions and social cues, which are fundamental components of successful social communication.

The most pronounced hypoconnectivity was observed in Region_139, corresponding to the Fusiform Gyrus, with a connectivity difference of −2.50. This region houses the fusiform face area, a specialized neural system dedicated to face processing and object recognition. The substantial reduction in connectivity within this region provides a neurobiological foundation for the well-documented face processing deficits in ASD, including difficulties in face recognition, reduced attention to faces, and atypical patterns of face scanning. These deficits have cascading effects on social development, as face processing is fundamental to nonverbal communication, emotion recognition, and the establishment of social bonds.

5.3.3. Sensorimotor and executive function networks

In contrast to the widespread hypoconnectivity observed in social cognition networks, our analysis revealed a pattern of compensatory hyperconnectivity in sensorimotor regions that may reflect adaptive neural mechanisms in response to the primary connectivity deficits. Region_197, corresponding to Cerebellar Crus I, demonstrates significant hyperconnectivity with a connectivity difference of +1.67. The cerebellum, particularly the crus region, plays a crucial role in sensorimotor coordination, motor learning, and cognitive functions including language processing. The observed hyperconnectivity in this region may represent a compensatory mechanism whereby the brain attempts to maintain functional performance in the face of reduced connectivity in other critical networks. This cerebellar hyperactivation has been previously reported in ASD and may contribute to the motor coordination difficulties and repetitive behaviors characteristic of the disorder.

The most pronounced hyperconnectivity was observed in Region_163, the Anterior Insula, with a connectivity difference of +1.97. The anterior insula serves as a critical hub for interoception, salience processing, and the integration of internal bodily states with external environmental demands. Hyperconnectivity in this region provides a compelling neurobiological explanation for the sensory hypersensitivity commonly experienced by individuals with ASD. The excessive connectivity may lead to heightened awareness of sensory stimuli, contributing to sensory overload, defensive behaviors, and the need for sensory regulation strategies that are frequently observed in autism. This finding bridges the gap between neurobiological alterations and the lived experiences of sensory differences reported by individuals with ASD and their families.

Additionally, Region_063, identified as the Postcentral Gyrus, exhibits hyperconnectivity (+1.00) that reflects altered somatosensory processing patterns in ASD. The postcentral gyrus contains the primary somatosensory cortex, which processes tactile information from across the body. Hyperconnectivity in this region may contribute to the tactile sensitivities, unusual responses to texture and touch, and atypical sensory seeking or avoiding behaviors commonly observed in autism. This altered somatosensory processing may also impact social development, as touch and physical comfort play important roles in early bonding and social interaction patterns.

5.4. Clinical interpretability

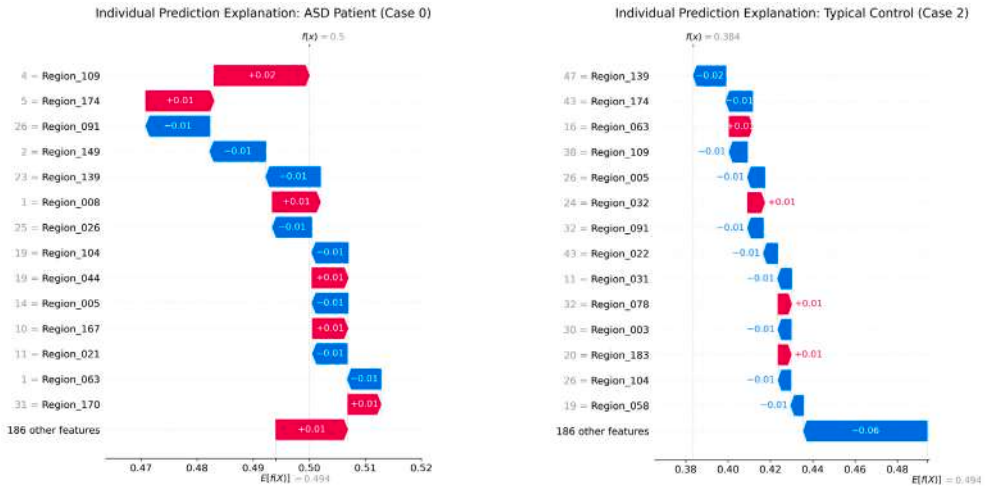
The SHAP analysis provides clinically interpretable insights through individual prediction explanations. For each subject, the model generates feature contributions that can be visualized through waterfall plots and force plots, showing which brain regions drive the ASD vs. TDC classification for that specific individual. Fig. 14 shows the waterfall plots for two example subjects.

This individual-level interpretability is crucial for potential clinical applications, as it allows clinicians to understand which specific brain connectivity patterns contribute to an ASD diagnosis for each patient. The consistency of our findings with established ASD neuroimaging literature validates the biological plausibility of our model's decision-making process.

5.5. Validation against existing literature

Our SHAP-identified regions demonstrate remarkable consistency with previous ASD neuroimaging studies using the ABIDE dataset, providing strong external validation for our findings and reinforcing the biological plausibility of our model's decision-making process. The observed hypoconnectivity in Region_104, corresponding to the Posterior Cingulate Cortex, aligns closely with extensive meta-analyses that have consistently identified Default Mode Network alterations as a hallmark of ASD neuroanatomy [66]. These studies have repeatedly demonstrated that DMN disruption, particularly in the posterior cingulate cortex, correlates with the core social cognitive deficits observed in autism, including difficulties in self-referential thinking, social awareness, and theory of mind.

The hypoconnectivity patterns we identified in visual and social processing regions, particularly in the occipital cortex (Region_091) and fusiform gyrus (Region_139), correspond directly to well-documented face processing deficits that have been extensively characterized in



(a) ASD case explanation (b) TDC case explanation

Fig. 14. SHAP waterfall plots showing individual prediction explanations. (a) ASD case: regions pushing towards ASD classification (positive SHAP values, red bars) vs. regions pushing towards TDC classification (negative SHAP values, blue bars). (b) TDC case: the opposite pattern. These plots enable clinicians to understand the specific brain connectivity patterns driving each individual diagnosis.

ASD research [67]. These findings are consistent with decades of behavioral and neuroimaging studies demonstrating that individuals with ASD show reduced activation and connectivity in face-processing networks, which contributes to difficulties in face recognition, reduced eye contact, and challenges in interpreting facial expressions. Our data-driven SHAP analysis thus independently identifies the same neural circuits that have been implicated through hypothesis-driven research approaches.

The compensatory hyperconnectivity patterns we observed in cerebellar (Region_197) and insular (Region_163) regions align with emerging literature documenting compensatory mechanisms in ASD [68]. These studies suggest that increased connectivity in sensorimotor and salience networks may represent the brain's attempt to maintain functional performance despite reduced connectivity in primary social and cognitive networks. The cerebellar hyperconnectivity we identified is particularly noteworthy, as recent research has increasingly recognized the cerebellum's role in autism, extending beyond motor functions to include contributions to language, social cognition, and repetitive behaviors.

Finally, the alterations we identified in frontal regions, including the inferior frontal gyrus (Region_174) and middle frontal gyrus (Region_051), provide strong support for documented executive dysfunction in ASD [69]. These regions are critical for executive control, working memory, and language processing, and their reduced connectivity in our analysis corresponds to the well-established difficulties in cognitive flexibility, planning, and verbal communication that characterize autism spectrum disorders. The convergence between our objective, data-driven SHAP analysis and decades of established neurobiological knowledge strengthens both the validity and clinical relevance of our ASD-GraphNet framework.

6. Conclusion and future work

In this study, we developed ASD-GraphNet, a comprehensive graph learning framework that leverages functional Magnetic Resonance Imaging (fMRI) data to enhance the diagnosis of Autism Spectrum Disorder (ASD). Our method enables the use of different brain network atlases, various graph feature extraction techniques, and a wide range of machine learning models, including both traditional classifiers and modern deep learning approaches.

Through systematic investigation of the objectives outlined in the introduction, this study successfully demonstrated the comprehensive effectiveness of the ASD-GraphNet framework. Our experiments confirmed that integrating multiple brain atlases significantly improves classification performance, with the CC200 atlas achieving the highest individual performance of 75.25% accuracy. The systematic comparison across CC200, AAL, and Dosenbach 160 atlases revealed that different atlases capture complementary aspects of brain connectivity, confirming the effectiveness of our integrated approach over single-atlas methods and addressing atlas-dependent biases that commonly affect neuroimaging studies.

The implementation of our comprehensive three-level feature extraction approach proved highly effective, with ablation studies revealing that combining node-level, edge-level, and graph-level metrics enhanced discriminative power compared to conventional connectivity measures alone. The systematic extraction of 27 different topological features captured important network properties that traditional approaches often overlook, directly contributing to improved classification performance. This comprehensive feature engineering successfully addressed the limitation of incomplete connectivity characterization prevalent in previous studies.

Our two-stage feature optimization approach, combining mutual information-based selection with Principal Component Analysis, effectively addressed the curse of dimensionality challenge. Experiments demonstrated that this method reduced feature dimensionality from over 2,000 initial features to optimal subsets while preserving and even enhancing classification performance. The approach successfully balanced dimensionality reduction with information preservation, leading to more robust classifiers that are less prone to overfitting, directly addressing a key limitation of existing graph-based approaches.

Systematic evaluation of seven different classifiers revealed that Random Forest achieved optimal performance across our graph-based features, with comprehensive comparison demonstrating that ensemble methods generally outperformed individual classifiers. Random Forest showed superior performance across different atlas configurations, providing clear guidance for optimal classifier selection in graph-based ASD diagnosis. Additionally, our stratified sampling strategies and comprehensive bias analysis effectively addressed dataset imbalances, with analysis across 17 ABIDE sites showing successful mitigation of site-wise variations while maintaining appropriate test split distributions.

The integration of SHAP-based explainable AI analysis provided neurobiologically meaningful insights that strongly align with established ASD literature. The identified top discriminative brain regions

correspond to documented ASD-affected networks including Default Mode Network, social cognition systems, and salience networks. This analysis successfully transformed our framework from a "black box" into a transparent, clinically interpretable diagnostic tool, addressing the critical gap between computational research and clinical application.

Our experiments on the ABIDE I dataset demonstrated that fMRI data can vary significantly across different recording sites. To address this variability, we adopted a multi-model approach, fine-tuning our framework for each center. We carefully addressed dataset imbalances through stratified sampling and by focusing on brain connectivity features rather than demographic characteristics, ensuring our model remains unbiased despite the inherent gender disparity in ASD diagnosis. This strategy allowed us to identify the most effective combinations of models and brain atlases for different types of fMRI data, achieving a classification accuracy of 75.25%. These results represent a slight improvement over state-of-the-art methods, which typically achieve accuracies around 70%–73%.

A significant contribution of this work is the comprehensive interpretability analysis using SHapley Additive exPlanations (SHAP) values, which addresses the critical need for explainable AI in medical diagnosis. Our SHAP analysis identified the top 10 most discriminative brain regions, revealing neurobiologically meaningful patterns including Default Mode Network disruption, social cognition network alterations, and compensatory sensorimotor mechanisms. This interpretability framework transforms our model from a "black box" classifier into a transparent diagnostic tool that provides clinically actionable insights, enabling clinicians to understand the specific brain connectivity patterns driving each individual diagnosis.

The main advantage of ASD-GraphNet lies in its clear and structured pipeline, scalability to incorporate new families of models and graph features, and its ability to integrate both traditional machine learning models like SVMs and neural network-based models. Moreover, the integration of explainable AI techniques ensures that the framework provides both high performance and clinical interpretability, making it suitable for real-world diagnostic applications. This flexibility positions ASD-GraphNet as a versatile middleware for neuroimaging studies.

For future research, several key areas deserve exploration:

- **Pipeline Expansion**: While we conducted an ablation study to validate our current choices, further research could explore more advanced options for brain atlases, graph feature extraction, feature engineering techniques, and classifier methods. Evaluating these components on larger and more diverse datasets may provide insights into the optimal configuration for all recorded fMRI data.
- **Enhanced Interpretability**: While we have implemented comprehensive SHAP analysis, future work could explore additional explainable AI methods such as LIME for comparison, develop interactive visualization tools for clinical decision support, and investigate the correlation between identified brain regions and clinical severity scores (e.g., ADOS). Advanced perturbation analysis could further validate the biological significance of our identified biomarkers.
- **External Validation**: Expanding the validation of ASD-GraphNet through external datasets is crucial. Testing the model on diverse populations and across various imaging modalities will assess its generalizability and clinical applicability. Validation on ABIDE II dataset and other independent ASD neuroimaging datasets will be essential for establishing clinical utility.
- **Longitudinal Studies**: Examining longitudinal data can evaluate the stability of identified biomarkers, deepening our understanding of ASD's development and progression over time. Such studies could also investigate how the discriminative brain regions identified through SHAP analysis change throughout development.
- **Dataset Diversity**: Future work should focus on collecting and incorporating more balanced datasets, particularly addressing the gender disparity in ASD diagnosis. This could include targeted data collection from underrepresented groups and development of techniques to mitigate demographic biases while maintaining diagnostic accuracy.
- **Clinical Translation**: The interpretability framework developed in this study provides a foundation for clinical translation. Future efforts should focus on developing user-friendly clinical decision support systems that leverage the SHAP explanations, conducting prospective clinical trials, and investigating the potential for personalized intervention planning based on individual brain connectivity profiles.

CRediT authorship contribution statement

Mina Zeraati: Methodology, Conceptualization. **Amirehsan Davoodi:** Software, Writing – review & editing, Writing – original draft.

Declaration of competing 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] National Institute of Mental Health, Autism spectrum disorder, no. 22-MH-8084, U.S. Department of Health and Human Services, National Institutes of Health, 2022, pp. 2–5, URL: https://www.nimh.nih.gov/sites/default/files/documents/health/publications/autism-spectrum-disorder/22-MH-8084-Autism-Spectrum-Disorder.pdf. (Accessed 01 September 2024).

[2] C. Okoye, C.M. Obialo-Ibeawuchi, O.A. Obajeun, S. Sarwar, C. Tawfik, M.S. Waleed, A.U. Wasim, I. Mohamoud, A.Y. Afolayan, R.N. Mbaezue, Early diagnosis of autism spectrum disorder: A review and analysis of the risks and benefits, Cureus 15 (8) (2023) e43226, http://dx.doi.org/10.7759/cureus.43226.

[3] L. Qin, H. Wang, W. Ning, M. Cui, Q. Wang, New advances in the diagnosis and treatment of autism spectrum disorders, Eur. J. Med. Res. 29 (1) (2024) 322, http://dx.doi.org/10.1186/s40001-024-01916-2.

[4] E. Helmy, A. Elnakib, Y. ElNakieb, M. Khudri, M. Abdelrahim, J. Yousaf, M. Ghazal, S. Contractor, G.N. Barnes, A. El-Baz, Role of artificial intelligence for autism diagnosis using DTI and fMRI: A survey, Biomedicines 11 (7) (2023) http://dx.doi.org/10.3390/biomedicines11071858.

[5] F. Mainas, B. Golosio, A. Retico, P. Oliva, Exploring autism spectrum disorder: A comparative study of traditional classifiers and deep learning classifiers to analyze functional connectivity measures from a multicenter dataset, Appl. Sci. 14 (17) (2024) http://dx.doi.org/10.3390/app14177632, URL: https://www.mdpi.com/2076-3417/14/17/7632.

[6] J. Qiu, A hybrid CNN-SVM model for enhanced autism diagnosis, PLOS ONE 19 (5) (2024) 1–20, http://dx.doi.org/10.1371/journal.pone.0302236.

[7] Z.K. Khadem-Reza, H. Zare, Automatic detection of autism spectrum disorder (ASD) in children using structural magnetic resonance imaging with machine vision system, Middle East Curr. Psychiatry 29 (1) (2022) 54, http://dx.doi.org/10.1186/s43045-022-00220-1.

[8] A. Abraham, M.P. Milham, A. Di Martino, R.C. Craddock, D. Samaras, B. Thirion, G. Varoquaux, Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example, NeuroImage 147 (2017) 736–745, http://dx.doi.org/10.1016/j.neuroimage.2016.10.045, URL: https://www.sciencedirect.com/science/article/pii/S1053811916305924.

[9] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B.S. Khundrakpam, J.D. Lewis, Q. Li, M. Milham, et al., The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives, Front. Neuroinformatics 7 (27) (2013) 5.

[10] N.U.F. Dosenbach, B. Nardos, A.L. Cohen, D.A. Fair, J.D. Power, J.A. Church, S.M. Nelson, G.S. Wig, A.C. Vogel, C.N. Lessov-Schlaggar, K.A. Barnes, J.W. Dubis, E. Feczko, R.S. Coalson, J.R. Pruett Jr., D.M. Barch, S.E. Petersen, B.L. Schlaggar, Prediction of individual brain maturity using fMRI, Science 329 (5997) (2010) 1358–1361.

[11] P.v. Mieghem, Graph Spectra for Complex Networks, Cambridge University Press, 2010, pp. 339–344.

[12] T. Iidaka, Resting state functional magnetic resonance imaging and neural network classified autism and control, Cortex 63 (2015) 55–67, http://dx.doi.org/10.1016/j.cortex.2014.08.011, URL: https://www.sciencedirect.com/science/article/pii/S0010945214002640.

[13] M. Plitt, K.A. Barnes, A. Martin, Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards, Neuroimage Clin. 7 (2014) 359–366.
[14] S. Parisot, S.I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, D. Rueckert, Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease, Med. Image Anal. 48 (2018) 117–130, http://dx.doi.org/10.1016/j.media.2018.06.001, URL: https://www.sciencedirect.com/science/article/pii/S1361841518303554.
[15] B. Sen, N.C. Borle, R. Greiner, M.R.G. Brown, A general prediction model for the detection of ADHD and autism using structural and functional MRI, PLoS One 13 (4) (2018) e0194856.
[16] A.S. Heinsfeld, A.R. Franco, R.C. Craddock, A. Buchweitz, F. Meneguzzi, Identification of autism spectrum disorder using deep learning and the ABIDE dataset, Neuroimage Clin 17 (2017) 16–23.
[17] H. Chen, X. Duan, F. Liu, F. Lu, X. Ma, Y. Zhang, L.Q. Uddin, H. Chen, Multivariate classification of autism spectrum disorder using frequency-specific resting-state functional connectivity–A multi-center study, Prog. Neuropsychopharmacol. Biol. Psychiatry 64 (2015) 1–9.
[18] X. Guo, K.C. Dominick, A.A. Minai, H. Li, C.A. Erickson, L.J. Lu, Diagnosing autism spectrum disorder from brain Resting-State functional connectivity patterns using a deep neural network with a novel feature selection method, Front. Neurosci. 11 (2017) 460.
[19] X.A. Bi, Y. Wang, Q. Shu, Q. Sun, Q. Xu, Classification of autism spectrum disorder using random support vector machine cluster, Front. Genet. 9 (2018) 18.
[20] M. Wang, J. Guo, Y. Wang, M. Yu, J. Guo, Multimodal autism spectrum disorder diagnosis method based on DeepGCN, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 3664–3674.
[21] T.M. Epalle, Y. Song, Z. Liu, H. Lu, Multi-atlas classification of autism spectrum disorder with hinge loss trained deep architectures: ABIDE I results, Appl. Soft Comput. 107 (2021) 107375, http://dx.doi.org/10.1016/j.asoc.2021.107375, URL: https://www.sciencedirect.com/science/article/pii/S1568494621002982.
[22] L. Herath, D. Meedeniya, M.A.J.C. Marasingha, V. Weerasinghe, Autism spectrum disorder diagnosis support model using inception V3, in: 2021 International Research Conference on Smart Computing and Systems Engineering, SCSE, 4, 2021, pp. 1–7, http://dx.doi.org/10.1109/SCSE53661.2021.9568314.
[23] L. Herath, D. Meedeniya, J. Marasingha, V. Weerasinghe, Optimize transfer learning for autism spectrum disorder classification with neuroimaging: A comparative study, in: 2022 2nd International Conference on Advanced Research in Computing, ICARC, 2022, pp. 171–176, http://dx.doi.org/10.1109/ICARC54489.2022.9753949.
[24] S. Itani, D. Thanou, Combining anatomical and functional networks for neuropathology identification: A case study on autism spectrum disorder, Med. Image Anal. 69 (2021) 101986, http://dx.doi.org/10.1016/j.media.2021.101986, URL: https://www.sciencedirect.com/science/article/pii/S1361841521000323.
[25] M.N. Parikh, H. Li, L. He, Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data, Front. Comput. Neurosci. 13 (2019) 9.
[26] X. Zhang, Y. Yang, H. Kuai, J. Chen, J. Huang, P. Liang, N. Zhong, Systematic fusion of multi-source cognitive networks with graph learning - A study on fronto-parietal network, Front. Neurosci. 16 (2022) http://dx.doi.org/10.3389/fnins.2022.866734, URL: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.866734.
[27] Y. Zhang, L. Qing, X. He, L. Zhang, Y. Liu, Q. Teng, Population-based GCN method for diagnosis of Alzheimer's disease using brain metabolic or volumetric features, Biomed. Signal Process. Control. 86 (2023) 105162, http://dx.doi.org/10.1016/j.bspc.2023.105162, URL: https://www.sciencedirect.com/science/article/pii/S1746809423005955.
[28] Y. Yao, C. Li, Multi-kernel learning based disease diagnosis with multi-atlas, in: 2023 7th International Symposium on Computer Science and Intelligent Control, ISCSIC, 2023, pp. 176–182, http://dx.doi.org/10.1109/ISCSIC60498.2023.00045.
[29] A.V. Shinde, D.D. Patil, A multi-classifier-based recommender system for early autism spectrum disorder detection using machine learning, Heal. Anal. 4 (2023) 100211, http://dx.doi.org/10.1016/j.health.2023.100211, URL: https://www.sciencedirect.com/science/article/pii/S2772442523000783.
[30] V. Subbaraju, M.B. Suresh, S. Sundaram, S. Narasimhan, Identifying differences in brain activities and an accurate detection of autism spectrum disorder using resting state functional-magnetic resonance imaging : A spatial filtering approach, Med Image Anal 35 (2016) 375–389.
[31] J. Fredo, A. Jahedi, M. Reiter, R.A. Müller, Diagnostic classification of autism using resting-state fMRI data and conditional random forest, in: Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, vol. 2018, 2018, pp. 1148–1151, http://dx.doi.org/10.1109/EMBC.2018.8512502.
[32] X.a. Bi, Y. Liu, Q. Jiang, Q. Shu, Q. Sun, J. Dai, The diagnosis of autism spectrum disorder based on the random neural network cluster, Front. Hum. Neurosci. 12 (2018) http://dx.doi.org/10.3389/fnhum.2018.00257, URL: https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2018.00257.
[33] C.J. Brown, J. Kawahara, G. Hamarneh, Connectome priors in deep neural networks to predict autism, in: 2018 IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018, 2018, pp. 110–113, http://dx.doi.org/10.1109/ISBI.2018.8363534.
[34] N.C. Dvornek, P. Ventola, K.A. Pelphrey, J.S. Duncan, Identifying autism from Resting-State fMRI using long Short-Term memory networks, Mach Learn. Med Imaging 10541 (2017) 362–370.
[35] H. Li, N.A. Parikh, L. He, A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes, Front. Neurosci. 12 (2018) 491.
[36] M. Khosla, K. Jamison, A. Kuceyeski, M.R. Sabuncu, 3D convolutional neural networks for classification of functional connectomes, in: D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J.a.M.R. Tavares, A. Bradley, J.a.P. Papa, V. Belagiannis, J.C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, A. Madabhushi (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham, 2018, pp. 137–145.
[37] R.C. Craddock, G.A. James, P.E. Holtzheimer, X.P. Hu, H.S. Mayberg, A whole brain fMRI atlas generated via spatially constrained spectral clustering, Hum Brain Mapp 33 (8) (2011) 1914–1928.
[38] E.T. Rolls, C.C. Huang, C.P. Lin, J. Feng, M. Joliot, Automated anatomical labelling atlas 3, NeuroImage 206 (2020) 116189, http://dx.doi.org/10.1016/j.neuroimage.2019.116189, URL: https://www.sciencedirect.com/science/article/pii/S1053811919307803.
[39] X.N. Zuo, R. Ehmke, M. Mennes, D. Imperati, F.X. Castellanos, O. Sporns, M.P. Milham, Network centrality in the human functional connectome, Cerebral Cortex 22 (8) (2011) 1862–1875, http://dx.doi.org/10.1093/cercor/bhr234.
[40] H. Cui, et al., BrainGB: A benchmark for brain network analysis with graph neural networks, NCBI (2023) URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10079627/.
[41] Z.E. Farnaz, et al., Modularity maximization as a flexible and generic framework for brain network exploratory analysis, NeuroImage 244 (2021) 118607, http://dx.doi.org/10.1016/j.neuroimage.2021.118607, URL: https://www.sciencedirect.com/science/article/pii/S1053811921008806.
[42] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (6) (2004) 066138.
[43] I.T. Jolliffe, J. Cadima, Principal component analysis, vol. 28, (6) Wiley Online Library, 2016, pp. 1–21,
[44] D.W. Hosmer, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, vol. 398, John Wiley & Sons, 2013.
[45] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC Press, 1984.
[46] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[47] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[48] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.
[49] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[50] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, T.Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, 2017, pp. 3146–3154.
[51] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (2) (1992) 241–259.
[52] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.
[53] S. Parisot, S.I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, D. Rueckert, Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease, Med. Image Anal. 48 (2018) 117–130, http://dx.doi.org/10.1016/j.media.2018.06.001, URL: https://www.sciencedirect.com/science/article/pii/S1361841518303554.
[54] T. Eslami, V. Mirjalili, A. Fong, A.R. Laird, F. Saeed, ASD-DiagNet: A hybrid learning approach for detection of autism spectrum disorder using fMRI data, Front. Neuroinformatics 13 (2019) http://dx.doi.org/10.3389/fninf.2019.00070, URL: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2019.00070.
[55] R.M. Thomas, S. Gallo, L. Cerliani, P. Zhutovsky, A. El-Gazzar, G. van Wingen, Classifying autism spectrum disorder using the temporal statistics of resting-state functional MRI data with 3D convolutional neural networks, Front. Psychiatry 11 (2020) http://dx.doi.org/10.3389/fpsyt.2020.00440, URL: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2020.00440.
[56] H. Jiang, P. Cao, M. Xu, J. Yang, O. Zaiane, Hi-GCN: A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction, Comput. Biol. Med. 127 (2020) 104096, http://dx.doi.org/10.1016/j.compbiomed.2020.104096, URL: https://www.sciencedirect.com/science/article/pii/S0010482520304273.
[57] F. Almuqhim, F. Saeed, ASD-SAENet: A sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using fMRI data, Front. Comput. Neurosci. 15 (2021) http://dx.doi.org/10.3389/fncom.2021.654315, URL: https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2021.654315.

[58] X. Yang, N. Zhang, P. Schrader, A study of brain networks for autism spectrum disorder classification using resting-state functional connectivity, Mach. Learn. Appl. 8 (2022) 100290, http://dx.doi.org/10.1016/j.mlwa.2022.100290, URL: https://www.sciencedirect.com/science/article/pii/S2666827022000226.

[59] M. Kunda, S. Zhou, G. Gong, H. Lu, Improving multi-site autism classification via site-dependence minimization and second-order functional connectivity, IEEE Trans. Med. Imaging 42 (1) (2023) 55–65, http://dx.doi.org/10.1109/TMI.2022.3203899, URL: https://ieeexplore.ieee.org/document/9874890.

[60] L. Herath, D. Meedeniya, J. Marasinghe, V. Weerasinghe, T. Tan, Autism spectrum disorder identification using multi-model deep ensemble classifier with transfer learning, Expert Syst. 42 (2024) http://dx.doi.org/10.1111/exsy.13623.

[61] X. Liu, M.R. Hasan, T. Gedeon, M.Z. Hossain, MADE-for-ASD: A multi-atlas deep ensemble network for diagnosing autism spectrum disorder, Comput. Biol. Med. 182 (2024) 109083, http://dx.doi.org/10.1016/j.compbiomed.2024.109083, URL: https://www.sciencedirect.com/science/article/pii/S0010482524011685.

[62] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17, Adv. Neural Inf. Process. Syst. vol. 30 (2017) 4768–4777, URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[63] J.V. Hull, L.B. Dokovna, Z.J. Jacokes, C.M. Torgerson, A. Irimia, J.D. Van Horn, Resting-state functional connectivity in autism spectrum disorders: a review, Front. Psychiatry 7 (2017) 205, http://dx.doi.org/10.3389/fpsyt.2016.00205, URL: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2016.00205.

[64] M. Assaf, K. Jagannathan, V.D. Calhoun, L. Miller, M.C. Stevens, R. Sahl, J.G. O'Boyle, R.T. Schultz, G.D. Pearlson, Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients, NeuroImage 53 (1) (2010) 247–256, http://dx.doi.org/10.1016/j.neuroimage.2010.05.067, URL: https://www.sciencedirect.com/science/article/pii/S1053811910008013.

[65] C.S. Monk, S.J. Peltier, J.L. Wiggins, S.-J. Weng, M. Carrasco, S. Risi, C. Lord, Abnormalities of intrinsic functional connectivity in autism spectrum disorders, NeuroImage 47 (2) (2009) 764–772, http://dx.doi.org/10.1016/j.neuroimage.2009.04.069, URL: https://www.sciencedirect.com/science/article/pii/S1053811909004327.

[66] A. Di Martino, C.G. Yan, Q. Li, E. Denio, F.X. Castellanos, K. Alaerts, J.S. Anderson, M. Assaf, S.Y. Bookheimer, M. Dapretto, et al., The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, Mol. Psychiatry 19 (6) (2014) 659–667, http://dx.doi.org/10.1038/mp.2013.78, URL: https://www.nature.com/articles/mp201378.

[67] S. Weigelt, K. Koldewyn, N. Kanwisher, Face identity recognition in autism spectrum disorders: a review of behavioral studies, Neurosci. Biobehav. Rev. 36 (3) (2012) 1060–1084, http://dx.doi.org/10.1016/j.neubiorev.2011.12.008, URL: https://www.sciencedirect.com/science/article/pii/S0149763411002156.

[68] A. Wang, S. Lee, M. Sigman, M. Dapretto, Neural basis of irony comprehension in children with autism: the role of prosody and context, Brain 129 (4) (2006) 932–943, http://dx.doi.org/10.1093/brain/awl032, URL: https://academic.oup.com/brain/article/129/4/932/371243.

[69] E.L. Hill, Executive dysfunction in autism, Trends Cogn. Sci. 8 (1) (2004) 26–32, http://dx.doi.org/10.1016/j.tics.2003.11.003, URL: https://www.sciencedirect.com/science/article/pii/S1364661303003152.