Cross-Attention Between Functional and Structural MRI for Autism Classification

Stanisław Wasilewski[1]

BSc Artificial Intelligence, Vrije Universiteit Amsterdam
Supervisor: Bob Borsboom
Student number: 2763732 • Submitted: 6 July 2025
wasilewski.sf@gmail.com
https://github.com/stasieniec/cross-attention-abide

Abstract. Autism Spectrum Disorder (ASD) affects 1 in 127 people worldwide, yet current diagnostic methods rely on behavioral assessments that are time-consuming and prone to bias. While machine learning approaches using neuroimaging data show promise, many existing models suffer from poor generalizability and evaluation methodologies that may inflate performance estimates. This thesis presents the first systematic investigation of bidirectional cross-attention between functional and structural MRI for ASD classification, exploring how different tokenization strategies affect model performance and generalizability.

This study develops eleven Transformer-based architectures that implement cross-attention mechanisms between fMRI connectivity patterns and sMRI structural features, using various tokenization approaches including ROI-based and network-based strategies. All models were evaluated using both standard k-fold cross-validation and leave-one-out cross-validation on 871 subjects from the ABIDE dataset to assess true generalizability across acquisition sites.

Results demonstrate that cross-attention consistently outperforms single-modality baselines, achieving 69.9% accuracy compared to 63.5% for fMRI-only models. However, complex tokenization strategies that improved performance under standard evaluation (70.1% peak accuracy) showed limited benefits under leave-one-out validation, revealing significant site bias effects that inflate traditional performance estimates. None of the approaches were able to achieve accuracy beyond a 70% performance ceiling, suggesting fundamental limitations in current neuroimaging-based ASD classification.

The findings highlight the importance of rigorous cross-site evaluation in neuroimaging research and demonstrate that while multimodal integration provides meaningful improvements, current approaches may be approaching their discriminatory limits for ASD classification.

Keywords: Autism Spectrum Disorder · Cross-attention · Multimodal fusion · Neuroimaging · Leave-one-out cross-validation.

1 Introduction

Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental condition affecting approximately 1 in every 127 people worldwide [25]. Characterized by social and communication impairments, along with distinctive behavioral patterns and sensory sensitivities [26], ASD profoundly impacts individuals' daily functioning and overall well-being. Current diagnostic approaches rely entirely on behavioral assessments, and are not free from errors, biases, and stigma. For example, women are diagnosed on average 8 years later than men [12]. This disparity is concerning, as late diagnosis is associated with poorer long-term quality of life [3].

Although structural, functional, and molecular differences have been identified in the brains of autistic individuals, a definitive neurobiological basis of ASD has not been established - no single brain characteristic could be used for clinical diagnosis [15]. Therefore, current diagnosis is based on a series of behavioral assessments, which are often associated with long wait times, high costs, and lack of clear clinical pathways for those who undergo it [20].

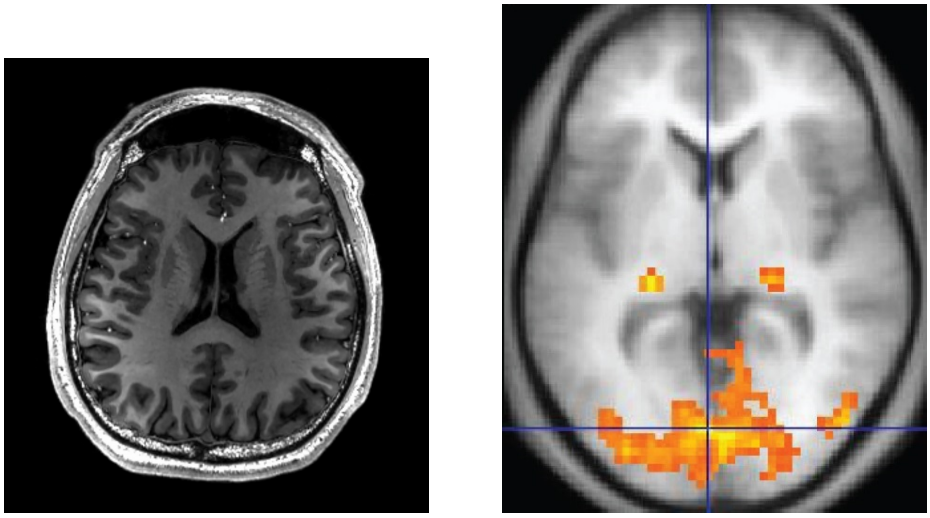
1.1 Promise of Machine Learning

Despite the lack of obvious neurobiological biomarkers for the identification of ASD, machine learning has shown potential to predict the diagnosis based on neuroimaging data. Resting-state functional magnetic resonance imaging brain scans (fMRI), which contain information on brain region activation, reveal functional connectivity patterns that appear to be more distinctly indicative of ASD [15]. However, structural magnetic resonance imaging brain scans (sMRI), which represent the anatomical structure of the brain, can also be used for ASD classification, with ever-so-slightly worse accuracy in single-modality settings [9]. Containing complementary information, which can be better conceptualized with the help of the two brain slice visualizations in Figure 1, models that use fMRI and sMRI features perform marginally better than single-modality models, even with simple feature fusion techniques [9]

By introducing the ABIDE I and ABIDE II datasets, the ABIDE initiative, a collaboration of 17 hospitals collecting neuroimaging data for ASD research [8], enabled the implementation of more "data-hungry" deep learning models in the ASD classification domain. Reporting accuracy as high as 85% [23], these models quickly proved their effectiveness.

However, in a recent study, Dong et al. [9] evaluated the most prominent of these models, to discover that their reported accuracy crumbles under more rigorous leave-one-out cross-validation - for example, 85% accuracy reported by Rakić et al. [23], even though evaluated with 10-fold cross-validation, shrinks down to 66.80% under experimental conditions that completely prevent site bias. Dong et al. established the performance of approximately 70% to be the upper limit of state-of-the-art models in this domain.

Next to heterogeneity of ASD and site biases in ABIDE data, another limitation of deep learning approaches is the black-box problem - interpretability



(a) T1-weighted sMRI slice **(b)** BOLD fMRI illustration

Fig. 1: Structural (a) and functional (b) brain images. (a) © Asnaebsa (2022), CC BY-SA 4.0 [2]. (b) © OpenStax (2016), CC BY 4.0 [21].

carries great value, especially in diagnostic tasks, and more explainable, simpler models with comparable performance are usually more valuable in clinical environments.

1.2 Transformer-based Models

Explainability is a strong reason for the promising appearance of Transformer-based architectures in ASD classification - attention mechanisms can often be tracked, allowing visualization of the most attended-to data features - features, which these models ground their decisions on and which could provide insights into eventually obtaining objective neurobiological biomarkers for ASD.

Existing Transformer-based models for ASD classification report high accuracies, with METAFormer reporting 85% thanks to masked pertaining phase [19]. However, none of these models use leave-one-out cross-validation, thus making it difficult to compare their effectiveness with other models or to evaluate them with a clinical perspective in mind.

1.3 Research Contribution and Questions

Although some of the existing Transformer models use multiple fMRI data derivatives as separate modalities, cross-attention between sMRI and fMRI has not yet been explicitly explored. Moreover, many of the existing multimodal

architectures use early or late fusion, with more complex settings remaining unexplored.

The goal of this thesis is to address this research gap by performing a series of experiments involving cross-attention between fMRI and sMRI, utilizing different tokenization strategies that are evaluated with leave-one-out cross-validation. The results should yield answers to the following two research questions:

- How does cross-attention between functional and structural MRI affect accuracy and generalizability on ASD classification, compared to single-modality models with comparable Transformer-based architecture?
- How do different tokenization techniques affect single-modality and multimodal cross-attention Transformer-based models in accuracy and generalizability for ASD classification?

2 Related Work

This section provides an overview of scientific achievements in the domain of ASD classification, with an emphasis on AI-based solutions, such as machine learning and deep learning.

2.1 Evolution of ASD Classification

During the last decade, advances in the field of machine learning have contributed significantly to the automatic classification of ASD. The availability of large datasets, such as ABIDE I and ABIDE II [8], enabled the application of data-demanding models in this domain.

Early Machine Learning Approaches Initial attempts at machine learning-based ASD classification faced challenges due to data limitations. Small datasets involving specific, often demographically homogeneous cohorts did not allow models to capture particular ASD characteristics encoded in the varying spectrum of observable differences, thus hindering generalizability [29]. For example, Ecker et al. used an SVM and a VBM for ASD classification, achieving 81% accuracy on a dataset of 44 subjects [11]. As outlined in later sections of this thesis, the pattern of high accuracy with limited generalizability is prevalent even on larger datasets and remains a challenge to this day.

ABIDE I, comprising 539 individuals with ASD and 573 neurotypical controls, established a new, universal standard for later research [8]. Sabuncu et al., who evaluated multiple classifiers on 650 instances from ABIDE I in 2015 and achieved an accuracy of 60%, confirmed the generalizability issues of previous approaches [24]. Two years later, Abraham et al. attained an accuracy of 67%, also on ABIDE I, exploring the importance of the feature selection process for fMRI [1]. In 2019, Parikh et al. evaluated nine machine learning models on 851 instances, achieving a best AUC of 64.6% [22]. The difference in results underscores two central issues in the ABIDE I dataset and in the process of

evaluation ASD classification in general: underlying heterogeneity in ASD, as well as biases introduced by different sites and measuring equipment, which contribute to variance between evaluation methods (e.g. cross-validation, stratified cross-validation, leave-site-out-cross validation) [29].

Deep Learning Approaches The introduction of deep learning techniques to the domain of ASD classification has quickly proven its potential, both for the classification task itself and for feature selection. Dvornek et al. [10] demonstrated that Long-Short-Term Memory (LSTM) network applied to raw fMRI time-series achieved 68% accuracy on all ABIDE I subjects. This highlights the ability of deep learning models to capture complex interactions among brain regions that were beyond the reach of hand-engineered features and shallow classifiers.

Subsequently, other architectures were explored. Heinsfield et al. [13] achieved 70% accuracy (with 10-fold cross-validation) using autoencoders for unsupervised feature selection and a multilayer perceptron (MLP) for supervised classification. Li et al. [16] introduced a 2-channel 3D CNN model for spatial and temporal features, which increased mean F-scores by 8.5% compared to traditional machine learning approaches. In 2020, Huang et al. [14] obtained 76.4% accuracy with a deep belief network and a graph-based fMRI connectivity feature selection process.

2.2 Transformers in Neuroimaging

Recent popularization of Transformers and Transformer-like architectures provides new possibilities for neuroimaging tasks such as ASD classification, both in performance and explainability.

Transformer Advantages for Neuroimaging The self-attention mechanism in the Transformer architecture enables models to capture long-range dependencies between brain regions. Unlike convolutional operations, self-attention is not limited by local constraints, which has proven to be particularly effective in modeling connectivity relations, as functional connectivity matrices from atlases such as CC200 represent connections without spatial relations.

Moreover, attention mechanisms offer interpretability opportunities by highlighting which sets of features are attended-to the most during classification. This might be invaluable for finding meaningful patterns and realizing the potential of artificial intelligence not only in classification, but also in identifying the sought-after ASD-defining biomarkers.

Relevant Transformer-Based Models for ASD Classification Several Transformer-based architectures have been developed for ASD classification, demonstrating the potential of attention mechanisms in this domain. Deng et al. [7] introduced the Spatial-Temporal Transformer (ST-Transformer), which

learns spatial and temporal features from raw fMRI sequences, achieving 71.0% accuracy on ABIDE I and 70.6% on ABIDE II with 10-fold cross-validation. Liu et al. [17] proposed Spatio-Temporal Cooperative Attention Learning (STCAl), employing co-attention between spatial and temporal features to reach 73.0% and 72.0% accuracy on ABIDE I and II respectively, also using 10-fold cross-validation. Wang et al. [28] developed RGTNet, a residual graph transformer that treats fMRI connections as graphs with ROIs as nodes, achieving 73.4% accuracy on 5-fold cross-validation.

More efficient approaches to processing fMRI have pushed performance boundaries further. Bannadabhavi et al. [4] developed the Community-Aware Hierarchical Transformer (Com-BrainTF), which exploits known fMRI functional communities through a two-level architecture, reporting 72.5% accuracy using a simple 70-10-20 train-validation-test split. Mahler et al. [19] achieved what appears to be current state-of-the-art performance with METAFormer, reaching 83% accuracy and 0.832 AUC on ABIDE-I. This approach combines multiple fMRI parcellation atlases (AAL, CC200, DOS160) and employs self-supervised pretraining with masked connectivity reconstruction, although without pretraining the model reaches only 63% accuracy.

Importantly, all these approaches employed standard k-fold cross-validation or train-test splits rather than leave-one-out cross-validation, which may lead to inflated performance estimates due to site bias effects as demonstrated by Dong et al. [9]. This evaluation methodology limitation makes direct comparison with rigorously evaluated models challenging and highlights the need for more stringent evaluation practices in this domain.

2.3 Generalizability, Robustness, and Evaluation Challenges

Despite significant advances in transformer-based approaches, ASD classification on the ABIDE dataset still faces limitations in current evaluation methodologies due to the condition's heterogeneity and biases from particular measuring environments. Moreover, unlike neurodegenerative conditions such as Alzheimer's disease, where pathological changes produce strong, consistent signals (enabling classification accuracy of 99.92%) [27], ASD involves more subtle alterations in brain connectivity patterns that vary significantly across individuals [15].

The Leave-One-Out Cross-Validation (LOOCV) Challenge A critical limitation in current ASD classification research is the lack of rigorous cross-site evaluation [9]. Most studies, including all Transformer-based approaches mentioned in 2.2 use k-fold cross-validation evaluation within combined multi-site datasets (with the exception of Com-BrainTF, which uses a simple 70-10-20 train-validation-test split [4]). k-fold CV can lead to inflated performance estimates if models learn site-specific artifacts rather than generalizable ASD biomarkers [9].

In 2025, Dong et al. [9] evaluated five of the most widely-used and best-performing models for ASD classification: graph convolutional networks, edge-variational graph convolutional networks, fully connected networks, autoencoders

with fully connected network, and support vector machines. The models were evaluated on the same ABIDE subset, with leave-one-out-cross-validation. Surprisingly, the results show that all models performed almost identically, with around 70% accuracy. According to the study, the difference between these models is not statistically significant. This raises important questions about generalizability of results reported from k-fold cross-validation. For example, the stacked autoencoders with FCN approach, which originally reported accuracy of 85% on 10-fold cross-validation [23], resulted in only 66.80% accuracy when evaluated on leave-one-out cross-validation [9].

2.4 Multimodal Fusion and Tokenization

Most ASD classification studies employ either early fusion (combining features before processing) or late fusion (ensembling separate models), with limited exploration of cross-attention between functional and structural MRI modalities. This represents a significant research gap, as these modalities capture complementary information about brain connectivity and anatomical structure.

Cross-attention has shown promise in other psychiatric conditions. Bi et al. [5] demonstrated its effectiveness for schizophrenia classification, where queries from one modality attend to keys and values from another, achieving 0.833 AUC and outperforming simple concatenation baselines. Zhou et al. [30] further developed this approach using self-attention within each modality followed by cross-attention between modalities, achieving 74% accuracy on schizophrenia datasets.

Tokenization strategies for neuroimaging transformers vary significantly, from patch-based approaches for structural MRI to ROI-based tokenization for functional data. Managing token count is crucial due to quadratic attention complexity, leading to approaches like adaptive token fusion and post-hoc merging [18,**?**]. In ASD classification, METAFormer demonstrates a multi-atlas ensemble approach, using three separate transformers that each process entire flattened connectivity matrices (6670, 19900, and 12720 features respectively) as single embedded inputs rather than individual tokens, with outputs combined through late fusion [19].

3 Methodology

3.1 Dataset and Preprocessing

ABIDE I Dataset This thesis utilizes a subset of ABIDE I from the Autism Brain Imaging Data Exchange (ABIDE) [8], a collaborative initiative of 17 neuroimaging research sites. ABIDE I provides resting-state functional MRI, structural MRI, as well as phenotypic data from individuals with ASD or neurotypical controls. The dataset spans an age range of 7-64 years with a median age of 14.7 years old. Moreover, it contains a bias in gender distribution, with approximately 90% of participants being male [8]. Specifically, a subset of 871 subjects from ABIDE I provided by the Preprocessed Connectomes Project (PCP) [6] was used.

Subject Matching and Data Integrity To ensure generalizable results, a subject-matching procedure was implemented. All experiments utilize the identical 871 matched subjects that exist for both modalities (fMRI and sMRI) and for which it was possible to identify acquisition sites, regardless of the actual available number of subjects for the given modality. This enables leave-one-out cross-validation and prevents unfair model comparisons.

fMRI Processing

Functional Connectivity Features Functional connectivity matrices are computed using the CC200 parcellation atlas, which defines 200 regions of interests (ROIs). Following established preprocessing protocols [19], Pearson Correlation Coefficiens are calculated between the mean time series of all ROI pair, yielding a 200×200 symmetric connectivity matrix. Each element $a_{n,m}$ of the matrix represents the degree to which the regions n and m are activated together, that is, if both always activate and deactivate simultaneously, the value of $a_{n,m}$ is 1. The upper triangle of each connectivity matrix (excluding the diagonal) was extracted and vectorized, resulting in 19900 features per subject (pre-tokenization). This representation captures relevant connectivity data while significantly reducing the amount of redundant information compared to, e.g., representing fMRI features based on raw voxel information.

fMRI Tokenization The baseline fMRI tokenization strategy (no tokenization) treats the whole connectivity vector as one token. Two distinct tokenization approaches were implemented:

- **ROI Connectivity Tokenization** Each of the 200 brain regions (ROIs) becomes an individual token, with its connectivity profile to all other regions (199 features) serving as the token representation. This strategy reconstructs the symmetric 200×200 connectivity matrix from flattened data, with each ROI receiving its connectivity vector to all other ROIs (excluding self-connection). This preserves spatial brain organization and enables region-specific attention patterns.
- **Full Connectivity Tokenization** The entire connectivity matrix is transformed into 8 tokens using different statistical and spatial transformations: raw connectivity, Fisher-Z transformed connectivity, rank-based connectivity, network-ordered connectivity, hemispheric connectivity, distance-based connectivity, and strength-ordered connectivity. Each token contains approximately 19,900 features, representing different analytical perspectives of the same connectivity data. In practice, Fisher-Z transformed connectivity serves as the primary representation due to its optimal statistical properties for correlation data.

sMRI Processing

Structural Features Structural MRI features were extracted using FreeSurfer cortical parcellation, taking inspiration from the extraction process of Dong et al. [9], yielding 800 anatomical measurements per subject. These features contain multiple types of brain morphometry including cortical thickness, surface area, cortical volume, subcortical volume, and white matter integrity measures.

sMRI Tokenization The baseline sMRI tokenization strategy (no tokenization) followed the same pattern as fMRI, treating the entire 800-feature vector as one token. Moreover, two tokenization approaches were selected as most promising:

- **Feature-Type Tokenization** The 800 structural features are grouped into five anatomically meaningful categories: cortical thickness, surface area, cortical volume, subcortical volume, and white matter integrity. Each category forms a separate token (160 features per token), enabling the model to learn relationships between different types of anatomical measurements.
- **Brain Network Tokenization** Features are organized according to 8 major brain networks: default mode network, executive control network, salience network, visual network, motor network, auditory network, frontal network, and limbic network. This organization(100 features per token) allows the model to capture network-specific patterns relevant to ASD.

3.2 Model Architectures

Baseline Models Three baseline architectures were implemented to establish performance benchmarks across single-modality and basic cross-attention approaches. These models process complete feature vectors without tokenization, providing direct comparisons for more complex architectures.

Table 1: Baseline Model Specifications

Model	Input	Params	Key Features
fMRI Baseline	19,900	5.3M	Scaled proj., GELU, enhanced encoders
sMRI Baseline	800	1.8M	Feature eng., dual pooling, stoch. depth
Cross-Att. Basic	19,900+800	7.2M	Bidirectional cross-attention

The baseline models demonstrate increasing architectural complexity from single-modality processing to bidirectional cross-attention, with parameter counts reflecting the additional computational requirements for multi-modal integration.

Tokenization Strategies To enable more sophisticated attention mechanisms, various tokenization strategies were developed that partition input features into meaningful groups. These approaches allow models to learn fine-grained relationships between different brain regions or anatomical measurements.

Table 2: Tokenization Strategy Overview

Strategy	Mod.	Tokens	Feat./Token	Rationale
None (Baseline)	fMRI	1	19,900	Complete connectivity vector
ROI Connectivity	fMRI	200	199	Spatial brain organization
Full Connectivity	fMRI	8	∼19,900	Multiple transformations
None (Baseline)	sMRI	1	800	Complete structural features
Feature-Type	sMRI	5	160	Anatomical groupings
Brain Network	sMRI	8	100	Network-specific patterns

The tokenization strategies were selected based on neurobiological principles, with fMRI tokenization preserving spatial brain organization and sMRI tokenization grouping anatomically or functionally related measurements. This systematic approach enables direct comparison of tokenization effects across modalities.

Cross-Attention Model Combinations Four cross-attention architectures systematically combine different fMRI and sMRI tokenization strategies to investigate optimal multi-modal integration approaches. These models implement bidirectional cross-attention where tokens from each modality attend to tokens from the other modality.

Table 3: Cross-Attention Model Combinations

Model	fMRI Strategy	sMRI Strategy	Tokens	Interactions
Full + Feature	Full Conn. (8)	Feature-Type (5)	13	40
Full + Network	Full Conn. (8)	Brain Net. (8)	16	64
ROI + Feature	ROI Conn. (200)	Feature-Type (5)	205	1,000
ROI + Network	ROI Conn. (200)	Brain Net. (8)	208	1,600

The cross-attention architectures implement the bidirectional attention mechanism described in Section 3.2, with token counts determining computational complexity. Cross-attention interactions scale quadratically as $\mathcal{O}(n_f \times n_s)$, resulting in models ranging from 40 (Full+Feature) to 1,600 (ROI+Network) attention computations per layer.

Bidirectional Cross-Attention The bidirectional cross-attention mechanism allows fMRI and sMRI tokens to exchange information through learned attention weights (Figure 2). Both modalities are first projected to a common dimension d_{model} through learned linear transformations with layer normalization.

Cross-attention is computed bidirectionally: fMRI tokens attend to sMRI tokens and vice versa. For each direction, queries are generated from one modality while keys and values come from the other:

$$\text{CrossAttn}_{f \to s} = \text{softmax}\left(\frac{\mathbf{Q}_f \mathbf{K}_s^T}{\sqrt{d_k}}\right)\mathbf{V}_s \tag{1}$$

After L cross-attention layers with residual connections, the final representations are concatenated and passed through a classification head.

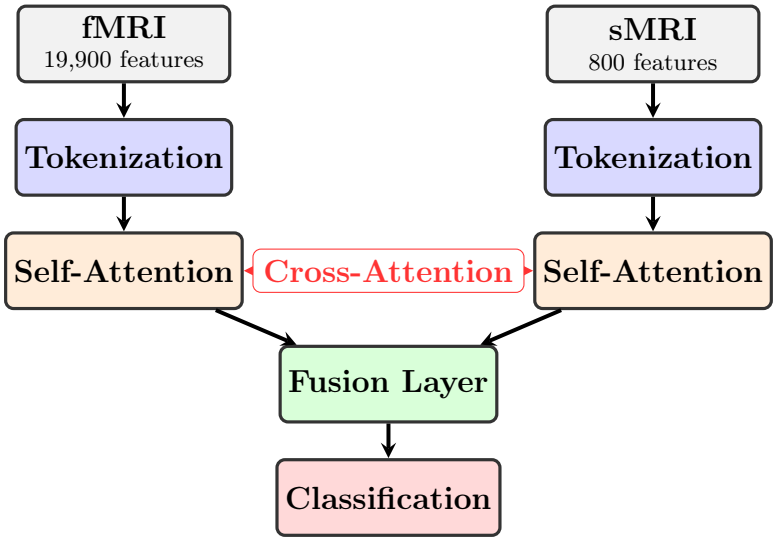


Fig. 2: Bidirectional cross-attention transformer architecture for multimodal autism spectrum disorder classification

Implementation Details Models use learned positional encodings for spatially ordered tokens, dropout regularization (0.1-0.4), and AdamW optimization with early stopping (patience=15). All parameters are initialized using Xavier initialization.

3.3 Hyperparameter optimization

A two-stage hyperparameter optimization was conducted prior to running experiments to ensure optimal model configurations. First, a grid search for the most important hyperparameters was employed on each of the eleven experimental setups. Subsequently, optimal configurations were validated on the training and validation data.

3.4 Evaluation Framework

Two evaluation regimes were employed in order to investigate their role in specific performance scores and variability. The performance of each model is measured in terms of accuracy, balanced accuracy, and Area Under the Curve (AUC) values. Moreover, results from both cases are averaged across the number of sites and random seeds.

Standard Stratified 5-Fold Cross-Validation The first evaluation system uses stratified 5-fold cross-validation with five random seeds for robust performance estimation. k-fold cross-validation is the most widely used evaluation framework in recent Transformer-based studies for ASD classification [9]. Along with stratification, which maintains class balance (ASD and Control) in each fold, early stopping was used with patience-based criteria to prevent overfitting. All models were tested for accuracy, balanced accuracy, and AUC.

Leave-One-Out Cross-Validation To address the critical generalizability challenges identified by Dong et al. [9], leave-one-out cross-validation was implemented, separately from the 5-fold regular cross-validation experiments. In this framework, 20 sites were used as separate folds (data from some hospitals are identified as separate sites in the dataset).

For each fold, 19 sites are used for training and validation purposes, and the model's performance is evaluated on the held-out site. This evaluation strategy directly tests model generalizability to new data acquisition environments, which provides realistic performance estimates for hypothetical clinical deployment scenarios. Similarly to 5-fold cross-validation, LOOCV also employs early stopping and patience.

4 Results

This section presents the findings from evaluating eleven Transformer-based architectures for ASD classification, as outlined in 3.

4.1 Standard 5-Fold Cross-Validation Results

Baseline Model Performance The baseline models established clear performance benchmarks across modalities (Table 4). The fMRI baseline achieved 63.5% ± 3.5% accuracy. In contrast, sMRI baseline reached 59% ± 1.2%, indicating that structural features alone provide less discriminative power for ASD classification in Transformer-like architectures, which is consistent with previous findings that functional connectivity patterns are more distinctly indicative of ASD [15].

The cross-attention baseline, which implements bidirectional attention between single fMRI and sMRI tokens, achieved 69% ± 1.9% accuracy with 0.736 ±

Table 4: Standard 5-Fold Cross-Validation Results

Architecture	Accuracy (%)	Balanced Acc. (%)	AUC
fMRI Baseline	63.5 ± 3.5	63.4 ± 3.6	0.696 ± 0.034
sMRI Baseline	59.6 ± 1.2	59.6 ± 1.1	0.633 ± 0.016
Cross-Attention Basic	69.9 ± 1.9	70.1 ± 1.9	0.736 ± 0.022
Tok. fMRI (Full)	63.7 ± 2.5	63.8 ± 2.6	0.704 ± 0.033
Tok. fMRI (ROI)	63.4 ± 1.0	63.4 ± 0.9	0.709 ± 0.014
Tok. sMRI (Feature)	59.0 ± 3.6	59.0 ± 3.5	0.634 ± 0.024
Tok. sMRI (Network)	57.4 ± 1.7	57.4 ± 1.9	0.608 ± 0.013
Tok. Cross-Att. (Full+Feature)	69.1 ± 2.1	69.0 ± 2.0	0.730 ± 0.017
Tok. Cross-Att. (Full+Network)	69.4 ± 1.8	69.3 ± 1.8	0.729 ± 0.015
Tok. Cross-Att. (ROI+Feature)	**70.1 ± 1.2**	**69.7 ± 1.3**	**0.737 ± 0.017**
Tok. Cross-Att. (ROI+Network)	68.3 ± 1.5	68.3 ± 1.4	0.718 ± 0.018

Note: Results averaged across 5 random seeds with 5-fold cross-validation (25 total evaluations per model). Bold indicates best performance. Tok. = Tokenized; Cross-Att. = Cross-Attention.

0.022 AUC. This 6.4 percentage point improvement over the best single-modality baseline provides strong evidence for the complementary information captured by functional and structural imaging modalities.

Tokenized Single-Modality Results Tokenization strategies showed minimal impact on single-modality performance (Table 4). Both fMRI tokenization approaches performed comparably to the baseline, while sMRI tokenization strategies actually degraded performance, with brain network tokenization showing the poorest results overall. These findings suggest that tokenization alone does not substantially improve single-modality ASD classification.

Tokenized Cross-Attention Results Cross-attention models with tokenization demonstrated the strongest performance across all architectures (Table 4). The ROI+Feature combination achieved the highest accuracy of 70.1% ± 1.2% with 0.737 ± 0.017 AUC, representing the peak performance in this study. This configuration combines 200 ROI-based fMRI tokens with 5 anatomical feature-type sMRI tokens, enabling fine-grained cross-modal attention between brain regions and structural measurements.

4.2 Leave-One-Out Cross-Validation Results

Generalizability Assessment Leave-one-out cross-validation revealed substantially different performance characteristics (Table 5). The cross-attention baseline maintained robust performance (69.7% vs 69.9% in standard CV), demonstrating superior generalizability across acquisition sites.

Table 5: Leave-One-Out Cross-Validation Results

Architecture	Accuracy (%)	Balanced Acc. (%)	AUC
fMRI Baseline	63.0 ± 6.1	62.5 ± 6.8	0.717 ± 0.089
sMRI Baseline	57.4 ± 11.6	58.6 ± 11.1	0.626 ± 0.121
Cross-Attention Basic	**69.7 ± 13.3**	**69.5 ± 12.6**	**0.736 ± 0.157**
Tok. fMRI (Full)	59.2 ± 9.2	57.7 ± 10.0	0.654 ± 0.133
Tok. fMRI (ROI)	64.5 ± 8.1	64.0 ± 7.5	0.691 ± 0.071
Tok. sMRI (Feature)	58.4 ± 9.7	58.0 ± 7.8	0.594 ± 0.099
Tok. sMRI (Network)	60.4 ± 9.5	60.5 ± 8.3	0.642 ± 0.094
Tok. Cross-Att. (Full+Feature)	62.7 ± 14.0	62.5 ± 12.2	0.700 ± 0.159
Tok. Cross-Att. (Full+Network)	63.9 ± 13.0	64.0 ± 13.7	0.720 ± 0.118
Tok. Cross-Att. (ROI+Feature)	63.8 ± 9.9	64.4 ± 9.8	0.697 ± 0.115
Tok. Cross-Att. (ROI+Network)	64.7 ± 15.7	63.9 ± 14.7	0.706 ± 0.160

Note: Results averaged across 20 ABIDE collection sites. Bold indicates best performance. Note the substantial increase in standard deviation compared to Table 4, highlighting site bias effects.

Single-modality models showed increased variability, with sMRI particularly sensitive to site-specific acquisition differences (standard deviation increased from 1.2% to 11.6%). Tokenized cross-attention models degraded under LOOCV evaluation, with the ROI+Feature combination dropping 6.3 percentage points from its peak standard CV performance.

Site Bias Impact Analysis Comparison between evaluation methodologies reveals significant site bias effects across all architectures. Standard deviations increased dramatically under LOOCV evaluation (from 1.0-3.6% to 6.1-15.7%), confirming the presence of site-specific artifacts that inflate performance estimates in standard k-fold validation.

The basic cross-attention model showed the most stable performance across evaluation methods, indicating advantages of simpler fusion strategies for cross-site generalization. Complex tokenized approaches exhibited higher variance, suggesting particular sensitivity to site-specific effects.

Evaluation Method Comparison The systematic comparison validates concerns about inflated performance estimates from k-fold validation [9]. While architecture rankings remained consistent, standard deviations increased 3-5 fold under LOOCV evaluation, revealing the true generalizability challenges in neuroimaging-based ASD classification.

Cross-attention approaches demonstrated superior robustness, with multimodal integration potentially offering resilience to site-specific artifacts. The consistent 70% performance ceiling aligns with established findings [9], indicating fundamental limitations in current neuroimaging features for ASD classification.

5 Discussion and Conclusion

This study presents the first systematic investigation of bidirectional cross-attention between functional and structural MRI for autism spectrum disorder classification. Through evaluation of eleven architectures using both standard k-fold and leave-one-out cross-validation, this work demonstrates the effectiveness of multimodal integration while revealing critical evaluation challenges in neuroimaging-based ASD classification.

5.1 Findings

Cross-Attention Effectiveness Cross-attention consistently outperforms single-modality models, achieving 6.4 percentage point improvements (69.9% vs 63.5% fMRI baseline) while maintaining superior generalizability. The bidirectional attention mechanism enables selective feature integration between modalities, capturing complementary information about brain organization. Importantly, cross-attention models showed stable average accuracy under leave-one-out evaluation (69.9% vs 69.7%), suggesting robustness to site-specific artifacts, though with notably increased variance (1.9% vs 13.3%).

Tokenization Strategy Impact Tokenization strategies showed limited benefits for single-modality models but enhanced multimodal performance under standard cross-validation. The ROI+Feature cross-attention combination achieved peak performance (70.1%) on standard CV. However, performance degradation under leave-one-out cross-validation undermines the contribution of complex tokenization strategies, as the basic cross-attention model demonstrated superior robustness. This suggests that tokenization benefits may be site-specific rather than generalizable.

5.2 Performance Ceiling and State-of-the-Art Comparison

The achievement of 69.9% accuracy without pretraining or extensive feature extraction from raw data is consistent with the established 70% ceiling identified by Dong et al. [9] under rigorous evaluation conditions. Unlike existing Transformer approaches that report higher accuracies using k-fold validation (METAFormer: 83% [19], STCAl: 73% [17]), this thesis' results demonstrate less ambiguous robustness with consistent performance across evaluation methods.

The convergence around 70% accuracy across different approaches suggests fundamental limitations in current neuroimaging-based ASD classification, likely reflecting autism's heterogeneity and imperfections of the ABIDE dataset.

5.3 Evaluation Methodology and Site Bias Implications

The dramatic increase in performance variability under leave-one-out cross-validation confirms substantial site bias in the ABIDE dataset. This finding

validates concerns about inflated performance estimates and provides evidence for adopting leave-one-out cross-validation as standard practice in neuroimaging research.

In the clinical context, the high cross-site variability poses significant challenges for reliable deployment, where consistent performance across different acquisition environments is crucial. The relative stability of simpler cross-attention approaches suggests that, while this technology is certainly not yet feasible for ASD classification, future clinical applications may benefit from prioritizing robustness over peak performance.

5.4 Limitations

Key limitations of this thesis include ABIDE dataset size and biases (gender imbalances, demographic heterogeneity), preprocessing dependencies (CC200 parcellation, FreeSurfer features), and computational constraints limiting the ability to process raw data or explore more sophisticated tokenization methods. Moreover, the tokenization strategies explored represent only a a subset of possible approaches, and the cross-attention mechanism introduces computational complexity that scales quadratically with token count.

The nature of ASD's functional and structural brain differences plays a role too - the 70% performance ceiling may reflect fundamental challenges in capturing autism's diverse etiologies and phenotypes through only neuroimaging patterns, suggesting that current approaches may be approaching their discriminative limits.

5.5 Future Work

Future work should involve evaluating existing Transformer-based models in this domain using leave-one-out cross-validation, exploring other tokenization and feature extraction techniques, which could possibly involve processing raw data files instead of the preprocessed derivatives, and introducing other modalities, such as some forms of the phenotypic data for subjects.

Moreover, a systematic analysis of attention patterns in the attention mechanisms of the models could potentially contribute to identifying consistent, generalizable, and explainable biomarkers, that could provide insights into autism spectrum disorder's neural basis.

5.6 Conclusion

This work demonstrates that bidirectional cross-attention provides meaningful improvements in ASD classification compared to single-modality approaches within the same model architecture framework. However, the consistent 70% performance ceiling highlights fundamental challenges that may extend beyond architectural innovations. The evaluation methodology reveals critical discrepancies between standard k-fold cross-validation estimates and true generalizability,

emphasizing the importance of robust evaluation practices in neuroimaging ASD research.

While current approaches may be approaching their discriminative limits for ASD classification, the superior robustness of cross-attention models suggests that multimodal integration remains a promising direction for future research.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G.: Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. NeuroImage **147**, 736–745 (2017). https://doi.org/https://doi.org/10.1016/j.neuroimage.2016.10.045, https://www.sciencedirect.com/science/article/pii/S1053811916305924
2. Asnaebsa: Cross-sectional t1-weighted mri of a healthy human brain produced at an ultra high-field mr of 7 tesla. https://commons.wikimedia.org/wiki/File:Brain_MRI_7T_slice.jpg (2022), image, Creative Commons Attribution – ShareAlike 4.0 International licence (CC BY-SA 4.0)
3. Atherton, G., Edisbury, E., Piovesan, A., Cross, L., Tchanturia, K., Happé, F.: Autism through the ages: A mixed methods approach to understanding how age and age of diagnosis affect quality of life. Journal of Autism and Developmental Disorders **52**, 3639–3654 (2022). https://doi.org/10.1007/s10803-021-05235-x, https://doi.org/10.1007/s10803-021-05235-x, accepted: 07 August 2021; Published: 04 September 2021; Issue Date: August 2022
4. Bannadabhavi, A., Lee, S., Deng, W., Li, X.: Community-aware transformer for autism prediction in fmri connectome (2023), https://arxiv.org/abs/2307.10181
5. Bi, Y., Abrol, A., Fu, Z., Calhoun, V.: Multivit: Multimodal vision transformer for schizophrenia prediction using structural mri and functional network connectivity data. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2023). https://doi.org/10.1109/ISBI53787.2023.10230385
6. Craddock, C., B.Y.C.C.C.F.E.A.J.A.K.B.L.J.L.Q.M.M.e.a.: The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Frontiers in Neuroinformatics 7, 27 (2013)
7. Deng, X., Zhang, J., Liu, R., Liu, K.: Classifying asd based on time-series fmri using spatial-temporal transformer. Computers in Biology and Medicine **151**, 106320 (2022). https://doi.org/10.1016/j.compbiomed.2022.106320, https://doi.org/10.1016/j.compbiomed.2022.106320, epub 2022 Nov 17
8. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keysers, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Müller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S., Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry **19**(6), 659–667 (June 2014). https://doi.org/10.1038/mp.2013.78, epub 2013 Jun 18

9. Dong, Y., Batalle, D., Deprez, M.: A framework for comparison and interpretation of machine learning classifiers to predict autism on the abide dataset. Human Brain Mapping **46**(5), e70190 (2025). https://doi.org/10.1002/hbm.70190, https://doi.org/10.1002/hbm.70190, published: April 1, 2025
10. Dvornek, N.C., Ventola, P., Pelphrey, K.A., Duncan, J.S.: Identifying autism from resting-state fmri using long short-term memory networks. In: Machine Learning in Medical Imaging (MLMI 2017). Lecture Notes in Computer Science, vol. 10541, pp. 362–370. Springer (2017). https://doi.org/10.1007/978-3-319-67389-9_42, https://doi.org/10.1007/978-3-319-67389-9_42, epub 2017 Sep 7
11. Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E.M., Brammer, M.J., Murphy, C., Murphy, D.G., Consortium, M.A.: Investigating the predictive value of whole-brain structural mr scans in autism: A pattern classification approach. NeuroImage **49**(1), 44–56 (2010). https://doi.org/10.1016/j.neuroimage.2009.08.024, https://doi.org/10.1016/j.neuroimage.2009.08.024
12. Gesi, C., Migliarese, G., Torriero, S., Capellazzi, M., Omboni, A.C., Cerveri, G., Mencacci, C.: Gender differences in misdiagnosis and delayed diagnosis among adults with autism spectrum disorder with no language or intellectual disability. Brain Sciences **11**(7), 912 (2021). https://doi.org/10.3390/brainsci11070912, https://www.mdpi.com/2076-3425/11/7/912, published: 9 July 2021
13. Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., Meneguzzi, F.: Identification of autism spectrum disorder using deep learning and the abide dataset. NeuroImage: Clinical **17**, 16–23 (2017). https://doi.org/10.1016/j.nicl.2017.08.017, https://doi.org/10.1016/j.nicl.2017.08.017, published: August 30, 2017
14. Huang, Z.A., Zhu, Z., Yau, C.H., Tan, K.C.: Identifying autism spectrum disorder from resting-state fmri using deep belief network. IEEE Transactions on Neural Networks and Learning Systems **32**(7), 2847–2861 (2021). https://doi.org/10.1109/TNNLS.2020.3007943, https://doi.org/10.1109/TNNLS.2020.3007943, epub 2021 Jul 6
15. Jiang, C.C., Lin, L.S., Long, S., Zhang, Y.Q., Lin, Y., Song, G.L., Yang, J.Q., Wang, X., Liu, L.M., Yuan, Y.: Signalling pathways in autism spectrum disorder: mechanisms and therapeutic implications. Signal Transduction and Targeted Therapy **7**, 229 (2022). https://doi.org/10.1038/s41392-022-01081-0, https://doi.org/10.1038/s41392-022-01081-0, received: 01 May 2022; Revised: 19 June 2022; Accepted: 23 June 2022; Published: 11 July 2022
16. Li, X., Dvornek, N.C., Papademetris, X., Zhuang, J., Staib, L.H., Ventola, P., Duncan, J.S.: 2-channel convolutional 3d deep neural network (2cc3d) for fmri analysis: Asd classification and feature learning. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1252–1255 (2018). https://doi.org/10.1109/ISBI.2018.8363798, https://doi.org/10.1109/ISBI.2018.8363798, epub 2018 May 24
17. Liu, R., Huang, Z.A., Hu, Y., Zhu, Z., Wong, K.C., Tan, K.C.: Spatial–temporal co-attention learning for diagnosis of mental disorders from resting-state fmri data. IEEE Transactions on Neural Networks and Learning Systems **35**(8), 10591–10605 (2024). https://doi.org/10.1109/TNNLS.2023.3243000
18. Lu, S.Y., Zhang, Y.D., Yao, Y.D.: A regularized transformer with adaptive token fusion for alzheimer's disease diagnosis in brain magnetic resonance images. Engineering Applications of Artificial Intelligence **155**, 111058 (2025). https://doi.org/ https://doi.org/10.1016/j.engappai.2025.111058, https://www.sciencedirect.com/science/article/pii/S0952197625010590

19. Mahler, L., Wang, Q., Steiglechner, J., Birk, F., Heczko, S., Scheffler, K., Lohmann, G.: Pretraining is all you need: A multi-atlas enhanced transformer framework for autism spectrum disorder classification (2023), https://arxiv.org/abs/2307.01759
20. Malik-Soni, N., Shaker, A., Luck, H., Mullin, A.E., Wiley, R.E., Lewis, M.E.S., Fuentes, J., Frazier, T.W.: Tackling healthcare access barriers for individuals with autism from diagnosis to adulthood. Pediatric Research **91**(5), 1028–1035 (2022). https://doi.org/10.1038/s41390-021-01465-y, https://doi.org/10.1038/s41390-021-01465-y, epub 2021 Mar 25
21. OpenStax: fmri illustration (figure 12.7) from *anatomy and physiology*. https://openstax.org/books/anatomy-and-physiology/pages/12-1-basic-structure-and-function-of-the-nervous-system (2016), image, Creative Commons Attribution 4.0 International licence (CC BY 4.0)
22. Parikh, M.N., Li, H., He, L.: Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data. Frontiers in Computational Neuroscience **Volume 13 - 2019** (2019). https://doi.org/10.3389/fncom.2019.00009, https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2019.00009
23. Rakić, M., Cabezas, M., Kushibar, K., Oliver, A., Lladó, X.: Improving the detection of autism spectrum disorder by combining structural and functional mri information. NeuroImage: Clinical **25**, 102181 (2020). https://doi.org/10.1016/j.nicl.2020.102181, https://doi.org/10.1016/j.nicl.2020.102181, epub 2020 Jan 17
24. Sabuncu, M.R., Konukoglu, E., for the Alzheimer's Disease Neuroimaging Initiative: Clinical prediction from structural brain mri scans: A large-scale empirical study. Neuroinformatics **13**, 31–46 (2015). https://doi.org/10.1007/s12021-014-9238-1, https://doi.org/10.1007/s12021-014-9238-1
25. Santomauro, D.F., Erskine, H.E., Mantilla Herrera, A.M., Miller, P.A., Shadid, J., Hagins, H., Addo, I.Y., Adnani, Q.E.S., Ahinkorah, B.O., Ahmed, A., Alhalaiqa, F.N., Ali, M.U., Al-Marwani, S., Almazan, J.U., Almustanyir, S., Alvi, F.J., Amer, Y.S.A.D., Ameyaw, E.K., Amiri, S., Andrei, C.L., Angappan, D., Antony, C.M., Aravkin, A.Y., Ashraf, T., Ayuso-Mateos, J.L., Barrow, A., Batra, K., Bemanalizadeh, M., Bhagavathula, A.S., Bhaskar, S., Bhatti, J.S.S., Bolla, S.R., Britton, G., Castaldelli-Maia, J.M., Catalá-López, F., Caye, A., Chattu, V.K., Chong, Y.Y., Ciobanu, L.G., Cortese, S., Cruz-Martins, N., Dachew, B.A., Dai, X., Darwish, A.H., Dashti, M., de la Torre-Luque, A., Diaz, D., Ding, D.D., Dy, A.B.C., Dziedzic, A.M., Ebrahimi Meimand, S., El Meligy, O.A.A., El Sayed, I., Elgar, F.J., Fagbamigbe, A.F., Faris, P.S., Faro, A., Ferreira, N., Filip, I., Fischer, F., Gandhi, A.P., Ganesan, B., Gebregergis, M.W., Gebrehiwot, M., Ghaderi Yazdi, B., Ghasemi, M.R., Ghasemzadeh, A., Gunturu, S., Gupta, V.B., Gupta, V.K., Halim, S.A., Hall, B.J., Han, F., Haro, J.M., Hasaballah, A.I., Hay, S.I., Hedley, D., Helfer, B., Hossain, M.M., Hwang, B.F., Ibrahim, U.I., Ilaghi, M., Islam, M.R., Islam, S.M.S., Iyer, M., Jaggi, K., Jahrami, H., Jamshidi, E., Khaleghi, A., Khan, A.A., Khan, M.J., Khidri, F.F., Kim, K., Koh, H.Y., Kumar, M., Landires, I., Le, L.K.D., Lee, S.W., Li, Z., Lim, S.S., Martinez-Raga, J., Marzo, R.R., Matthew, I.L., Maugeri, A., Mestrovic, T., Mitchell, P.B., Mohammed, S., Mokdad, A.H., Monasta, L., Montazeri, F., Mrejen, M., Mughal, F., Murray, C.J.L., Myung, W., Nauman, J., Newton, C.R.J., Nguyen, H.L.T., Nri-Ezedi, C.A., Nwatah, V.E., Oladunjoye, A.O., Olufadewa, I.I., Ordak, M., Otstavnov, N., Palma-Alvarez, R.F., Parikh, R.R., Park, S., Pasovic, M., Patel, J., Pereira, M., Pereira, M.O., Phillips, M.R., Polanczyk, G.V., Pourfridoni, M., Puvvula, J., Radfar, A., Rahim, F., Rahman, M., Rahman, M.A., Rahmani, A.M., Rahmati,

M., Ratan, Z.A., Rhee, T.G., Ronfani, L., Roy, P., Saddik, B.A., Saghazadeh, A., Sakshaug, J.W., Salehi, S., Samuel, V.P., Sankararaman, S., Saravanan, A., Satpathy, M., Schumacher, A.E., Schwebel, D.C., Šekerija, M., Shafiee, A., Shahabi, S., Shamim, M.A., Silva, J.P., Solomon, Y., Sumpaico-Tanchanco, L.B.C., Swain, C.K., Tabarés-Seisdedos, R., Temsah, M.H., Tromans, S.J., Tzivian, L., Varma, R.P., Vinueza Veloz, A.F., Vinueza Veloz, M.F., Walde, M.T., Waqas, M., Wickramasinghe, N.D., Yesodharan, R., Yon, D.K., Youm, Y., Zaman, B.A., Zeng, Y., Zielińska, M., Whiteford, H.A., Brugha, T., Scott, J.G., Vos, T., Ferrari, A.J.: The global epidemiology and health burden of the autism spectrum: findings from the global burden of disease study 2021. The Lancet Psychiatry **12**(2), 111–121 (2025). https://doi.org/10.1016/S2215-0366(24)00363-8, https://www.sciencedirect.com/science/article/pii/S2215036624003638, funding: Queensland Health and the Bill & Melinda Gates Foundation

26. Skuse, D.: Autism – 25 years on: A lot has changed! Clinical Child Psychology and Psychiatry **25**(3), 721–725 (2020). https://doi.org/10.1177/1359104520929729, https://doi.org/10.1177/1359104520929729
27. Sorour, S.E., El-Mageed, A.A.A., Albarrak, K.M., Alnaim, A.K., Wafa, A.A., El-Shafeiy, E.: Classification of alzheimer's disease using mri data based on deep learning techniques. Journal of King Saud University - Computer and Information Sciences **36**(2), 101940 (2024). https://doi.org/https://doi.org/10.1016/j.jksuci.2024.101940, https://www.sciencedirect.com/science/article/pii/S1319157824000296
28. Wang, Y., Long, H., Bo, T., Zheng, J.: Residual graph transformer for autism spectrum disorder prediction. Computer Methods and Programs in Biomedicine **247**, 108065 (2024). https://doi.org/https://doi.org/10.1016/j.cmpb.2024.108065, https://www.sciencedirect.com/science/article/pii/S0169260724000610
29. Xu, M., Calhoun, V., Jiang, R., Yan, W., Sui, J.: Brain imaging-based machine learning in autism spectrum disorder: Methods and applications. Journal of Neuroscience Methods **361**, 109271 (2021). https://doi.org/10.1016/j.jneumeth.2021.109271, https://doi.org/10.1016/j.jneumeth.2021.109271, epub 2021 Jun 24
30. Zhou, Z., Orlichenko, A., Qu, G., Fu, Z., Calhoun, V.D., Ding, Z., Wang, Y.P.: An interpretable cross-attentive multi-modal mri fusion framework for schizophrenia diagnosis (2024), https://arxiv.org/abs/2404.00144