# Multi-Modal Explainable Machine Learning for Automated Detection of Autistic Spectrum Disorder

Meryem Ben Yahia, Moncef Garouani, Julien Aligon

# Multi-Modal Explainable Machine Learning for Automated Detection of Autistic Spectrum Disorder

By

**Meryem Ben Yahia**

Supervisors:

**Moncef Garouani**
**Julien Aligon**

Machine Learning and Data Mining
Université Jean Monnet



# Multi-Modal Explainable Machine Learning for Automated Detection of Autistic Spectrum Disorder


## Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by diverse symptoms affecting social interaction, communication, and behavior. Diagnosing ASD is challenging due to variability among individuals and the lack of clear biomarkers. Although artificial intelligence (AI) has enhanced diagnostic accuracy, the "black-box" nature of many models limits their clinical use. Addressing current limitations, this paper introduces a multimodal ASD detection framework using deep neural networks (DNN) with explainable AI (xAI) to enhance model transparency. Our proposed approach achieves a mean 5-fold accuracy of 99.66% for the fMRI-based model and 99.83% for the multimodal-based model, which leverages both fMRI and phenotypic data, surpassing existing methods while offering better interpretability.



**Keywords:** *ASD; multimodal diagnostic frameworks; explainable AI; deterministic brain atlas; functional connectivity; connectivity matrix; deep neural network.*


## 1. Introduction

This research journey began with approaching ASD detection through the lens of a machine learner, rather than a medical practitioner. Early on, it became clear that before any solution could be developed, it was essential to define the problem—a task that often proves as challenging as solving it. However, the broader research objective extended well beyond understanding existing approaches to the problem. It centered on a crucial question: how can the power of AI be effectively harnessed in medicine, and why has its deployment been so limited despite its potential?

There were many unknowns when this study effort first started.It was unclear if machine learning techniques could detect autistic symptoms reliably. Which data modalities could be used to detect autism was another unanswered question. Using a trial-and-error process, we weaved a web of inquiry and experimentation that gradually revealed insights. The focus of the endeavor expanded on the study of the intersection of artificial intelligence and medicine, the consideration of patients' individual characteristics, and uncovering the obstacles that prevent these technologies from being fully realized in clinical practice.

Machine learning has already shown great promise in psychiatry. For example, studies have demonstrated that algorithms using clinical and neuroimaging data can predict antidepressant responses with 80% accuracy, outperforming traditional methods (Chekroud et al., 2016). Similarly, machine learning models have been used to identify patterns in brain imaging data, enabling accurate classification of mental health disorders like schizophrenia and bipolar disorder, with accuracies reaching up to 87% (Vieira et al., 2020).

Despite these advances, the question remains: if machine learning holds such promise, why isn't it more widely used in clinical practice? The answer lies in a significant challenge—the lack of interpretability in machine learning models. Clinicians require models that are not only accurate but also transparent and understandable to ensure patient safety and informed decision-making (Ghassemi et al., 2021). The lack of interpretability and the ensuing mistrust are the main obstacles to the real application of machine learning in medicine. Moreover, they need technology that is adaptable and cognizant of many patient-specific variability in modalities.

Our aim in doing this research was to bridge the gap between AI's promise and its practical, dependable applications in healthcare. The intention was to contribute to making technology understandable, accessible, and truly transformative for patient care. To this end, the study is structured to provide a comprehensive exploration of both the theoretical and practical dimensions of applying multimodal explainable machine learning for automated detection of ASD. Section 3 surveys ASD research, covering its historical of its understanding and diagnosis, and ASD detection machine learning models. Section 4 outlines our methodology and model architecture (Figure 1). Section 6 then provides presents a SHAP-based analysis comparing multimodal and fMRI models, with detailed gender-based and subgroup analyses of brain regions using atlas-based visualizations. Lastly, the conclusion and future work sections provide a summary of our findings and suggest avenues for additional research.

## 2. Context of the Project

The research project was conducted during a five-month internship at the Computer Science Research Institute of Toulouse (IRIT) in France, specifically within the Generalised Information Systems team[1] (SIG), one of the most prominent teams in the institute. The SIG team focuses on developing methods and tools that enable efficient information access, simplify analysis, and support decision-making. This research was supervised by Dr. Moncef Garouani[2], whose work specializes in the automatic selection and parametrization of machine learning algorithms and AI explainability, and Dr. Julien Aligon[3], who focuses on post-hoc methods in prediction explanation. The primary objectives and contributions of this research are detailed in the subsequent sections.

### 2.1. Informative Feature Selection

The goal was to develop a feature selection strategy that prioritizes quality-controlled, informative variables for effectively differentiating ASD. For this purpose, we used the Preprocessed Autism Brain Imaging Data Exchange I (ABIDE I) dataset. Our approach incorporated metrics from automated and manual assessment protocols to refine and clean the data. For the fMRI data, Recursive Feature Elimination (RFE) was used to isolate and preserve the most essential features critical for distinguishing ASD. A detailed description of the dataset is provided in section 4.1, with a summary in Table 1.

### 2.2. Multi-modal Data Fusion

A key objective of our research was to improve ASD diagnosis accuracy by incorporating various data types and accounting for individual variations. We achieved this by transforming fMRI data and combining it with phenotypic data into a unified vector for holistic analysis. Details of this data transformation method are outlined in section 4.3.

### 2.3. Explainable AI

To ensure that clinicians can trust and comprehend the outcomes of the automated ASD diagnosis system, a primary research goal was to offer transparent and comprehensible insights into the decision-making process. To explain the output of our machine learning models, we used SHapley Additive exPlanations (SHAP), a game-theoretic method. The approach is detailed in Section 6 .

[1]https://www.irit.fr/en/departement/dep-data-management/sig-team/

[2]https://mgarouani.fr/

[3]https://www.irit.fr/ Julien.Aligon/

## 3. Literature Review and Survey of the State-of-the-Art

Throughout history, there has been a major evolution in the classification and understanding of what is now known as Autism Spectrum Disorder (Wing, 1997). The first seminal and systematic description of early infantile autism was put forth by psychiatrist Leo Kanner in 1943 through meticulous observations of a group of eleven children (Kanner, 1943).

Following Kanner's first research, theories were proposed by the psychiatric literature to account for the basic deficiencies in autism, with the majority of these theories treating the condition as a mental illness. In contrast, today, it is understood that autism is a complicated neurodevelopmental disorder that predominantly affects the brain, and other systems of the body. The discoveries of structural imaging studies have contributed to the near universal recognition of autism as a brain-based condition rather than a behavioral one, a distinction that is subtle but important (Minshew & Williams, 2007). Due to the latter, cognitive and neurological models have been preferred over emotional and inter-subjective ones in ASD research in recent decades (Harris, 2018).

Nonetheless, the vast majority of cases (around 85%) of ASD are classified as non-syndromic[4] and idiopathic[5], meaning that the etiology is unknown and they do not fit into any particular category of recognized genetic and neurological syndromes (National Institute of Health, 2019).

### 3.1. Autism Disorder Spectrum Diagnosis

Autism Spectrum Disorder encompasses a wide range of related conditions, including unique symptoms and traits. It is characterized as "a syndrome composed of subgroups" rather than a singular disorder, and thus presents differently across individuals (Maser & Akiskal, 2002). Given the complexity, diversity, lifelong nature, and high prevalence of ASD, affecting approximately 1 in 36 children (Maenner et al., 2023), comprehensive diagnostic approaches are central to accurately identifying and addressing the symptoms associated with the disorder.

The contemporary diagnostic approach for ASD primarily involves subjective interviews and a detailed review of the patient's behavior and developmental history by the physician. Tools such as the Autism Diagnostic Interview-Revised (ADI-R) and the Diagnostic and Statistical Manual

[4]Non-syndromic refers to a condition that occurs without the additional presence of other recognizable clinical features that would typify a specific syndrome.

[5]Idiopathic describes a condition or disease whose cause is not known or understood. In medical terms, idiopathic conditions are those for which the genetic basis remains unclear or unidentifiable with current technology and knowledge.

of Mental Disorders, Fifth Edition (DSM-5) are used in the evaluation process (Lordan et al., 2021). Owing to the overreaching range of symptoms, there are no precise diagnostic standards that apply to every person with ASD. Furthermore, experts continue to dispute on the most accurate criteria of making the diagnosis (Hus & Segal, 2021).

In light of these uncertainties, applications of machine learning to medicine hold significant promise. For example, in a study by Koutsouleris et al. (2018), machine learning models trained on functional, neuroimaging, and combined baseline data to predict social outcomes at 1 year achieved up to 83% accuracy in patients at high risk for psychosis, outperforming human prognostication.

### 3.2. Autism Detection, fMRI and Multi-Modality

The invention and accessibility of non-invasive brain imaging methods has made it possible to gain a deeper comprehension of the neuronal circuitry responsible for a variety of neurological disorders. Magnetic Resonance Imaging (MRI) has been used to identify a variety of neuropsychiatric and neurodegenerative disorders, including schizophrenia (Jafri et al., 2008), Alzheimer's (Chen et al., 2011), and so forth.

In their 2023 review, Alharthi and Alzahrani conducted a comprehensive examination of ASD diagnosis using MRI techniques in scientific literature. Their extensive search was conducted across various conferences and journals from 2020 to 2023, and was meticulous in gauging the methodologies and conclusions related to the diagnosis, detection, and classification of ASD. The analysis focused on machine learning-based approaches in ASD diagnosis, particularly deep learning-based methods such as Multi-Layer Perceptrons (MLP), Convolutional Neural Networks (CNN), Autoencoders (AE), Graph Convolutional Networks (GCN), and Graph Attention Networks (GAT), among other models, all utilizing MRI modalities. The maximum accuracy of the reviewed methods ranged from 54.79% (Sharif & Khan, 2022) to 99.19% (Kim et al., 2021). Furthermore, this review played a crucial role in choosing the best methodologies. In-depth topic expertise is necessary to select appropriate features and data modifications. In fMRI, the brain is depicted as thousands of voxels—small cubes whose activities are monitored over time as time series data. Working with such data is an extraordinarily challenging research challenge because of the brain's complex structure, nonlinear separability, high dimensionality, and the sequential changes in traceable signals inside each voxel.

For this reason, the methodology of our proposed architecture and the data transformation were a combination and enhancement of state-of-the-art methods. In particular, we improved upon the DNN architecture outlined "Deep Learning Approach to Predict Autism Spectrum Disorder Using Multi-site Resting-State fMRI" by (Subah et al., 2021).

## 4. Methodological Contribution

In this section, we outline the data preparation and transformation processes, as well as propose a deep learning method that integrates phenotypic data with fMRI connectivity measures from the ABIDE I dataset. The proposed method effectively integrates multimodal data and employs RFE to reduce dimensionality, serving as input for our neural networks. The outputs are interpreted using SHAP-based visualizations, which offer interpretable insights into the contributions of various features. The complete methodology is illustrated in Figure 1.

### 4.1. The dataset: Autism Brain Imagine Network ABIDE I

Our study used data from the Autism Brain Imaging Data Exchange I, a 2012 initiative involving 17 international sites. It included 1112 subjects, 539 with ASD and 573 controls, aged 7 to 64 years (median age 14.7). The dataset and its data legend is open access and can be found at the following links: **Access Link — Data Legend**.

*Table 1.* Summary of ABIDE I Dataset

| **Category** | **ABIDE I** |
|---|---|
| **Participants** | 1112 total (539 ASD, 573 controls) |
| **Age Range** | 7-64 years (median: 14.7) |
| **Number of Sites** | 17 international sites |
| **Imaging Data** | - Resting State Functional MRI<br>- Structural MRI |
| **Phenotypic Data** | - Composite Phenotypic File<br>- Phenotypic Data Legend |

The phenotypic data, in Table 1, is under a CSV file format with 106 columns (features), and 1112 rows (subjects).

### 4.2. Data Cleaning of Phenotypic Data

To maintain dataset integrity, we removed patients with corrupt files or missing FILE_ID values, indicative of missing fMRI scans. Despite these efforts, a sparsity of 32.71% remained. For biological data, traditional imputation methods for addressing missing values can compromise the validity of statistical analyses, as highlighted by Sterne et al. (2009), making them unreliable for accurate data interpretation. Therefore, to avoid biases and preserve the essential characteristics of the dataset, we framed the issue of missing data as an optimization problem, in which we sought to maximize the feature-to-example ratio while minimizing missing-example-to-feature ratio.

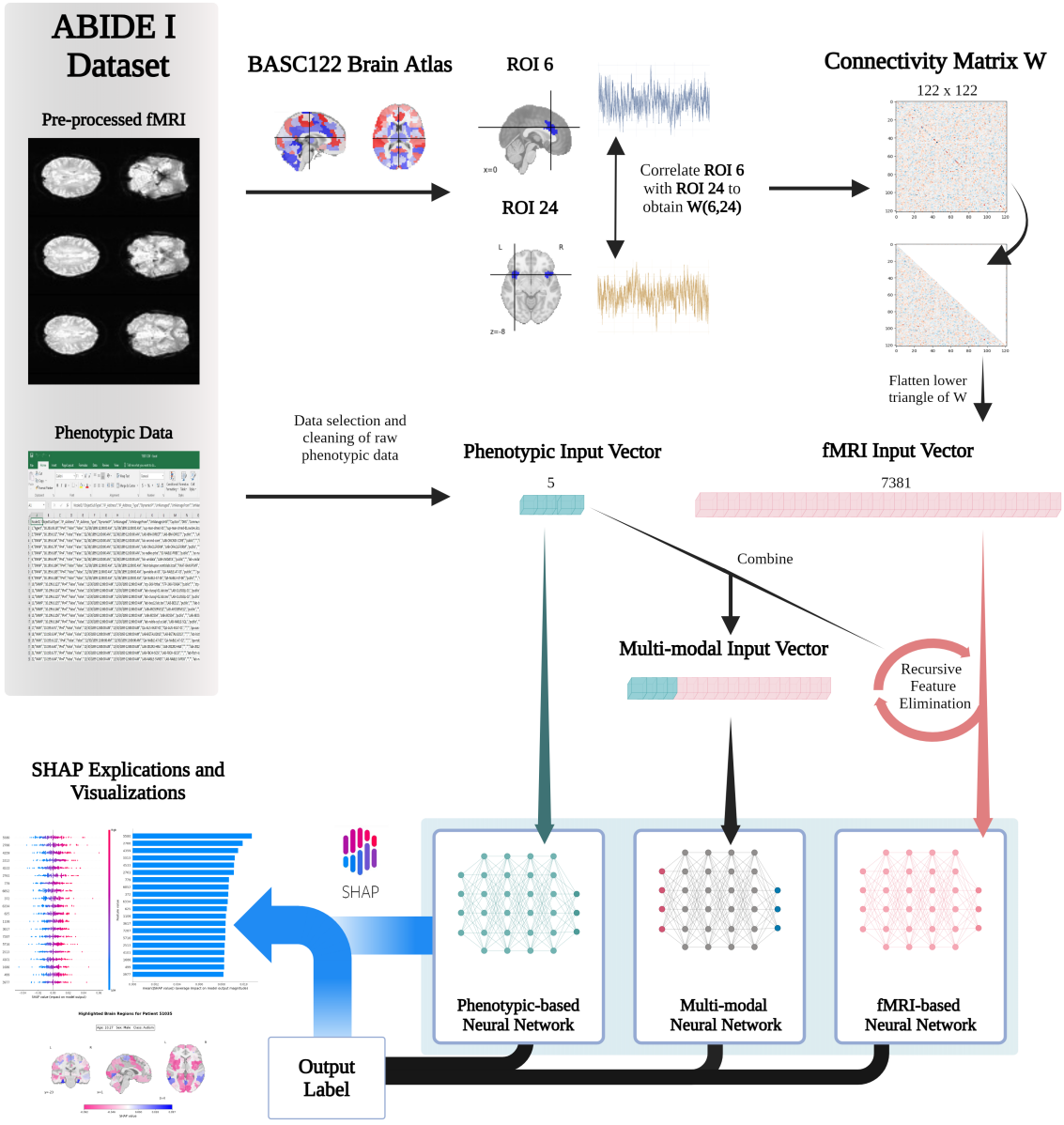


*Figure 1.* Architecture of the Proposed Method for Autism Spectrum Detection with xAI.

The objective function is defined as follows:

**Maximize**

$$\mathcal{L} = \frac{\text{num\_cols}}{\text{total\_rows}} - \frac{\text{num\_cols}}{\text{thresh}}$$

where:

- num_cols: Number of features.
- total_rows: Total number of observations.
- thresh: Minimum threshold for the number of non-missing examples per feature.

**Subject to the constraint:**

$$\text{total\_rows} \geq 2 \times (\text{num\_cols} - 1)$$

The optimization process retained 12 columns, as shown in Table 2.

Furthermore, we eliminated uninformative columns, such as FIQ, PIQ, and VIQ test types, due to their high levels of missingness and weak correlation with the corresponding test scores. The final set of phenotypic features, described in Table 2 is: AGE_AT_SCAN, SEX, FIQ, VIQ, and PIQ.

*Table 2.* Descriptions, Types, and Coding of Key Features

| Feature | Description | Type | Coding |
|---|---|---|---|
| **DX Group** | Diagnostic group | Categorical | 1: ASD, 2: Control |
| **Age at scan** | Age at the time of the scan (years) | Numeric | N/A |
| **Sex** | Biological sex | Categorical | 1: Male, 2: Female |
| **FIQ** | Full-Scale Intelligence Quotient | Numeric | N/A |
| **VIQ** | Verbal Intelligence Quotient | Numeric | N/A |
| **PIQ** | Performance Intelligence Quotient | Numeric | N/A |

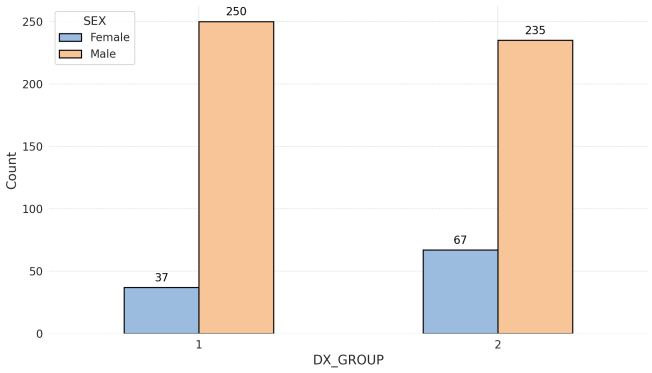


*Figure 2.* Count of Males and Females in each DX_GROUP

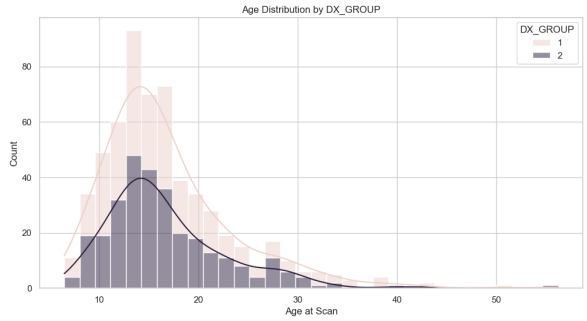


*Figure 3.* Age Distribution by DX_GROUP

The final dataset exhibits a pronounced male predominance in both the ASD and control groups, with 250 males and 37 females in the ASD group, and 235 males and 67 females in the control group (Figure 2), which aligns with the higher prevalence of ASD in males in existing epidemiological data (Fombonne, 2009). Regarding age distribution (Figure 3), the ASD group is predominantly younger, with a concentration of individuals between 5 and 15 years old, with peaks around 10 years, whereas the control group spans a wider age range, extending from childhood into late adulthood with peaks with a peak at 15 years.

### 4.3. Data Transformation of fMRI Data

The study concentrated on specific brain Regions of Interest (ROIs) rather than analyzing the time series from every voxel. We leveraged the BASC122 (Bootstrap Analysis of Stable Clusters) brain atlas[6] which defines 122 networks, to delineate these regions from imaging data (Bellec et al., 2010; Liu et al., 2009).

Functional connectomes[7] were constructed using the tangent embedding of the Ledoit-Wolf regularized covariance estimator, implemented through the Nilearn library (Nilearn). This process involves selecting specific voxels at each time point from 4D fMRI scans using 3D masks, thereby converting the 4D data into a 2D time-series representation. From this time-series data, a symmetric tangent connectivity matrix was generated and simplified by retaining only the lower triangular values. These values were then flattened into a 1D feature vector of size 7,831. Each element in the vector represents the interaction between a pair of distinct ROIs in the BASC122 atlas.

Lastly, to diminish dimensionality, RFE was applied with a logistic regression estimator to obtain the top 500 features.

### 4.4. Classification Using a Deep Neural Network

The DNN architecture, illustrated in Figure 4, includes two hidden layers and was adapted for three datasets: phenotypic, multimodal, and fMRI. Hyperparameters were finedtuned using Random Search and Hyperband, for each modality used, with the optimal configuration validated via stratified 5-fold cross-validation.

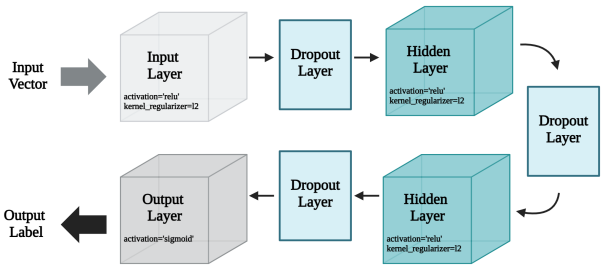


*Figure 4.* Architecture of the Deep Neural Network Classifier

To prevent overfitting, each hidden layer was followed by a dropout layer with an optimal rate and used the Rectified Linear Unit (ReLU) activation function. The output layer

[6]A brain atlas is a detailed map that identifies and categorizes different regions and structures of the brain, often used as a reference in neuroscience for studying brain anatomy and function.

[7]Functional connectomes are representations of the functional connections in the brain, illustrating how different regions interact with one another during rest or task performance.

employed a sigmoid function, while L2 regularization and training with a small batch size enhanced generalization and mitigated overfitting with high-dimensional data.

The network's parameters, including the weights $W_i$ and biases $b_i$ for each hidden layer $i$, , were tuned using the Adam optimizer with an optimal learning rate. The latter targeted minimizing the loss function, specifically the binary cross-entropy, defined as follows:

$$J = -\frac{1}{m}\sum_{i=1}^{m}[y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))]$$

where $m$ represents the total number of samples, $y_i$ is the true label, and $p(y_i)$ is the predicted probability of $y_i$ belonging to a specific class (e.g., ASD or control).

### 4.5. Explainability using SHAP

SHAPley values were first introduced by Shapley et al. (1953) as a solution in cooperative game theory to fairly distribute the total gains among players in an alliance. In machine learning, SHAPley (SHAP) values quantify the contribution of individual features to a model's prediction.

In this work, we employed the KernelExplainer (shap-0.46.0), a model-agnostic method within the SHAP framework, which approximates SHAPley values by treating the model as a black box and using perturbations to estimate feature contributions, as shown in Figure 5.

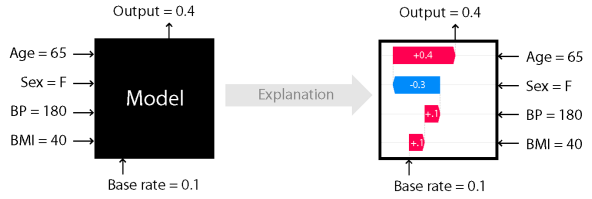


*Figure 5.* SHAP Values for Model Output Explanation (Lundberg et al., 2024)

The SHAPley value for a particular feature $i$ is mathematically defined as (Molnar, 2019):

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!}\,[f(S \cup \{i\}) - f(S)]$$

where:

- $N$: Set of all features.
- $S$: Subset of $N$ excluding feature $i$.
- $|S|$: Number of elements in subset $S$.
- $f(S)$: Prediction using only features in $S$.
- $f(S \cup \{i\})$: Prediction with feature $i$ added to $S$.
- $\phi_i$: SHAPley value of feature $i$.

KernelExplainer operates by generating a series of perturbed instances around the data point of interest, then analyzing how the model's predictions change when features are omitted to obtain a detailed breakdown of feature contributions.

## 5. Model Performance Evaluation

In this section, we present the evaluation of the DNN architecture outlined in section 4.4, with their mean accuracy from 5-fold cross-validation summarized in Table 3.

*Table 3.* Mean Accuracy across 5-Folds for fMRI, Multimodal and Phenotypic Models

| Fold | fMRI | Multimodal | Phenotypic |
|---|---|---|---|
| 1 | 99.15% | 99.15% | 68.64% |
| 2 | 100.00% | 100.00% | 61.86% |
| 3 | 100.00% | 100.00% | 62.39% |
| 4 | 99.15% | 100.00% | 62.39% |
| 5 | 100.00% | 100.00% | 63.25% |
| **Mean Accuracy** | 99.66% | 99.83% | 63.71% |
| **Standard Deviation** | ±0.42 | ±0.34 | ±2.51 |

To assess the performance of our DNN, we evaluated the fMRI, phenotypic, and multimodal models corresponding to each fold using 5-fold cross-validation and saved the one achieving the highest accuracy among all folds as the best-performing model to use for the performance evaluation. The training and evaluation were conducted on Google Colab, utilizing an Intel(R) Xeon(R) CPU @ 2.20GHz (8 virtual CPUs), 51 GB of RAM, and an NVIDIA Tesla T4 GPU with 15 GB of GPU memory. The computational setup substantially expedited the training process, with the maximum training time across model being at most 60 seconds.

### 5.1. Phenotypic Model

The phenotypic model, developed using only five features (AGE_AT_SCAN, SEX, FIQ, VIQ, PIQ), underperformed despite extensive fine-tuning, indicating these features were insufficient for accurate ASD prediction.

*Table 4.* Classification Report for Phenotypic-based Model

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.49 | 0.39 | 0.43 | 54 |
| 1 | 0.56 | 0.66 | 0.60 | 64 |
| **Accuracy** | | 0.53 (118) | | |
| **Macro Avg** | 0.52 | 0.52 | 0.52 | 118 |
| **Weighted Avg** | 0.53 | 0.53 | 0.53 | 118 |

Table 4 shows the model's accuracy at 53%, akin to random guessing in binary classification. It has a precision of 0.49, recall of 0.39, and F1-score of 0.43 for Class 0. The confusion matrix (Figure 6), with 33 false positives and 22 false negatives, indicates difficulty in distinguishing between the two classes.

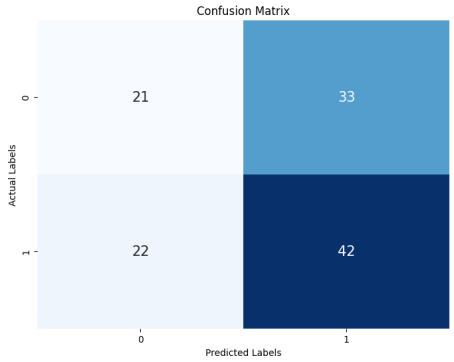


*Figure 6.* Confusion Matrix for Phenotypic-based Model

The model's limited performance is attributed to its restricted feature set, which comprises basic demographic variables that are inherently insufficient for diagnosing any disorder.

### 5.2. fMRI Model

The fMRI-based model outperforms the phenotypic model by far. From the classification report in Table 5, both recall and precision for each class achieve a perfect score of 1.00, resulting in an F1-score of 1.00 across all metrics.

*Table 5.* Classification Report for fMRI-based Model

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 54 |
| 1 | 1.00 | 1.00 | 1.00 | 64 |
| **Accuracy** | | 1.00 (118) | | |
| **Macro Avg** | 1.00 | 1.00 | 1.00 | 118 |
| **Weighted Avg** | 1.00 | 1.00 | 1.00 | 118 |

With no false positives or false negatives, the confusion matrix (Figure 7) shows that every sample was correctly classified. The ROC curve (Figure 8) further supports this, with an Area Under the Curve (AUC) of 1.00.

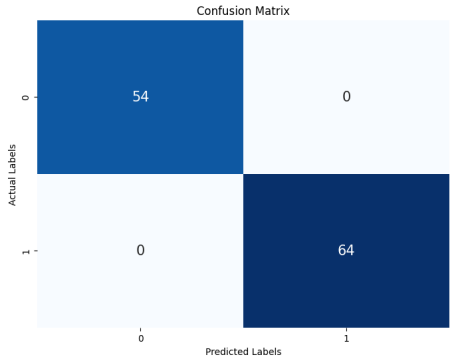


*Figure 7.* Confusion Matrix for fMRI-based Model

Figures 9 and 10, depict the accuracy and loss plots during training, demonstrate a rapid convergence. The model attains near-perfect accuracy at an early stage and sustains this performance until the end of the training session.

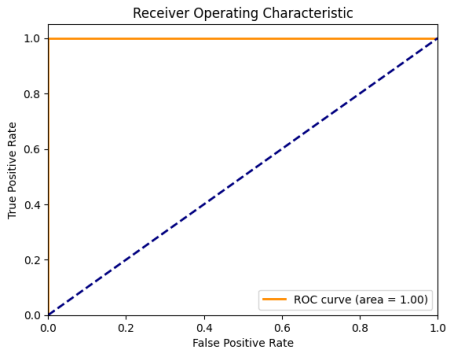


*Figure 8.* ROC AUC for fMRI-based Model

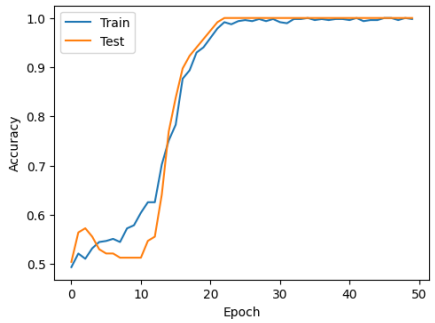


*Figure 9.* Training Accuracy for fMRI-based Model

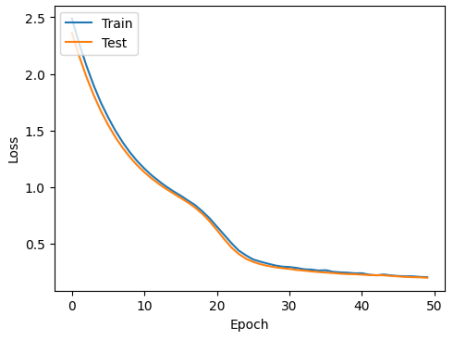


*Figure 10.* Training Loss for fMRI-based Model

Several factors contribute to our model's high performance. Primarily, the CPAC pipeline preprocessing improved scan quality by reducing artifacts from breathing, head motion, and heartbeat. Crucially, feature selection significantly improved the model's performance. Before RFE, the model, despite fine tuning, achieved a maximum accuracy of 69.8%. The current fMRI-based model (Table 3) has a 5-fold mean accuracy of 99.66% with a standard deviation of 0.42.

### 5.3. Multimodal Model

The Multimodal approach offers the most stable model, with an added ability to incorporate a wider variety of features. As shown in Table 3, the model achieves a mean accuracy of 99.83% with a standard deviation of 0.34.

*Table 6.* Classification Report for Multimodal Model

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.98 | 1.00 | 0.99 | 54 |
| 1 | 1.00 | 0.98 | 0.99 | 64 |
| **Accuracy** | | 0.99 (118) | | |
| **Macro Avg** | 0.99 | 0.99 | 0.99 | 118 |
| **Weighted Avg** | 0.99 | 0.99 | 0.99 | 118 |

The precision and recall values are nearly perfect for both classes (Table 6). The ROC curve (Figure 12) further underscores this, with an AUC of 1.00, indicating near-perfect class discrimination. The confusion matrix (Figure 11) reinforces these results, with only one misclassification out of 118 cases, accurately identifying all Class 0 instances and only one error in Class 1 predictions, demonstrating the model's robustness.

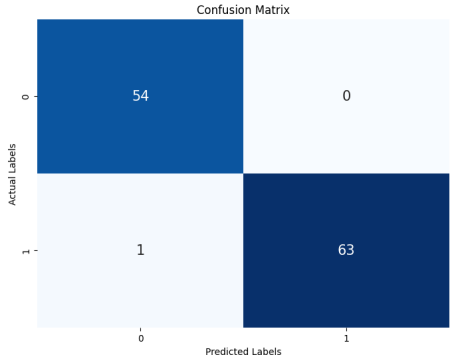


*Figure 11.* Confusion Matrix for Multimodal Model

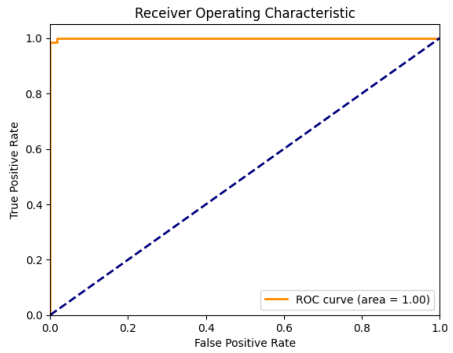


*Figure 12.* ROC AUC for Multimodal Model

## 6. SHAP Interpretation

In this section, we will showcase the results obtained using SHAP to interpret the models' predictions. Given that the phenotypic model did not perform well, it will not be discussed further. Rather, the focus will be on comparing the multimodal and fMRI models, with an emphasis on the latter one.

For the sake of disclosure and transparency, the brain visualizations in this section should be taken as indicative rather than definitive. They offer helpful visual support rather than a precise view of brain region importance. Similarly for the anatomical and functional labeling of the BASC122, as the labels are approximated and do not fully capture the ROIs.

The importance of brain regions in visualizing uses SHAP score estimates obtained by mapping each feature in the fMRI input vector to its corresponding pair ROIs in BASC122. Since BASC122 regions lack labels, we approximated their functional and anatomical designations by aligning them with regions in established brain atlases, including Yeo's networks (Yeo et al., 2011), the AAL atlas (Rolls et al., 2020), the Harvard-Oxford atlas[8], and the Juelich atlas[9].

### 6.1. Comparison of fMRI and Multimodal SHAP Values

In each figure, we displayed the top 20 features. Beginning with the multimodal model's global summary (Figure 13), FIQ emerges as the most influential feature, exhibiting the highest SHAP values. PIQ and VIQ also play significant roles, though their influence is less pronounced. From AGE_AT_SCAN onward, the remaining features show smaller mean SHAP values, indicating a more moderate impact on the model's predictions.

The Beeswarm plot (Figure 14) provides a detailed visual representation of how each feature contributes to the multimodal model's output across all instances. Each point represents a SHAP value for a particular feature. Red values indicate positive impact on the model predicting the ASD class, while the blue values indicate a negative contribution (control class).

FIQ is once again stands out as the most influential feature, with a wide spread and clear separation of SHAP values, showing significant contributions both positively and negatively depending on the instance. PIQ and VIQ also show notable contributions, but with less variance compared to FIQ. The clustering of points near zero for these features suggests that while they are important, their impact is more moderate and consistent across different instances. The remaining features exhibit a tighter cluster of SHAP values

[8]https://neurovault.org/collections/262/
[9]https://julich-brain-atlas.de/

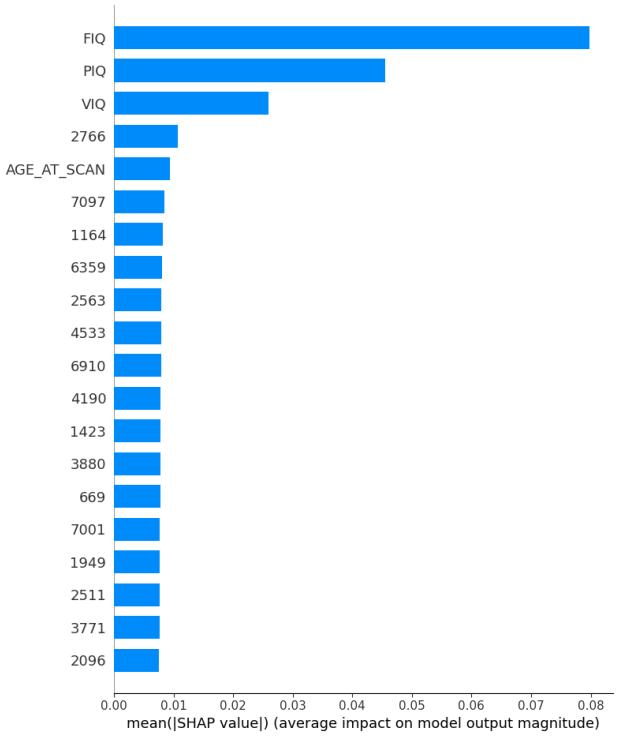


*Figure 13.* Global Summary Plot for Multimodal-based Model

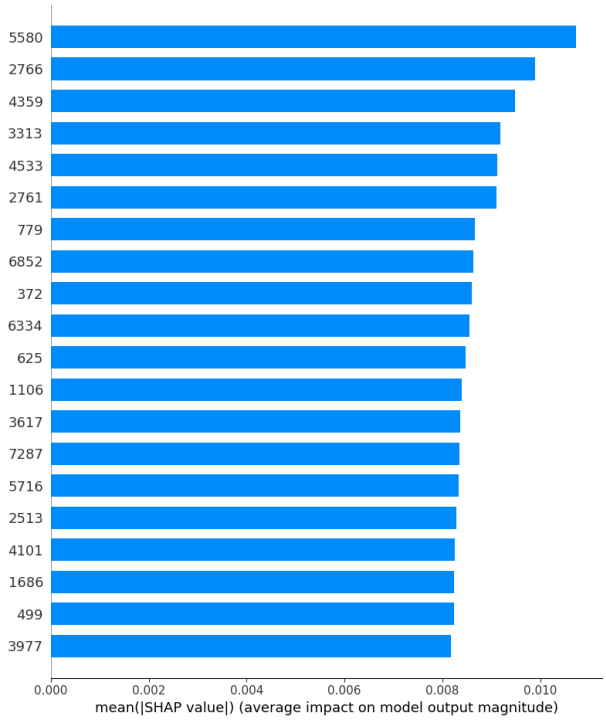


*Figure 15.* Global Summary Plot for fMRI-based Model


*Figure 14.* Beeswarm for Multimodal-based Model

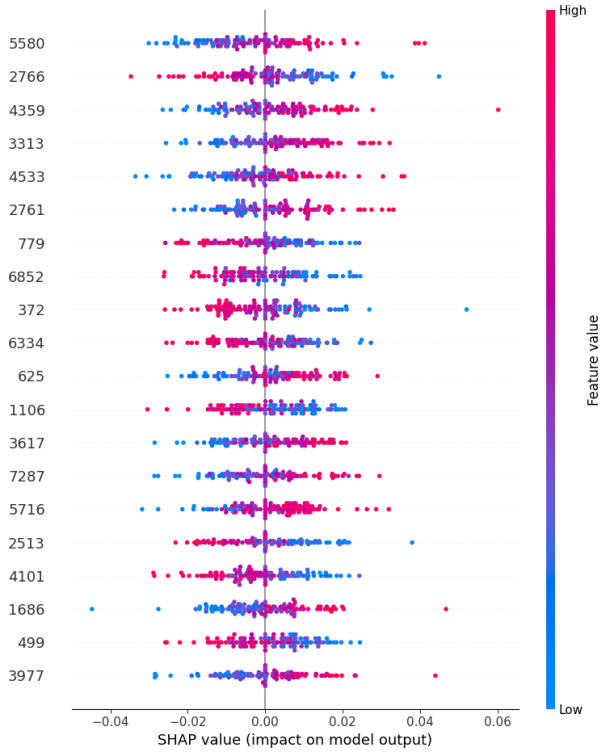


*Figure 16.* Beeswarm for fMRI-based Model

around zero, indicating that their influence on the model's predictions is generally weaker and more stable.

In contrast, the fMRI model's global summary plot (Figure 15) shows a different set of fMRI dominant features, with brain regions such as 5580, 2766, and 4359 being the most influential. Feature 5580 is quite prominent in the fMRI model, yet it is absent from the top features in the multimodal model. It indicates that 5580 is especially significant when IQ measurements are excluded. When comparing the global summary plots of the fMRI model (Figure 13) to those of the multimodal model (Figure 15), the fMRI model shows a consistent distribution of SHAP values. It reveals that a larger range of impacts are captured by the fMRI model from its features, which may be because the brain imaging data is more variable. We note, however, that features 2766 and 4533 are included in the top 20 most influential characteristics of both the fMRI model and the multimodal model.

In summary, the comparison shows that the multimodal model is mainly driven by cognitive features like FIQ, PIQ, and VIQ, while the fMRI model emphasizes specific brain regions that are less influential in the multimodal analysis. We hypothesize that FIQ, PIQ, and VIQ can effectively substitute for the absent regions in the multimodal model, as these metrics capture a broad range of brain functions that the fMRI model isolates into specific regions. Additionally, SHAP results indicate that further refinement, such as reapplying RFE, could have further reduced the feature vector and isolated more informative features.

### 6.2. Analysis of Influential Features across fMRI and Multimodal Models

Features 5580, 2766, and 4533 stand out as particularly significant. Feature 5580 is the most influential in the fMRI-based model, while feature 2766 appears in both models, ranking second in the fMRI model and fourth in the multimodal model, where it is the highest-ranked non-phenotypic feature. Although feature 4533 is not as highly ranked, it is the only other feature present in the top 20 of both models. Given their prominence, these features warrant further investigation to better understand their impact on the model's predictions. For additional insight, we have visualized features 5580, 2766, and 4533, along with their approximate anatomical and functional mappings, in Figures 18, 19, and 20. In our analysis, we consider these features as activations, assuming they correspond with the activation of their respective pairs of regions.

Starting with feature 5580, figure 18 highlights the brain regions 107 and 15. Region 107 is associated with visual network[10] (shown in 17), while Region 15 linked to the default mode network (DMN)[11] and prefrontal network[12].

Figure 19 highlights the brain regions linked to feature 2766, focusing on Region 75, part of the frontoparietal and default mode networks, responsible for cognitive control, and Region 65, involved in the default mode and executive control networks. Region 75 includes areas within the left cerebral white matter, associated with cognitive functions like working memory and attention.

Figure 20 highlights the brain regions activated by feature 4533, with Region 96, part of the dorsal attention and default mode networks, involved in emotional regulation and cognitive processes, and Region 68, within the somatomotor network[13], crucial for motor control, emotional regulation, and language processing.

To put it simply, feature 5580 links areas related to visual processing, executive functioning, and referential reasoning. While feature 2766's regions are both linked to cognitive tasks and working memory. In addition, feature 4533 pairs a region related to emotional control with another associated with language processing. All of these areas, assuming that there is atypical activity in them, can be linked to ASD symptoms. Individuals with ASD often struggle with maintaining eye contact (related to the visual processing network), understanding language, and executive functioning, which affects impulse control and can lead to repetitive behaviors (Pelphrey et al., 2011).

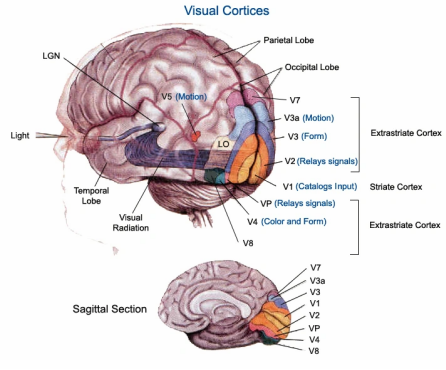


*Figure 17.* The Six Regions of the Visual Cortex (Brain for AI Wiki, n.d.)

[10]The Visual Network is a collection of brain regions involved in processing visual information, including areas like the occipital lobe and calcarine cortex.

[11]The Default Mode Network is a network of interconnected brain regions active during rest and involved in self-referential thinking, memory retrieval, and mind-wandering.

[12]The Prefrontal Network comprises brain regions within the prefrontal cortex that are critical for executive functions, decision-making, and social behavior.

[13]The somatomotor network is a brain network responsible for integrating sensory input and coordinating motor activities, including voluntary movements and body sensation processing.

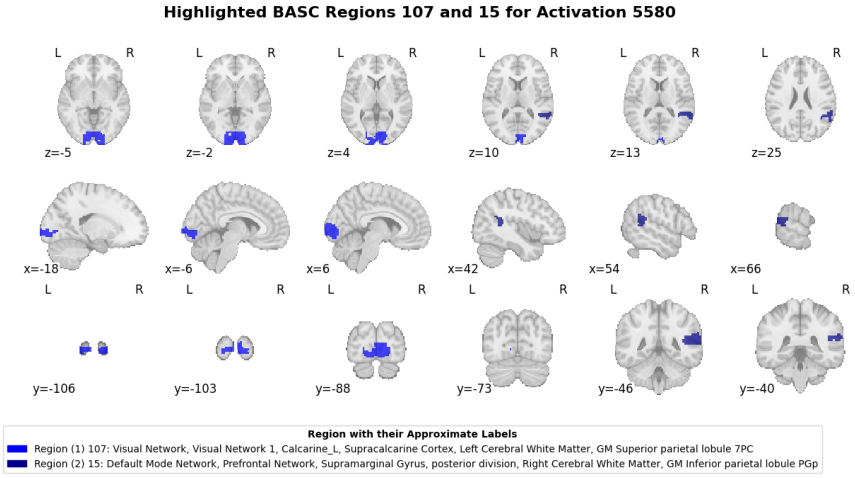


*Figure 18.* Visualization of Activation 5580

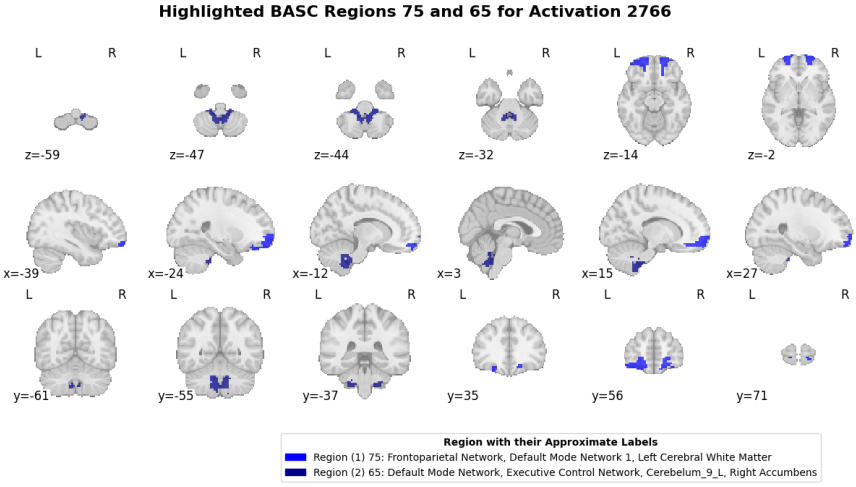


*Figure 19.* Visualization of Activation 2766

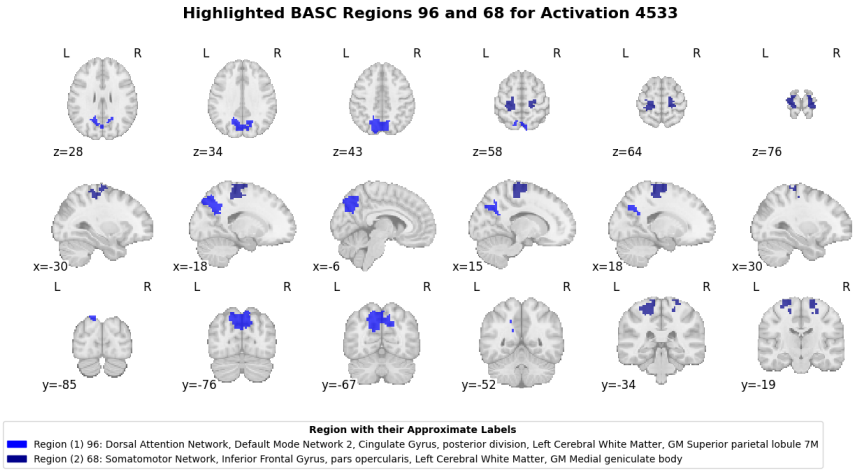


*Figure 20.* Visualization of Activation 4533

### 6.3. Analysis by Gender for fMRI-based Model

Research has shown significant sex differences in the brain processes underlying autism spectrum disorder. In a study by (Kendall, 1938) analyzed sex-specific changes in intrinsic functional connectivity using a multicenter resting-state fMRI dataset using the same dataset as the one used for this research, ABIDE. Their findings revealed that ASD-related alterations in functional connectivity differ markedly between males and females. Specifically, females with ASD exhibited patterns of hyper-connectivity that aligned more closely with the higher connectivity levels typically seen in males, suggesting neural masculinization[14]. Conversely, males with ASD showed patterns of hypo-connectivity that resembled the lower connectivity levels typically seen in females, indicating neural feminization[15]. Therefore, it was crucial to investigate sex-specific differences in predictions, as highlighted in the literature. For brevity, this discussion will focus exclusively on the fMRI-based model.

Figure 21 plot represents the SHAP values for the top features in a model trained on male participants (both ASD and non-ASD) using fMRI data. Feature 5580 is present once again at the top, along with 2766 and 2761. Features such as 4533, 4359, and 625 are also influential, although slightly less than the top three. Instead than mostly depending on one feature, the model's predictions appear to be evenly impacted by a variety of brain regions.

The Beeswarm plot (Figure 22) illustrates the SHAP values for the top features influencing the fMRI-based model's predictions for males. Feature 5580 remains the most influential, with a wide spread of SHAP values across both sides of the zero line. Features 2766 and 2761 also show significant variability, indicating their strong influence on the model. The red gradient suggests that higher values of these features tend to push predictions towards the ASD class. The spread of SHAP values highlight the variable impact depending on individual differences.

Similar to the male SHAP values, feature 5580 is the most influential in the female model, suggesting that this brain region plays a critical role in the model's predictions for females as well (Figure 23). Unlike the male group, where features like 2761 and 4533 were more prominent, the female model emphasizes different regions, such as 3313 and 4965.

Figure 23 for the female participants provides a more granular view of how each feature contributes to the fMRI-based

[14]Neural masculinization is the process where the brain develops male-typical traits, mainly due to the influence of sex hormones like testosterone during key stages of brain development.

[15]Neural feminization is the process where the brain develops female-typical traits, usually due to lower levels of testosterone and the influence of hormones like estrogen, during critical periods of brain development.

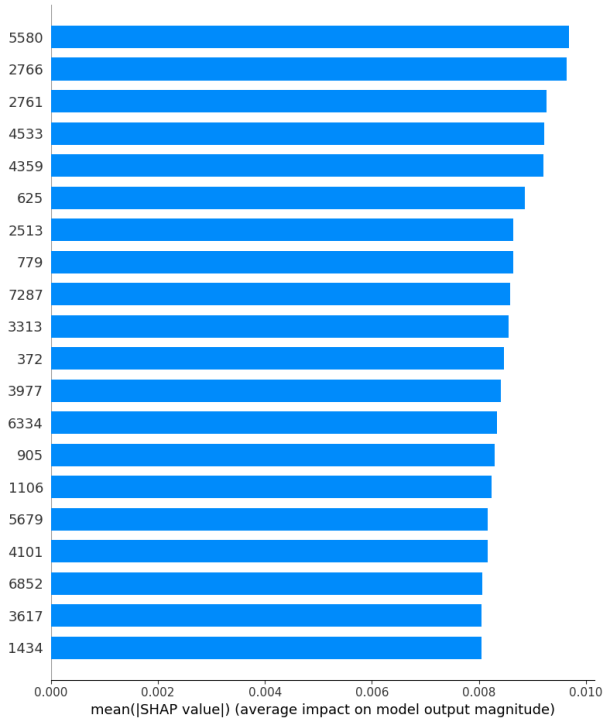


*Figure 21.* Global Summary Plot for Males in the fMRI-based Model

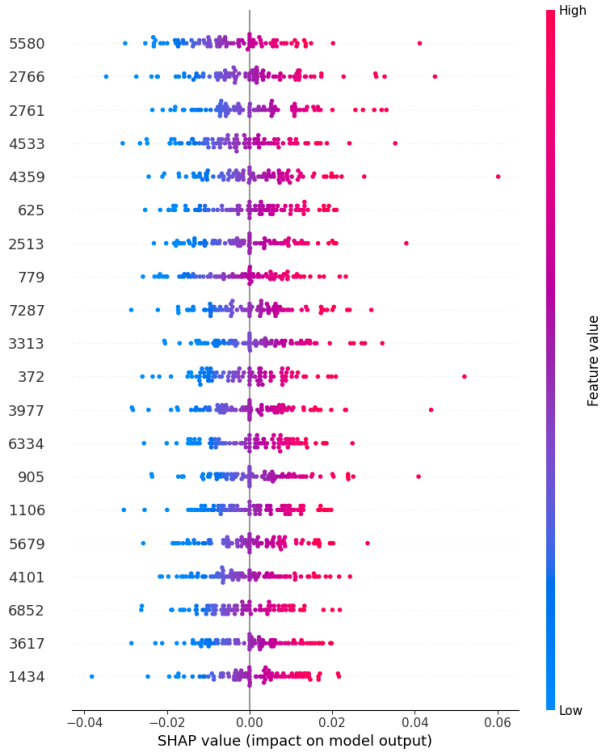


*Figure 22.* Beeswarm for Males in the fMRI-based Model

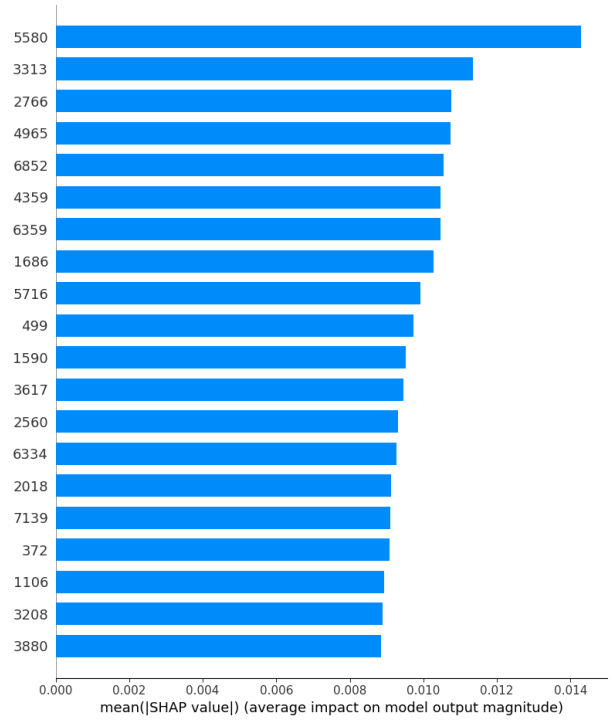


*Figure 23.* Global Summary Plot for Females in the fMRI-based Model

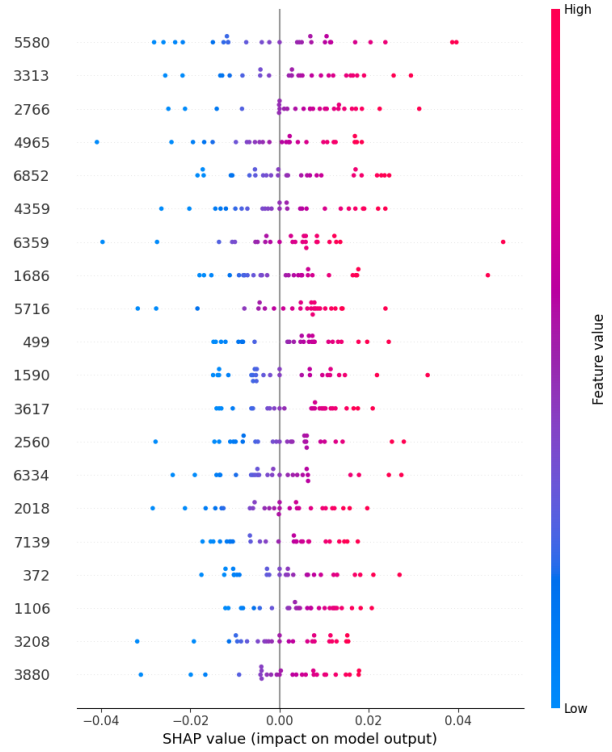


*Figure 24.* Beeswarm for Females in the fMRI-based Model

model's predictions. As with the males, Feature 5580 is highly influential in the females, with SHAP values spread across both positive and negative impacts. Compared to the male model, the female model shows different feature having more impact (e.g., 3313 and 4965) and a wider distribution of SHAP values for certain features, suggesting more variability in how these regions influence ASD-related outcomes in females.

The SHAP analysis of fMRI-based models for males and females reveals that while some features are influential across both genders, there are distinct differences in the brain areas contributing to ASD predictions. In males, regions such as 2761 and 4533 play a more prominent role, whereas in females, regions like 3313 and 4965 are key contributors. The findings suggest that incorporating sex-specific distinctions into predictive models could improve their accuracy and offer deeper insights into the unique neurobiological underpinnings of ASD across genders, thereby enhancing model interpretability.

### 6.4. Sampling of ASD vs. Non-ASD Subjects

The objective of this sampling is to capture individual variations in predictions across both ASD and control groups, accounting for differences in cognitive abilities and sex. The sample set includes four participants (Table 7): two with ASD and two controls, each group comprising one male and one female. The female ASD participant (ID: 50276), aged 16.8, has exceptionally high cognitive scores, while the male ASD participant (ID: 50606), aged 16.42, exhibits extremely low IQ scores. Among the controls, the male participant, aged 19.75, has IQ scores in the low average range, whereas the female participant, aged 15.75, has high average IQ scores.

*Table 7.* Grouped data by DX_GROUP with Autism and Control categories

| Group | ID | Age | Sex | FIQ | VIQ | PIQ |
|---|---|---|---|---|---|---|
| Autism | 50276 | 16.8 | Female | 146.5 | 145 | 148 |
| | 50606 | 16.42 | Male | 41 | 42 | 37 |
| Control | 50467 | 19.75 | Male | 89 | 89 | 90 |
| | 50572 | 15.75 | Female | 105 | 126 | 82 |

#### 6.4.1. ASD: Male vs. Female

The SHAP waterfall plot (Figure 25) and brain region visualization (Figure 26) analyze a 16.8-year-old female with autism. The waterfall plot identifies features 4965, 3880, and 2018 as the most significant contributors, with positive SHAP values driving the prediction toward ASD. Although features like 3681 and 7372 exert negative SHAP values, slightly nudging the prediction toward non-ASD, the cumulative positive contributions from other features result in a high ASD probability (f(x) = 0.952). The aggregation

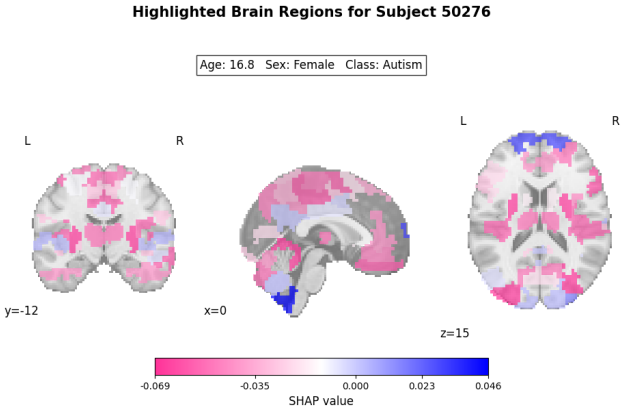


*Figure 25.* Waterfall Plot for Patient 50276

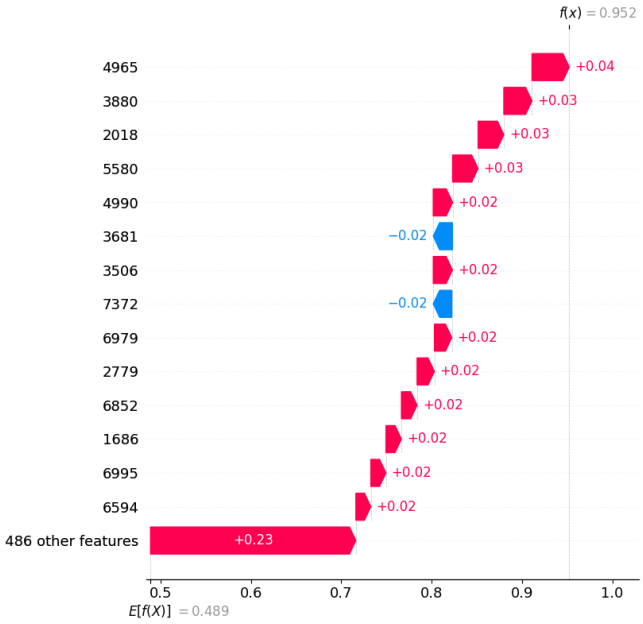


*Figure 26.* Highlighted Brain Regions for Patient 50276

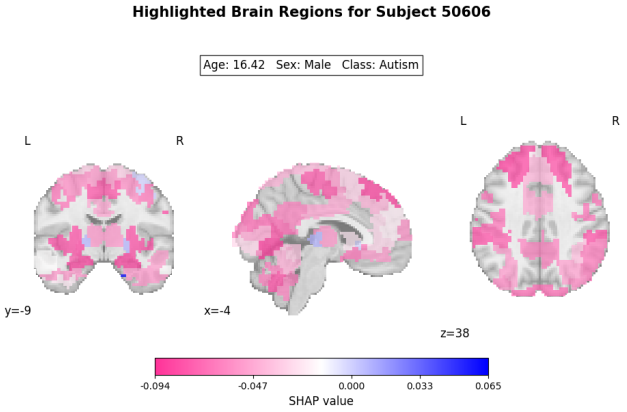


*Figure 27.* Waterfall Plot for Patient 50606

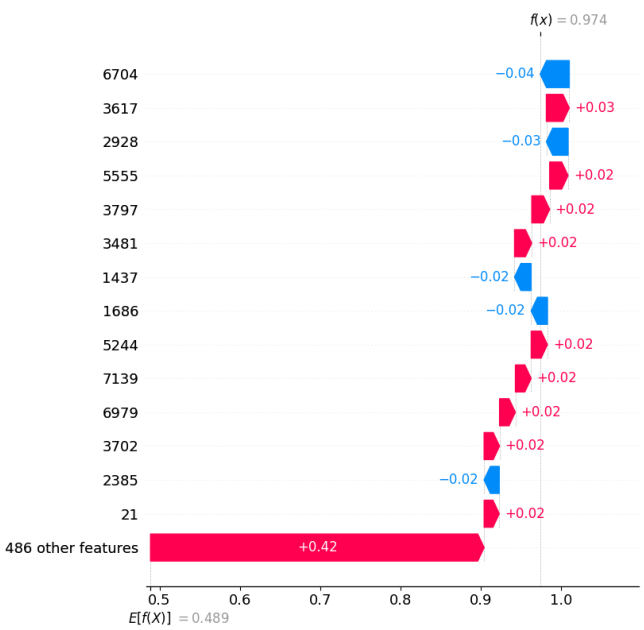


*Figure 28.* Highlighted Brain Regions for Patient 50606

of the remaining 486 features further reinforces the ASD prediction. In Figure 26, the predominance of pink regions signifies a strong influence toward ASD. The blue regions, although fewer, are highly concentrated, suggesting a significant counteracting influence concentrated in one region for a non-ASD prediction. In contrast, the SHAP waterfall plot (Figure 28) and brain region visualization (Figure 27) for a 16.42-year-old male reveal a high predicted probability (f(x) = 0.974) for ASD. Despite five of the top features displaying negative SHAP values, subtly pulling the prediction away from ASD, the brain visualization is predominantly pink, indicating a majority of brain regions aligned with an ASD classification. Thus, it suggests that the aggregation of all features plays a more decisive role than individual top features.

### 6.4.2. Non-ASD: Male vs. Female

The waterfall plot (Figure 31) for the 15.75-year-old female reveals that the top features are almost evenly split between negative and positive SHAP values, each exerting nearly equal influence. Despite this balance, the corresponding brain visualization (Figure 32) predominantly displays blue regions. Similarly, the waterfall plot (Figure 30) for the 19.75-year-old male control exhibits a comparable trend, yet his brain visualization (Figure 31) shows some areas with concentrated pink regions. In contrast to the ASD sample, where the top features in the waterfall plots predominantly were predominantely red, the control sample's features were more evenly distributed between red and blue. This pattern of distribution suggests that the model's predictions for the

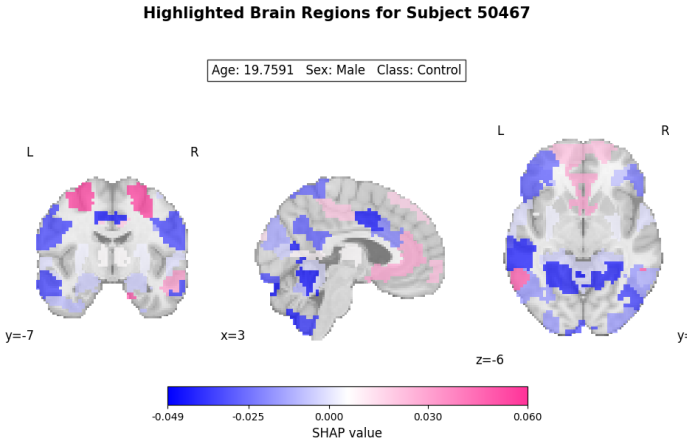


*Figure 29.* Waterfall Plot for Patient 50467

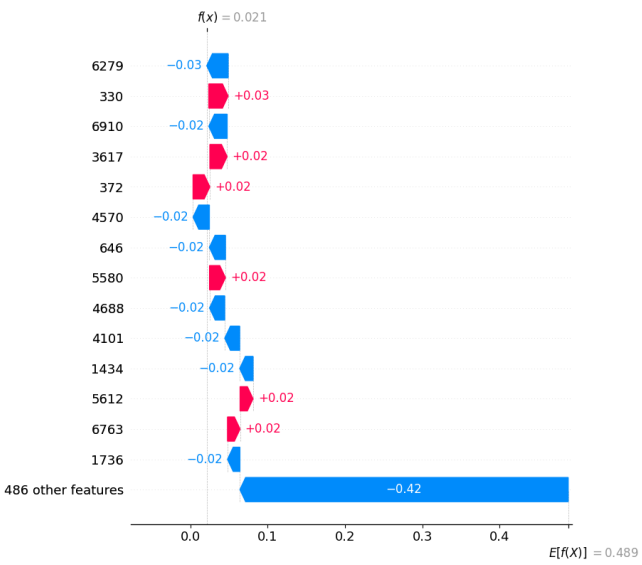


*Figure 30.* Highlighted Brain Regions for Patient 50467

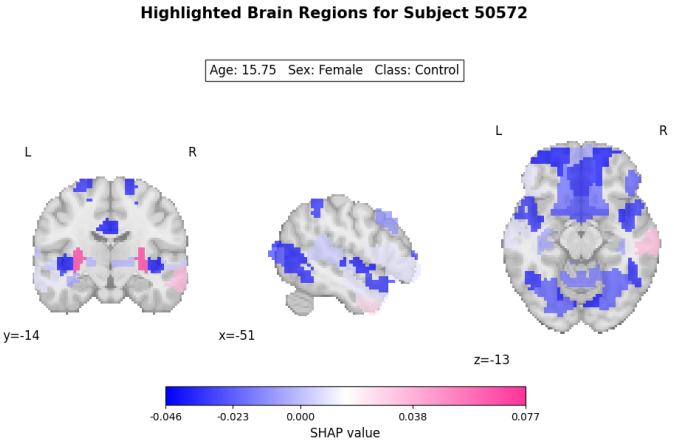


*Figure 31.* Waterfall Plot for Patient 50572

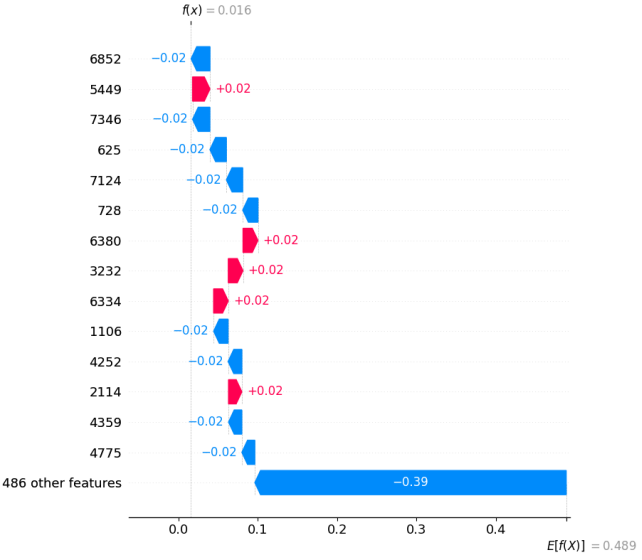


*Figure 32.* Highlighted Brain Regions for Patient 50572

control group rely on a more balanced combination of fMRI features, rather than being influenced by specific features as seen in the ASD group.

## 7. Conclusion and Perspectives

In this study, we explored the interpretability of machine learning models applied to fMRI, phenotypic, and multimodal data, with a focus on SHAP-based interpretability in predicting ASD. The fMRI and multimodal models achieved near-perfect accuracy, underscoring their efficacy, while the phenotypic model underperformed, suggesting that demographic variables alone are inadequate for reliable ASD diagnosis. Our findings highlight that integrating multiple modalities significantly enhances model performance.

The SHAP analysis revealed significant variability in feature importance across the models. In the Multimodal Model, FIQ, PIQ, and VIQ emerged as dominant contributors, suggesting that these readily accessible cognitive metrics, when combined with fMRI connectivity features, effectively encapsulate brain functions. Conversely, the fMRI model exhibited an evenly distributed pattern of feature importance.

The study acknowledges several shortcomings, particularly in the interpretation of results. An important limitation lies in the approximate nature of brain labeling used to interpret SHAP findings, which may impact the precision of the conclusions. Additionally, the absence of specialized neurological expertise limited the contextualization of these interpretations. Time constraints further hindered a fine-grained analysis and the exhaustion of potential experiments.

## 8. Future Works

This work serves as a preliminary exploration ahead of a potential research paper. Moving forward, we aim to validate our methods through expert review and further immersion in relevant literature to better link our findings to established brain structures. Additionally, we plan to extend our analysis by experimenting on atlases with functional labeling. Enhancing our SHAP analysis will involve exploring feature interactions, performing dependence versus independence tests, and conducting detailed subgroup analyses across variables such as sex, age category, IQ type, and stratified pairings.

## 9. Acknowledgement

I would like to begin by acknowledging my father's memory that remains a constant source of motivation.I am profoundly grateful to have been his daughter and to follow the path his love led me to.

I am deeply appreciative of my supervisors at IRIT, Dr. Garouani and Dr. Aligon, whose guidance and support were pivotal in my journey. Their mentorship was instrumental in helping me persevere and achieve my goals. I am especially thankful for their kindness in providing me with a comfortable workspace, and ensuring I felt supported throughout this process. Their care allowed me to thrive.

My sincere gratitude extends to the faculty of the MLDM program, whose exceptional expertise taught me a lot. I am particularly thankful to Dr. Habrard for his thoughtful supervision during my internship.

I also wish to acknowledge the brilliant researchers, hobbyists, and innovators who have contributed to the development of the technologies and concepts that were fundamental to my research.

To the friends I made during the MLDM program, I am profoundly appreciative of you. You were always there to answer my calls, and to bring laughter on difficult days—thank you.

Lastly, I extend my heartfelt thanks to my MLDM classmates. The more I learned about their passions, sacrifices, and hardships, the more my own passion grew. I felt that we all united in this journey of life, as people, as machine learners, taking small meaningful steps towards making the world a better place. It's the essence of what anyone who engages in research strives to do.

## References

Alharthi, A. G. and Alzahrani, S. M. Do it the transformer way: A comprehensive review of brain and vision transformers for autism spectrum disorder diagnosis and classification. *Computers in Biology and Medicine*, 167:107667, Dec 2023. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2023.107667. URL https://doi.org/10.1016/j.compbiomed.2023.107667. Epub 2023 Nov 3.

Bellec, P., Rosa-Neto, P., Lyttelton, O. C., Benali, H., and Evans, A. C. Multi-level bootstrap analysis of stable clusters in resting-state fmri. *NeuroImage*, 51(3):1126–1139, 2010. doi: https://doi.org/10.1016/j.neuroimage.2010.02.082. URL https://www.sciencedirect.com/science/article/pii/S1053811910002697.

Brain for AI Wiki. Visual cortex. https://brain-for-ai.fandom.com/wiki/Visual_cortex?file=Visual_cortex.png, n.d. Accessed: 2024-08-22.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., and Corlett, P. R. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry*, 3(3):243–250, 2016.

Chen, G., Ward, B. D., Xie, C., Li, W., Wu, Z., Jones, J. L., Franczak, M., Antuono, P., and Li, S.-J. Classification of alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional mr imaging. *Radiology*, 259(1):213–221, 2011. doi: 10.1148/radiol.10100734. URL https://doi.org/10.1148/radiol.10100734. PMID: 21248238.

Fombonne, E. Epidemiology of pervasive developmental disorders. *Pediatric Research*, 65(6):591–598, June 2009. doi: 10.1203/PDR.0b013e31819e7203.

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3 (11):e745–e750, 2021.

Harris, J. Leo kanner and autism: a 75-year perspective. *International review of psychiatry (Abingdon, England)*, 30: 1–15, 04 2018. doi: 10.1080/09540261.2018.1455646.

Hus, Y. and Segal, O. Challenges surrounding the diagnosis of autism in children. *Neuropsychiatric Disease and Treatment*, 17:3509–3529, 2021. doi: 10.2147/NDT.S282569. URL https://www.tandfonline.com/doi/abs/10.2147/NDT.S282569.

Jafri, M. J., Pearlson, G. D., Stevens, M., and Calhoun, V. D. A method for functional network connectivity among spatially independent resting-state components in schizophrenia. *NeuroImage*, 39(4):1666–1681, 2008. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2007.11.001. URL https://www.sciencedirect.com/science/article/pii/S1053811907010282.

Kanner, L. Autistic disturbances of affective contact, the nervous child, 2, 1943.

Kendall, M. G. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93, 06 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL https://doi.org/10.1093/biomet/30.1-2.81.

Kim, B.-H., Ye, J. C., and Kim, J.-J. Learning dynamic graph representation of brain connectome with spatio-temporal attention. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4314–4327. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/22785dd2577be2ce28ef79febe80db10-Paper.pdf.

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., Paolini, M., Chisholm, K., Kambeitz, J., Haidl, T., Schmidt, A., Gillam, J., Schultze-Lutter, F., Falkai, P., Reiser, M., Riecher-Rössler, A., Upthegrove, R., Hietala, J., Salokangas, R. K. R., Pantelis, C., Meisenzahl, E., Wood, S. J., Beque, D., Brambilla, P., and Borgwardt, S. Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis. *JAMA Psychiatry*, 75(11):1156–1172, 2018. ISSN 2168-6238. doi: 10.1001/jamapsychiatry.2018.2165.

Liu, H., Stufflebeam, S. M., Sepulcre, J., Hedden, T., and Buckner, R. L. Evidence from intrinsic activity that asymmetry of the human brain is controlled by multiple factors. *Proceedings of the National Academy of Sciences*, 106(48):20499–20503, December 2009. doi: 10.1073/pnas.0908073106. URL https://pnas.org/doi/full/10.1073/pnas.0908073106.

Lordan, R., Storni, C., and Benedictis, C. A. D. Autism spectrum disorders: Diagnosis and treatment. In Grabrucker, A. M. (ed.), *Autism Spectrum Disorders*, chapter 2. Exon Publications, Brisbane (AU), 2021. doi: 10.36255/exonpublications.autismspectrumdisorders.2021.diagnosis. URL https://www.ncbi.nlm.nih.gov/books/NBK573609/.

Lundberg, S. M. et al. *SHAP: SHapley Additive exPlanations Documentation*, 2024. URL https://shap.readthedocs.io/en/latest/. Accessed: 2024-08-22.

Maenner, M. J., Warren, Z., Williams, A. R., and et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2020. *MMWR Surveillance Summaries*, 72(No. SS-2):1–14, 2023. doi: 10.15585/mmwr.ss7202a1. URL http://dx.doi.org/10.15585/mmwr.ss7202a1.

Maser, J. and Akiskal, H. Spectrum concepts in major mental disorders. *The Psychiatric Clinics of North America*, 25(4):xi–xiii, December 2002. doi: 10.1016/S0193-953X(02)00034-5.

Minshew, N. J. and Williams, D. L. The new neurobiology of autism: Cortex, connectivity, and neuronal organization. *Archives of Neurology*, 64(7):945–950, 2007. doi: 10.1001/archneur.64.7.945. Erratum in: Arch Neurol. 2007 Oct;64(10):1464.

Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com, 2019. Available online at https://christophm.github.io/interpretable-ml-book/.

National Institute of Health. About autism. National Human Genome Research Institute, May 2019. URL https://www.genome.gov/Genetic-Disorders/Autism. Accessed: 16 April 2024.

Nilearn. Nilearn: Statistical analysis for neuroimaging in python—machine learning for neuroimaging. https://nilearn.github.io/index.html.

Pelphrey, K. A., Shultz, S., Hudac, C. M., and Vander Wyk, B. C. Research review: Constraining heterogeneity: The social brain and its development in autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, 52(6):631–644, 2011. doi: 10.1111/j.1469-7610.2010.02349.x.

Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., and Joliot, M. Automated anatomical labelling atlas 3. *NeuroImage*, 206:116189, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.116189. URL https://www.sciencedirect.com/science/article/pii/S1053811919307803.

Shapley, L. S. A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pp. 307–318. Princeton University Press, 1953.

Sharif, H. and Khan, R. A. A novel machine learning based framework for detection of autism spectrum disorder (asd). *Applied Artificial Intelligence*, 36(1):2004655, 2022. doi: 10.1080/08839514.2021.2004655. URL https://doi.org/10.1080/08839514.2021.2004655.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338, 2009. doi: 10.1136/bmj.b2393.

Subah, F. Z., Deb, K., Dhar, P. K., and Koshiba, T. A deep learning approach to predict autism spectrum disorder using multisite resting-state fmri. *Applied Sciences*, 11 (8), 2021. ISSN 2076-3417. doi: 10.3390/app11083636. URL https://www.mdpi.com/2076-3417/11/8/3636.

Vieira, S., Pinaya, W. H., and Mechelli, A. Using machine learning and structural neuroimaging to detect first episode psychosis: Reconsidering the evidence. *Schizophrenia Research*, 214:60–70, 2020.

Wing, L. The history of ideas on autism: Legends, myths and reality. *Autism*, 1(1):13–23, 1997. doi: 10.1177/1362361397011004. URL https://doi.org/10.1177/1362361397011004.

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., and Buckner, R. L. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, September 2011. doi: 10.1152/jn.00338.2011. URL https://doi.org/10.1152/jn.00338.2011. Epub 2011 Jun 8.