EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mathematisch-
Naturwissenschaftliche
Fakultät

Master Thesis

# Towards Interpretable Brain Biomarker Extraction using Deep Learning for fMRI Prediction


Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Lucas Mahler, lucas.mahler@student.uni-tuebingen.de
lucas.mahler@tuebingen.mpg.de


Bearbeitungszeitraum: 01.03.2023 - 01.09.2023

Gutachter: PD Dr. Gabriele Lohmann,
Max-Planck-Institut für biologische Kybernetik
Gutachter: Dr. Christian Baumgartner, Universität Tübingen

# Selbstständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Masterarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

---

Lucas Mahler (Matrikelnummer xxxxxxx), August 31, 2023

# Abstract

Mental disorders are a major public health concern, affecting millions of people worldwide. However, finding reliable biomarkers for diagnosis and treatment remains a challenge. Autism spectrum disorder (ASD) is a common psychiatric condition characterized by atypical cognitive, emotional, and social patterns. Timely and accurate diagnosis is critical for effective interventions and improved outcomes in individuals with ASD. In this study, we propose a novel Multi-Atlas Enhanced Transformer framework, METAFormer, for ASD classification. Our framework utilizes resting-state functional magnetic resonance imaging data from the ABIDE I dataset. METAFormer employs a multi-atlas approach, where flattened connectivity matrices from the AAL, CC200, and DOS160 atlases serve as input to transformer encoders. In particular, we show that self-supervised pre-training, which involves the reconstruction of masked values from the input, significantly improves classification performance without the need for additional or separate training data. Through cross-validation, we evaluate the proposed framework and show that it outperforms the state-of-the-art on the ABIDE I dataset, with an average accuracy of 83.7% and an AUC score of 0.832. We also perform a comprehensive analysis of the model's interpretability using several commonly used methods. Our quantitative evaluation of the quality of the explanations provided by these methods shows that DeepLIFT provides the most robust and faithful explanations. Using DeepLIFT, we extract the most important features for ASD classication which also consistetly align with current literature.

# Kurzfassung

Psychische Störungen stellen ein großes Problem für die öffentliche Gesundheit dar und betreffen weltweit Millionen von Menschen. Die Suche nach verlässlichen Biomarkern für Diagnose und Behandlung bleibt jedoch eine Herausforderung. Autismus (ASD) ist eine häufige psychische Störung, die durch atypische kognitive, emotionale und soziale Muster gekennzeichnet ist. Eine frühe und genaue Diagnose ist entscheidend für effektive Interventionen und bessere Behandlungsergebnisse bei Menschen mit ASD. In dieser Arbeit schlagen wir ein neuartiges Multi-Atlas Enhanced Transformer Framework, METAFormer, für die Klassifikation von ASD vor. Unser System verwendet funktionelle Magnetresonanztomographie-Daten aus dem ABIDE I Datensatz. METAFormer verwendet einen Multi-Atlas-Ansatz, bei dem Korrelationsmatrizen aus den Atlanten AAL, CC200 und DOS160 als Input für Transformer-Encoder dienen. Insbesondere zeigen wir, dass self-supervised Pretraining, das die Rekonstruktion maskierter Werte aus dem Input beinhaltet, die Klassifikationsperformance signifikant verbessert, ohne dass zusätzliche oder separate Trainingsdaten benötigt werden. Durch Cross-Validation evaluieren wir das vorgeschlagene System und zeigen, dass es den Stand der Technik im ABIDE I Datensatz mit einer durchschnittlichen Genauigkeit von 83.7% und einer AUC von 0.832 übertrifft. Darüber hinaus führen wir eine umfassende Analyse der Interpretierbarkeit des Modells unter Verwendung mehrerer gängiger Methoden durch. Unsere quantitative Bewertung der Erklärungsqualität dieser Methoden zeigt, dass DeepLIFT die robustesten und glaubwürdigsten Erklärungen liefert. Mit DeepLIFT extrahieren wir die wichtigsten Merkmale für die Klassifizierung von ASD, die auch mit der aktuellen Literatur übereinstimmen.

# Acknowledgments

I would like to express my sincere gratitude to my supervisors, Dr. Christian Baumgartner and Dr. Gabriele Lohmann, for their invaluable guidance and mentorship throughout this research. Their expertise has been instrumental in shaping the outcome of this thesis.

I would also like to thank my friends and colleagues for their support and collaborative spirit. Their insights and discussions have greatly enriched this project and made the journey more rewarding.

To my family, thank you for your unwavering belief in me. Your encouragement was my driving force.

# Contents

# 1 Introduction

## 1.1 Background and Context

Humans have always been curious about the mysteries of the human mind. From the early civilizations, pondering about the nature of mind and consciousness, to the ancient Egyptians attributing consciousness to the heart, and the Greeks speculating about the connection between the brain and the mind. It was not until the Renaissance that systematic anatomical studies began to provide insights into the physical substrate of human cognition. Pioneers such as Leonardo DaVinci sought to understand medieval psychology and to localize sensory and motor functions in the brain. His discoveries marked a sharp departure from medieval anatomical progress and embodied the dawn of the modern scientific era [1].

Until the 20th century, however, the only way to study the brain was invasively. In 1924, Hans Berger was the first to record human brain activity by placing electrodes on the scalp, marking a turning point that allowed researchers to circumvent the physical limitations of the human brain and observe brain activity in vivo [2]. The post-World War II era saw rapid advances in neuroimaging. Because the brain is composed mostly of soft tissue, X-rays couldn't be used effectively. Another milestone was the development of computed tomography (CT), pioneered by Oldendorf, Hounsfield, and Cormack in the early 1960s [3]. CT is an advanced medical imaging technique that uses X-rays to produce cross-sectional images of the human body. By analyzing the differential absorption of X-rays by different tissues, CT produces detailed visualizations [4]. However, CT has certain risks and limitations, such as ionizing radiation, potential risk of cancer, contrast agent reactions, limited soft tissue details, and inapplicability during pregnancy [5, 6, 7].

In 1946, both Bloch and Purcell independently discovered the phenomenon of nuclear magnetic resonance (NMR) [8]. NMR is a quantum-based spectroscopic method in which atomic nuclei with magnetic properties resonate when exposed to a strong external magnetic field, resulting in the absorption and emission of radiofrequency electromagnetic radiation. It became an important technique for non-destructive chemical analyses [8]. In 1971, Damadian [9] was the first to use NMR to detect tumors for medical applications. At the same time, Paul Lauterbur was working on the potential use of NMR in medicine and presented the first 2D NMR image of a water-filled object. NMR imaging uses the response of atomic nuclei to a magnetic field to reveal structures and properties. Hydrogen nuclei, which are abundant in the body, have a distinct magnetic behavior that lends itself to imaging. Nuclei

have a property called spin, which allows them to align parallel or antiparallel to a magnetic field. Hydrogen nuclei align due to energy differences between spin states. Radiofrequency (RF) pulses at the Larmor frequency excite the nuclei, causing the magnetization to rotate 90 degrees from its initial position. Gradients in the magnetic field induce a resonance at different frequencies across a slice, generating spatial information. A Fourier transform converts this data into a spatial map in the frequency domain. RF pulses and gradients create slices by exciting a specific plane of nuclei. Gradient fields cause dephasing across the slice, after which transverse magnetization is detected. The two-dimensional Fourier Transform is derived from NMR spectroscopy and uses phase shifts to encode spatial data [10].

In 1977, the first human subject was scanned using MRI, accelerating its commercialization into the 1980s. Since then, MRI has seen significant improvements, including higher field strengths for improved image resolution and signal-to-noise ratio, 3D imaging, faster acquisition times, and techniques to visualize specialized anatomical structures such as white matter tracts or blood vessels [11]. MRI plays a crucial role in modern science and medicine by providing non-invasive, high-resolution images that enable precise diagnoses, deeper insights into biological processes, and advances in understanding complex diseases and neurological functions.

In clinical practice, MRI has proven to be an invaluable tool in guiding the diagnosis and treatment of various medical conditions. One important area where MRI plays a critical role is in the assessment of acute ischemic stroke, the second leading cause of death worldwide. Available treatments for stroke require rapid administration [12]. MRI is essential in assessing stroke and determining the extent of brain damage. It helps differentiate between ischemic strokes (caused by reduced blood flow) and hemorrhagic strokes (caused by bleeding), which is essential for guiding treatment decisions [13]. The use of MRI as a diagnostic tool has led to significant advances in understanding the substrate and mechanisms underlying diseases such as multiple sclerosis. The development and evolution of focal lesions and diffuse damage can now be more precisely characterized. High-field and ultrahigh-field MRI techniques have facilitated the detection of gray and white matter damage and have provided insight into the topographic relationship between damage and venous blood vessels [14].

MRI has become a powerful tool for monitoring treatment, assessing safety, and predicting disease progression in multiple sclerosis [15]. MRI also plays an important role in assessing brain damage resulting from traumatic events such as concussions. It can detect contusions, diffuse axonal injury, and other structural abnormalities associated with traumatic brain injury (TBI) [16, 17, 18]. In addition, MRI serves as a primary tool for the diagnosis and characterization of brain tumors. It provides detailed information about tumor size, location, and composition, which aids in treatment planning and surgical guidance. Early diagnosis facilitated by MRI is critical for treatment outcomes and improving patient survival [19, 20]. In addition, MRI helps identify structural causes of epilepsy, such as cortical dysplasia, tumors, and vascular malformations. These findings guide treatment decisions, including

surgery for drug-resistant epilepsy [21, 22]. MRI finds widespread use in the diagnosis and monitoring of neurodegenerative diseases such as Alzheimer's disease [23, 24, 25], Parkinson's disease [26, 27, 28], and Huntington's disease [29]. It effectively reveals changes in brain volume, cortical thickness, and patterns of atrophy.

In the field of neurosurgery, MRI provides detailed preoperative images that assist neurosurgeons in planning procedures, identifying tumor boundaries, avoiding critical structures, and minimizing surgical risks [30, 31]. The sensitivity of MRI is generally equal to or superior to other imaging modalities in detecting abnormalities. In addition, MRI has the ability to detect clinically silent abnormalities or incidental findings [32].

MRI is a powerful tool capable of examining various tissue properties, including microstructure, metabolism, composition, and morphology. Despite its potential, few quantitative measures have been derived to discriminate between healthy and diseased subjects [33].

Biomarkers, a portmanteau of biological markers, are objective indications of a medical condition that can be accurately and reproducibly observed from outside the patient [34]. In the context of the brain, a biomarker could be represented by metrics such as hippocampal volume or cortical thickness. The utility and relevance of biomarkers [33] is evident in several contexts: 1) In basic science, biomarkers enhance the understanding of normal biological processes and disease progression by revealing features not readily accessible by other means. 2) In drug or device development, biomarkers serve as safety, diagnostic, prognostic, response, or predictive indicators. 3) In clinical practice, biomarkers inform patient management decisions.

The journey of a biomarker from discovery to patient management in daily clinical practice involves several steps: Assay development, technical validation, biological validation, clinical validation, clinical utility and cost effectiveness. In neurodegenerative diseases, for example, changes in neuroanatomy can be used to predict the onset, development or progression of symptoms. In diseases such as Alzheimer's disease, neuropathology is believed to develop before the onset of dementia, motivating the search for predictive biomarkers. A longitudinal study by [35] found that cortical thinning occurs in asymptomatic individuals, suggesting it as a potential biomarker for early neurodegeneration. Similarly, using structural MRI, [25] identified the baseline volume of the entorhinal cortex and its rate of decline as possible predictors of the onset of Alzheimer's disease.

Attempts to identify biomarkers in Parkinson's disease have led to a better understanding of its underlying causes. Changes in whole-brain cortical and subcortical gray matter identified by [36] predicted the development of freezing of gait. Another study by [37] concluded that biomarkers for various stages of Parkinson's disease are available and potentially useful when combined in multimodal clinical assessments.

Uncovering biomarkers for psychiatric disorders has been challenging, primarily due to the lack of prominent structural differences between patients and healthy controls [38]. Functional Magnetic Resonance Imaging (fMRI) emerged in the early 1990s as a non-invasive method for visualizing brain activity. It uses changes in blood oxygenation to map brain activity based on blood oxygenation level-dependent contrast [39, 40]. This technological advance opened new avenues for scientific exploration by allowing the study of brain function over time with high spatial resolution. In general, two main types of fMRI experiments can be distinguished: 1) Task-based fMRI, in which a subject performs a specific task (such as attending to stimuli) while being scanned. 2) Resting-state fMRI (rs-fMRI), which measures the subject's brain activity under standard conditions.

fMRI has become an invaluable tool in cognitive neuroscience, psychology and medical research. It can be used to explore brain networks, uncover the neural basis of various disorders, and improve our understanding of brain-behavior relationships. Functional connectivity (FC) analysis [41] plays a crucial role in fMRI data analysis, as it examines the statistical dependencies and temporal correlations between different brain regions. Rather than looking at isolated abnormalities in specific regions, brain disorders often result from disrupted communication and abnormal interactions between regions. FC analysis allows researchers to explore the network-level abnormalities associated with different disorders. This analysis involves dividing the brain into regions of interest (ROIs) and quantifying the correlations between their time series using various mathematical measures.

## 1.2 Motivation

Mental disorders have emerged as a major global health problem, contributing to a significant proportion of disability worldwide. According to the GBD 2019 [42], roughly 5% of the global burden of disease is attributable to mental disorders and 16% of the global diisability-adjusted life years (DALYs) are due to mental disorders. The most common mental disorders include depression, anxiety, schizophrenia, bipolar disorder and autism spectrum disorder (ASD) [43]. These disorders are characterized by a complex interplay of abnormal thoughts, perceptions, emotions, behaviors, and interpersonal relationships. The multifaceted nature of these conditions results from a combination of genetic, biological, psychological and environmental factors, and their etiology remains incompletely understood.

However, advances in functional neuroimaging techniques, particularly fMRI, have revealed altered patterns of FC as potential biomarkers. In contrast to structural changes, FC variations provide insights into the underlying neural dynamics of psychiatric disorders.

The goal of this thesis is to address the critical challenge of extracting interpretable brain biomarkers from fMRI data using deep learning for the prediction of psychiatric

disorders, with a particular focus on autism spectrum disorder. This disorder in particular is in dire need of improved diagnostic methods, as current diagnoses rely solely on behavioral symptoms and lack objective biomarkers. Autism Spectrum Disorder is a term used to describe early-onset deficits in social communication and repetitive sensorimotor behaviors [44]. Altered brain development and early neural reorganization are thought to contribute to ASD. Currently, there are no reliable biomarkers for ASD, so diagnoses must be based solely on behavioral observations. The American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) defines ASD as a spectrum disorder with two core domains: Social communication and restricted, repetitive, or unusual sensorimotor behaviors. This definition includes subtypes such as Asperger's syndrome and pervasive developmental disorder under a single umbrella.

ASD is often accompanied by other psychiatric disorders such as attention deficit hyperactivity disorder [45]. In children, two key elements of diagnosis are a detailed developmental history and the child's interactions with parents and unfamiliar adults during structured and unstructured assessments. In addition, tools such as the Autism Diagnostic Observation Schedule [46] and the Autism Diagnostic Interview-Revised [47] have been developed for diagnosis. However, these assessments have been criticized for their lack of objectivity and transparency [48]. Currently, the diagnosis of ASD relies heavily on behavioral observations and historical information, which is challenging and time consuming.

Given these limitations, there is an urgent need for a rapid, cost-effective, and objective diagnostic method that can accurately identify ASD. Such a method would lead to more timely interventions and improved outcomes for individuals affected by ASD.

Several factors make autism spectrum disorder an ideal focus for this research. First, its prevalence, with 1 in 160 children affected worldwide, highlights the importance of developing more accurate and reliable diagnostic tools. Second, the lack of objective biomarkers for ASD limits the ability to fully understand the underlying neurobiological mechanisms of the disorder. In addition, the availability of high-quality datasets with large numbers of subjects from different sites provides an excellent opportunity to develop robust and generalizable models. Previous work in this area has also suggested the potential to achieve promising results through the application of deep learning methods.

**Research Question.** The primary research questions driving this thesis are whether we can train a deep learning model to accurately predict ASD from resting-state fMRI data, and whether we can make progress in extracting interpretable and meaningful biomarkers from the learned representations of the model. To effectively address our research question, we will follow common data preprocessing strategies and utilize preprocessed connectome data. We will then propose a

novel multi-atlas based transformer architecture as well as an established reference architecture for predicting ASD from resting-state fMRI data. Using different feature attribution methods, we aim to gain insight into the inner workings of the models. Using quantitative evaluation metrics, we will compare the quality of the provided explanations and determine the most appropriate feature attribution method for extracting potential biomarkers.

By addressing these critical questions, this thesis aims to provide a more accurate and reliable diagnosis of ASD and to gain insights into the underlying pathology from the learned representations of the models. Ultimately, the proposed research could have significant implications for improving our understanding of Autism Spectrum Disorder and potentially other psychiatric disorders, leading to improved early diagnosis and personalized treatment strategies.

## 1.3 Related Works

As deep learning models are increasingly adopted across domains, the need for interpretable and explainable approaches has increased dramatically [49]. Not only do regulations such as the European General Data Protection Regulation make black-box approaches almost impossible to use in business, but there's also a lack of trust and understandability. This is even more pronounced in the medical field, where explainable and interpretable AI would help promote transparency and trust in AI-based applications [50]. Healthcare provides fertile ground for the impactful application of AI, with large amounts of data readily available and potentially enormous benefits [51, 52]. There have been many recent advances in interpretable and explainable AI methods. For example, feature attribution methods attempt to identify which features contribute most to model predictions, such as DeepLIFT [53], Integrated Gradients [54]. Gradient-based methods such as Grad-CAM [55] determine importance by analyzing gradients with respect to specific input regions. Shapley value-based methods [56] assign importance scores to each feature based on its contribution to the output prediction by considering its contribution to all possible feature combinations. Local Interpretable Model-Agnostic Explanations (LIME) [57] create interpretable surrogate models for local explanations. Counterfactual explanations [58] provide examples of inputs that, if changed, would lead to different predictions. The question, however, is how good the explanation provided by a particular method actually is. Qualitative, often subjective, measures of explanatory quality are often insufficient for objective evaluation. A number of metrics have been proposed, and the Explanable AI toolkit Quantus [59] groups these evaluations into six categories:

1. Fidelity: Determines how closely a given explanation follows the prediction of a model such as [60, 61].

2. Robustness: measures how stable given explanations are to small changes in

the input, such as [62, 60].

3. Localization: determines whether there is a region of interest (ROI) of high importance [63, 64].
4. Complexity: measures how many features are needed to explain a prediction [65].
5. Randomization: measures how the quality of the explanation decreases with the introduction of randomness into the model parameters [66].
6. Axiomatic: whether certain axioms are satisfied by the explanation method [54].

**Interpretability in Medical Imaging.** Leveraging advances in computer vision over the past decade, explicable and interpretable deep learning has found great acceptance in the medical imaging field. Applications in tumor detection, classification, and segmentation are numerous. For example, [67] use post-hoc explainability methods on an autoencoding neural network to segment brain tumors and provide explanations using class activation maps. Similarly, [68] use a novel visual transformer model for the same task. [69] extract important regions from a CNN trained on prostate tumor segmentation using saliency maps.

Explanatory AI has also found applications in cancer. For example, [70] propose a case-based reasoning approach that automatically provides quantitative and qualitative visual explanations to help medical experts understand case similarities. [71] extracted important regions for prostate cancer classification from deep learning models using LIME [57]. In another work, [72] proposes a pathology whole slide diagnosis approach for carcinoma identification. A CNN detects tumor regions, which are characterized by a subsequent CNN; a recurrent neural network provides a textual description. Thus, the final report consists of detected tumor regions with visually highlighted ROIs and a textual description.

Generative AI has also contributed to interesting developments in explicable AI. [73] use StyleGAN to propose a new interpretability method that shows how the input image would need to be modified to produce different predictions. [74] use generative adversarial networks to learn a map-generating function from unlabeled training data that can generate medical disease effect maps for mild cognitive impairment and Alzheimer's disease.

In addition to these areas, other segments of medical imaging have also seen successful applications of explicable and interpretable AI, such as medical image retrieval [75, 76], or diagnosis of COVID-19 and pneumonia [77, 78].

**ASD Classification.** In recent years, machine learning approaches have been widely applied to the problem of ASD classification using rs-fMRI data. The majority of these studies use functional connectivity obtained from a predefined atlas as

input to their classifiers. A considerable amount of work has used classical machine learning algorithms such as support vector machines and logistic regression to classify ASD [79]. However, these methods have limitations because they are typically applied to small datasets with specific protocols and fixed scanner parameters, which may not adequately capture the heterogeneity present in clinical data. 3D convolutional neural networks [80, 81, 82] have also been applied to preprocessed fMRI data, [83] have used 2D CNNs on preprocessed fMRI data. However, these approaches are also limited by the fact that they used only small homogeneous datasets.

More recent work has attempted to overcome the homogeneity limitation and has used deep learning approaches to classify ASD based on connectomes. Multilayer perceptrons are well suited to vector-based representations of connectomes and have thus seen some use in ASD classification [84, 85]. Graph convolutional models are also well suited and have yielded high accuracies [86, 87]. Other approaches have used 1D CNNs [88], or variants of recurrent neural networks [89, 90], and probabilistic neural networks have also been proposed [91].

However, ASD classification is not limited to fMRI data, and there has been work using, for example, EEG [92] or more novel imaging approaches such as functional near-infrared spectroscopy [93].

**Interpretability in ASD Classification.** [94] used an ensemble of 300 CNNs for ASD prediction, extracted saliency maps, and performed feature visualization of the learned class representation, revealing temporal and cerebellar connections to be of high importance. [95] trained 3D models on ABIDE using Guided Grad-CAM [55] to extract activation maps. They found that left lateralized regions involved in language processing were highly relevant. [96] extracted class activation maps from a 3D resnet trained on structural data. They identified important regions near the hippocampus, corpus callosum, thalamus, and amygdala. In a task-fMRI study, [80] used a custom feature ablation experiment on the 2CC3D CNN [97] to extract meaningful features. They found that visual regions were the most discriminative between ASD and typical controls. [98] used a spatiotemporal deep neural network for ASD classification. Their results showed that regions associated with the default mode network were most important for classification. Also using 3D CNNs, [99] incorporated subject-specific Group Independent Component Analysis-derived functional network maps to improve the interpretability of their approach. [100] used multi-view graph convolutional neural networks to classify ASD and extracted important regions from multi-task graph embeddings constrained by prior knowledge of network structures associated with ASD. [101] used graph neural networks for ASD classification on ABIDE I. They used network community clustering and graph saliency mapping to identify important regions and connections. Similarly, [96] incorporated invertible blocks into their GCN, allowing them to reconstruct the dynamic features of the input from the output of the network.

## 1.4 Overview & Contributions

In chapter 2, we first provide an overview of the interpretability methods under consideration and two methods for quantitatively assessing the quality of the explanations they provide. We then provide a brief description of the dataset used in this work and the preprocessing steps we applied in chapter 3. In chapter 4 we describe the reference architecture and present a novel multi-atlas enhanced transformer architecture, METAFormer, and a pretraining strategy. In chapter 5 we describe the ASD classification task, the experimental setup, the evaluation regime, and the results of the classification task. Then, in chapter 6, we present the design of our interpretability experiments, their results, and the results of the quantitative and qualitative analyses. Finally, in chapter 7 we discuss the results of our experiments, address limitations and provide an outlook for future work, and end with a concluding section.

**Contributions.** The contributions of this work are as follows:

1. We propose a novel multi-atlas enhanced transformer architecture, METAFormer, which outperforms the state-of-the-art in ASD classification on ABIDE-I.
2. We present a pre-training strategy for METAFormer that significantly improves performance.
3. We perform a comprehensive interpretability analysis to extract potential biomarkers for ASD and quantitatively evaluate the quality of the explanations provided by the interpretability methods to determine the most appropriate method for our task.
4. We contextualize the extracted potential biomarkers with the current literature.

# 2 Interpretability Methods

In this section, we present commonly used interpretability methods to assign importance to features. Sections 2.1 to 2.5 provide a brief introduction and theoretical background. Section 2.6 will introduce an additional method to visualize features. One question that naturally arises when using these methods is how good the explanation provided by a method really is. This will be addressed by section 2.7.

## 2.1 Feature Ablation

Feature ablation is a perturbation-based approach used to compute feature weights in neural networks. The main idea behind feature ablation is to systematically remove or replace each input feature with a given baseline or reference, and then compute the difference in the model's output with respect to the original input.

The feature ablation process can be described as follows: Each feature in a given input is independently perturbed by replacing it with a previously selected baseline value The model output is then recorded for each perturbed input. By comparing the model output to the original output, the importance or contribution of each feature can be quantified based on how much the model prediction changes due to feature ablation.

Feature ablation provides a straightforward approach to calculating feature weights and is conceptually simple to understand. However, for models with large feature vectors, feature ablation can be computationally expensive because it requires running the model multiple times, once for each perturbed input. Despite its computational cost, feature ablation remains a useful technique in interpretability studies, especially for understanding the relative importance of individual features in the model's decision-making process.

## 2.2 Integrated Gradients

Integrated Gradients [54] is a method for attributing the predictions of a deep neural network to its input features. The key idea behind Integrated Gradients is to compute line integral over a linear function between a baseline input $x'$ and an input of interest $x$ of the corresponding gradients. By cumulating these gradients, this provides a way to assign attributions to individual features in the input data.

Formally, let us consider a function $f : \mathbb{R}^n \to [0,1]$ that represents a deep neural network, and an input $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. An attribution of the prediction at input $x$ relative to a baseline input $x'$ is represented as a vector $a_f(x, x') = (a_1, \ldots, a_n) \in \mathbb{R}^n$, where $a_i$ is the contribution of feature $x_i$ to the prediction $f(x)$.

The Integrated Gradient for the $i$-th dimension of an input $x$ with respect to a baseline $x'$ is defined as follows:

$$\text{IntegratedGradients}_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{2.1}$$

Here, $\frac{\partial f(x)}{\partial x_i}$ represents the gradient of the prediction function $f(x)$ with respect to the $i$-th dimension of the input.

The intuition behind Integrated Gradients is to approximate the change in the model's prediction from the baseline $x'$ to the input $x$ along the straight line path connecting the two. By integrating the gradients along this line, Integrated Gradients provides a way to attribute the contribution of each feature to the final prediction.

Integrated Gradients satisfies both the sensitivity and implementation invariance axioms. It ensures that when attributing the prediction to input features, if two inputs and baselines differ in a feature but have different predictions, the differing feature receives a non-zero attribution. In addition, Integrated Gradients provides consistent attributions for functionally equivalent networks, thereby avoiding sensitivity to unimportant aspects of the model architecture.

## 2.3 Saliency

Saliency maps are a visualization technique used to understand the influence of individual pixels in an image on the classification decision made by a Convolutional Neural Network (CNN). Unlike using class posteriors returned by the softmax layer, saliency maps employ unnormalized class scores. The maximization of the class posterior probability for a specific class can be achieved by minimizing the scores of other classes.

**Image-Specific Class Saliency Visualization.** Given an image $I_0$, a target class $c$, and a classification CNN with a class score function $S_c(I)$, the goal is to rank the pixels of $I_0$ based on their influence on the score $S_c(I_0)$. The magnitude of the derivative with respect to the input image indicates which pixels need to be changed the least to induce a maximal change in the output class score. It is reasonable to expect that these pixels correspond to the location of the object in the image.

**Class Saliency Extraction.** The class saliency map $M \in \mathbb{R}^{m \times n}$ is computed by finding the derivative $w$ through back-propagation. The saliency map is then obtained by rearranging the elements of the vector $w$. To derive a single class saliency value for each pixel $(i, j)$, the maximum magnitude of $w$ across all channels is taken: $M_{i,j} = \max_c |w_{h(i,j,c)}|$, where $h(i, j, c)$ denotes the index of the element corresponding to pixel $(i, j)$ in channel $c$ of $w$.

**Relation to Deconvolutional Networks.** Saliency maps are related to Deconvolutional Neural Networks [102]. Deconvolutional networks aim to reconstruct the input of the $n$-th layer, denoted as $x_n$, by computing the gradient of the visualized neuron activity $f$ with respect to $x_n$. Thus, Deconvolutional Networks effectively correspond to gradient backpropagation through the CNN.

Except for the ReLU layer, computing an approximate feature map reconstruction $R_n$ using Deconvolutional Networks is equivalent to computing the derivative $\partial f / \partial x_n$ using backpropagation. This implies that gradient-based visualization, such as saliency maps, can be seen as a generalization of Deconvolutional Networks as it can be applied to any layer, not just convolutional ones.

## 2.4 DeepLIFT

Neural networks have often been considered as black boxes, creating a barrier to their adoption in certain domains due to the lack of interpretability. DeepLIFT [53] is a proposed method that addresses this issue by assigning importance scores to the inputs of a neural network, shedding light on the features that significantly influence the model's output.

One of the unique aspects of DeepLIFT is that it frames the question of feature importance in terms of differences from a baseline state. This baseline input serves as a default or neutral input, chosen based on the problem at hand. DeepLIFT explains the difference in the model's output concerning this baseline output in terms of the difference in input from the baseline input. The summation-to-delta property plays a crucial role in assigning contribution scores $C_{\Delta x_i \Delta t}$:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t \tag{2.2}$$

Here, $t$ represents some target output neuron of interest, $x_1, x_2, ..., x_n$ are neurons in intermediate layers that are necessary and sufficient to compute $t$, and $\Delta t$ is the difference from the baseline, i.e., $\Delta t = t - t^0$. The term $C_{\Delta x_i \Delta t}$ represents the amount of difference-from-baseline in $t$ that is attributed to the difference-from-baseline of $x_i$. Notably, $C_{\Delta x_i \Delta t}$ can be nonzero even when $\frac{\partial t}{\partial x_i} = 0$, which circumvents

the fundamental limitations of gradients. Neurons can transmit information even when their gradient is zero, and DeepLIFT benefits from the continuous nature of differences-from-baseline, avoiding the discontinuities caused by bias terms.

**Multipliers and Chain Rule.** DeepLIFT introduces the concept of multipliers, defined as:

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x_i \Delta t}}{\Delta x} \tag{2.3}$$

The multiplier $m_{\Delta x \Delta t}$ represents the contribution of $\Delta x$ to $\Delta t$ divided by $\Delta x$. It is analogous to the idea of partial derivatives but computed over finite differences instead of infinitesimal differences.

DeepLIFT utilizes the chain rule for multipliers, which can be expressed as:

$$m_{\Delta x_i \Delta t} = \sum_j m_{\Delta x_i \Delta y_j} \cdot m_{\Delta y_j \Delta t} \tag{2.4}$$

Here, $x_1, ..., x_n$ are neurons in the input layer, $y_1, ..., y_n$ are neurons in a hidden layer, and $t$ is the target neuron. Given multipliers for each neuron to its immediate successor, multipliers for any neuron to a given target can be efficiently computed via backpropagation.

**Baseline Values.** The baseline of a neuron is its activation on the baseline input. For a neuron $y$ with inputs $x_1, x_2, ...$, where $y = f(x_1, x_2, ...)$, and given baseline activations $x_1^0, x_2^0, ...$ of inputs, the baseline activation $y^0$ of the output is:

$$y^0 = f\left(x_1^0, x_2^0, ...\right) \tag{2.5}$$

baselines for all neurons can be found by choosing a baseline input and forward-propagating activations through the network. The choice of baseline is critical to obtain insightful results from DeepLIFT and often relies on domain-specific knowledge.

In some situations, it is essential to treat positive and negative contributions differently:

$$\Delta y = \Delta y^+ + \Delta y^- \tag{2.6}$$

$$C_{\Delta y \Delta t} = C_{\Delta y^+ \Delta t} + C_{\Delta y^- \Delta t} \tag{2.7}$$

**Rescale Rule.** DeepLIFT applies the rescale rule to nonlinear transformations such as ReLU. For a nonlinear transformation $y = f(x)$ with $x$ as input and because $y$ has only one input, the summation-to-delta property is utilized, resulting in $C_{\Delta x \Delta y} = \Delta y$ and $m_{\Delta x \Delta y} = \frac{\Delta y}{\Delta x}$.

Thus, for positive and negative contributions:

$$\Delta y^+ = \frac{\Delta y}{\Delta x} \Delta x^+ = C_{\Delta x^+ \Delta y^+} \tag{2.8}$$

$$\Delta y^- = \frac{\Delta y}{\Delta x} \Delta x^- = C_{\Delta x^- \Delta y^-} \tag{2.9}$$

Leading to:

$$m_{\Delta x^+ \Delta y^+} = m_{\Delta x^- \Delta y^-} = m_{\Delta x \Delta y} = \frac{\Delta y}{\Delta x} \tag{2.10}$$

**Target Layer.** To compute contributions of the final layer preceding the softmax/sigmoid activation functions and avoid attenuation caused by the summation-to-delta property, DeepLIFT normalizes the contributions to the linear layer by subtracting the mean contribution to all classes:

$$C'_{\Delta x \Delta c_i} = C_{\Delta x \Delta c_i} - \frac{1}{n} \sum_{j=1}^{n} C_{\Delta x \Delta c_j} \tag{2.11}$$

Here, $n$ is the number of classes, $C_{\Delta x \Delta c_i}$ is the unnormalized contribution to class $c_i$ in the linear layer, and $C'_{\Delta x \Delta c_i}$ is the normalized contribution.

## 2.5 Shapley Additive Explanations

We can explain the predictions of ML models by assuming that each input feature is a player in a game and the prediction is the payoff. Shapley values, introduced by [103], is a coalitional game theory method that formalizes how to distribute the payoff fairly among players. And in terms of machine learning, each feature in a given model input is considered a player in a game, that is, the task of the model, which cooperates with other features, called coalitions, to produce a payoff, i.e., the output value of the model. For each coalition, the contribution of each feature is calculated by comparing the model's output when the feature is present in the coalition to the output when it's not. Thus, the contribution of a feature in a coalition is the difference between the model output with and without that

feature. In other words, the Shapley value is the average marginal contribution of a feature across all possible coalitions [104]. However, since all possible coalitions of feature values must be computed for each feature to obtain the exact Shapley value, this becomes computationally expensive very quickly. [56] proposed a Monte Carlo based approximation to compute Shapley values $\Phi_j$ for a given feature $j$ more efficiently:

$$\Phi_j = \frac{1}{M} \sum_{m=1}^{M} \left( f\left(x_{+j}^m\right) - f\left(x_{-j}^m\right) \right) \tag{2.12}$$

where $f\left(x_{+j}^m\right)$ is the model output for a given input $x$, where the feature values preceding position $j$ are taken from the original input $x$ and the subsequent values, excluding $x_j$, are taken from a randomly sampled data point $z$. Conversely, for $f\left(x_{-j}^m\right)$, all feature values following position $j$ are taken from the randomly sampled data point $z$, including position $j$. This is done for a total of $M$ iterations for each feature to obtain Shapley values for all features for a current input $x$.

### 2.5.1 SHAP Values

Shapley Additive ExPlanation (SHAP) values [105] are a unified approach to explain the output of any machine learning model. SHAP rephrases Shapley values as additive feature attribution method to explain individual predictions. According to [104] the SHAP explanation can be defined as:

$$g(z') = \Phi_0 + \sum_{j=1}^{M} \Phi_j z'_j \tag{2.13}$$

where, $g$ is the explanation model and $z' \in \{0,1\}^M$ is the coalition vector, where an entry of 1 represents a feature being present and 0 represents the absence of a feature. $M$ is the maximum size of a coalition and $\Phi_j \in \mathbb{R}$ is the Shapley value, or feature attribution, for feature $j$. To compute the Shapley values, only some features are present and some are absent. As the explanation model is a linear model of coalitions it is easier to calculate $\Phi$ as for a given input of interest $x$ the coalition vector $x'$ contains only 1's and thus the formula 2.13 becomes:

$$g(x') = \Phi_0 + \sum_{j=1}^{M} \Phi_j \tag{2.14}$$

### 2.5.2 DeepLIFT-SHAP

DeepLIFT-SHAP, or DeepSHAP [105], uses additional knowledge about the compositional nature of deep networks to approximate Shapley values more efficiently.

This is done under the assumption that the input features are independent and that the deep model is linear. DeepLIFT linearizes the nonlinear components of the neural network by backpropagation rules as shown in equation 2.10.

Since DeepLIFT is an additive feature attribution, it is a suitable compositional approximation of Shapley scores. DeepLIFT-SHAP computes Shapley values for smaller components of the network and then combines them to obtain Shapley values for the entire network. This composition approach allows for fast computation of Shapley values for the entire model assuming the components have been linearized.

### 2.5.3 GradientSHAP

GradientSHAP[105] is an extension of integrated gradients. It uses the gradients of the model output with respect to the input features to approximate Shapley values.

To compute the expected values of the model output, we use a baseline that represents an estimate of the expected value of the model output, which serves as a starting point for computing the Shapley values. Given the baseline, we compute the expected value of the model output, denoted as $E[f(x)]$, where $f(x)$ is the output of the model for input $x$. This expected value represents the average prediction of the model over the baseline. For a given input instance $x$, we compute the gradients of the model output with respect to the input features To approximate the Shapley values, we linearize the model output around the expected values using the computed gradients. Therefore, the approximation of the Shapley value for each feature can be computed as

$$\Phi_i = \sum_{j=1}^{N} (x_j - E[x_j]) \cdot \nabla_{x_j} f(x) \tag{2.15}$$

where $\Phi_i$ is the Shapley value for the $i$-th feature, $N$ is the number of features, $x_j$ is the value of the $j$-th feature for the input instance $x$, and $E[x_j]$ is the expected value of the $j^{th}$ feature calculated from the baseline dataset.

This gives an approximation of the Shapley values for each feature. To improve the accuracy, we can use different baseline instances for each feature and average the results. This can be done by sampling multiple times from the baseline distribution and computing the Shapley values for each sample, then averaging the results. This process is called Monte Carlo sampling.

## 2.6 Feature Visualization

A conventional approach for qualitatively comparing features extracted by the first layer of a deep neural network (DNN) involves examining the learned filters, which are essentially the linear weights in the input-to-first-layer weight matrix. However, this technique is limited to the first layer and lacks generality.

The objective of feature visualization, first proposed by Erhan et al.[106] is to explore methods for visualizing the computations performed by units in arbitrary layers of a deep network. This visualization should ideally be in the input space, efficient to compute, and broadly applicable across various layers.

A significant finding in this context is that the response of an internal unit to input images appears to be (almost) unimodal. This implies that each unit tends to respond strongly to specific input patterns, leading to distinct visualizations.

**Maximizing Activation.** The core idea of feature visualization is to search for input patterns of bounded norm that maximize the activation of a given hidden unit. Since the activation function of a unit in the first layer is a linear function of the input, the input pattern is directly proportional to the filter itself. This stems from the understanding that the pattern to which a unit responds maximally could serve as a good first-order representation of the unit's functionality.

To achieve this, one can look for input samples, either from the training or test set, that maximally excite a given unit. This raises the challenge of determining the number of samples to retain for each unit and devising a method to combine these samples in a meaningful way. Ideally, identifying commonalities among these samples is desired.

To generalize this approach, the problem is treated as an optimization task: maximize the activation of a specific unit. Mathematically, this can be expressed as:

$$x^{+} = \arg\max_{\|x\|=\phi} h_{ij}(\Theta, x) \tag{2.16}$$

where $h_{ij}(\Theta, x)$ represents the activation of the unit. Although this optimization problem is non-convex, it can be approached by performing gradient ascent in input space. The idea is to compute the gradient of $h_{ij}(\Theta, x)$ with respect to $x$ and move $x$ in the direction of this gradient.

This process can result in two possible scenarios: either the same minimum is found when starting from different initializations, or two or more local minima are identified. In both cases, the unit can be characterized by the found minima.

This visualization technique is applicable to any network where gradients can be calculated for, and it offers insights into the function of individual units across various layers.

## 2.7 Quantitative Evaluation

After obtaining importance scores for each feature using the various attribution methods described in sections 2.1 through 2.5, a crucial question arises: How reliable

and robust are these attributions? The evaluation of attribution methods can be divided into two classes: objective measures and subjective measures. While the notion of attribution is inherently human-centric, most attribution evaluations have been subjective, such as qualitative displays of attribution examples and crowd-sourced evaluations of human satisfaction with attributions. However, objective measures are more theoretically sound and can help improve attributions by optimizing their objective measures.

### 2.7.1 Explanation Infidelity

The efficacy of an explanation can be measured by its capacity to accurately encapsulate variations in the predictor function when subjected to significant perturbations. Infidelity [60] serves as a metric for evaluating this accuracy, quantified as the mean squared error between the magnitudes of input perturbations' influence on model explanations and the corresponding alterations in the prediction function.

This concept can be formalized within the context of a general supervised learning framework: Consider an input space $X \subseteq \mathbb{R}^d$ and an output space $Y \subseteq \mathbb{R}$, where a black-box predictor $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ makes predictions $f(x)$ for a given test input $x \in \mathbb{R}^d$. Feature attribution is defined by a function $\Phi : F \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$, which, given the black-box predictor $f$ and a test point $x$, assigns importance scores $\Phi(f, x)$ to individual input features.

Given a black-box function $f$, an explanation functional $\Phi$, and a random variable $I \in \mathbb{R}^d$ with a probability measure $\mu_I$ representing relevant perturbations, infidelity can be mathematically described as follows:

$$INFD(\Phi, f, x) = \mathbb{E}_{I \sim \mu_I}\left[I^T \Phi(f, x) - (f(x) - f(x - I))\right]^2 \tag{2.17}$$

The central objective of infidelity is to ascertain the explanation's ability to accurately model the impact of a perturbation $I$ applied to a specific test input $x$, resulting in the modified input $x - I$. Ideally, the model should exhibit sensitivity to such perturbations, leading to an output alteration of $f(x) - f(x - I)$. The infidelity metric, drawing inspiration from Taylor Series approximations, necessitates that the inner product of the explanation and the perturbation, $I^T \Phi(f, x)$, approximates the model's responsiveness. As the perturbation is inherently stochastic, infidelity involves an expectation computation with respect to this random perturbation.

### 2.7.2 Explanation Sensitivity

One important aspect in evaluating an attribution is its sensitivity [60]. Sensitivity measures the degree to which an attribution is affected by insignificant perturbations from a given test point. It is natural to desire attributions with low sensitivity, as high sensitivity could lead to differing explanations with minor variations in input,

potentially leading to distrust in the attributions. Lower sensitivity can be viewed as a notion of simplicity, which is desirable in explanations.

To measure sensitivity, we can consider the gradient of the attribution function with respect to the input:

$$[\nabla_x \Phi(f(x))]_j = \lim_{\epsilon \to 0} \frac{\Phi(f(x + \epsilon e_j)) - \Phi(f(x))}{\epsilon} \tag{2.18}$$

where $j \in \{1, ..., d\}$, and $e_j \in \mathbb{R}^d$ is the $j$-th coordinate basis vector with the $j$-th entry being one and all others zero. This gradient quantifies how the explanation changes as the input is varied infinitesimally along each coordinate direction. We can then compute a scalar-valued summary of the sensitivity vector as $\|\nabla_x \Phi(f(x))\|$. For a more robust measure, we can consider a locally uniform bound:

$$SENS_{GRAD}(\Phi, f, x, r) = \sup_{\|\delta\| \leq r} \|\nabla_x \Phi(x + \delta)\| \tag{2.19}$$

which is closely related to local Lipschitz continuity [107] around $x$:

$$SENS_{LIPS}(\Phi, f, x, r) = \sup_{\|\delta\| \leq r} \frac{\|\Phi(x) - \Phi(x + \delta)\|}{\|\delta\|} \tag{2.20}$$

If an attribution has locally uniformly bounded gradients, it is also locally Lipschitz continuous. However, local Lipschitz continuity can be unbounded in deep networks, making max-sensitivity a more robust estimation:

$$SENS_{MAX}(\Phi, f, x, r) = \max_{\|y - x\| \leq r} \|\Phi(f, y) - \Phi(f, x)\| \tag{2.21}$$

Moreover, if an attribution score is bounded, the max-sensitivity is always finite, making it more reliable for estimation through Monte Carlo sampling.

# 3 Data

The following chapter provides an overview of the data used and the preprocessing applied in this paper. Section 3.1 first describes the ABIDE I dataset [108] and provides an overview of the demographics of the dataset. Section 3.2 describes the preprocessing pipeline used to obtain the preprocessed data from the raw ABIDE I dataset. Section 3.3 describes the three atlases used in this thesis to parcellate the brain into regions of interest. Finally, section 3.4 describes the computation of functional connectivity between the regions of interest.

## 3.1 ABIDE Dataset

Our experiments are conducted on the ABIDE I dataset [108], which is a publicly available dataset containing both structural MRI and rs-fMRI data from individuals with ASD and typical controls (TC) from 17 different research sites. The raw data set includes a total of 1112 subjects, of which 539 are diagnosed with ASD and 573 are TC. Subjects range in age from 7 to 64 years, with a median age of 14.7 years across all groups. The ABIDE I dataset is considered one of the most comprehensive and widely used datasets in the field, offering a combination of MRI, rs-fMRI and demographic data.

The ABIDE I dataset has significant heterogeneity and variation that should be taken into account. It includes data from different research centers around the world, resulting in variations in scanning protocols, age groups, and other relevant factors. Consequently, the analysis and interpretation of the ABIDE I dataset is challenging due to this inherent heterogeneity.

## 3.2 Preprocessing Pipeline

We use the ABIDE I dataset provided by the Preprocessed Connectomes Project (PCP) [109] for our analysis. The PCP provides data for ABIDE I using various preprocessing strategies. In this work, we use the preprocessed data from the DPARSF pipeline [110]. The DPARSF pipeline is based on SPM8 and includes the following steps: The first 4 volumes of each fMRI time series are discarded to allow for magnetization stabilization. A slice timing correction is performed to correct for differences in acquisition time between slices. The fMRI time series are then

| Site | ASD | | TC | | Total |
|---|---|---|---|---|---|
| | Mean Age±std | Count | Mean Age±std | Count | |
| CALTECH | 27.44±10.30 | M:15, F:4 | 28.02±10.89 | M:14, F:4 | 37 |
| CMU | 30.33±8.50 | M:3 , F:0 | 25.50±6.36 | M:1 , F:1 | 5 |
| KKI | 9.56 ±1.40 | M:9, F:3 | 10.08±1.12 | M:20, F:7 | 39 |
| LEUVEN_1 | 21.86±4.11 | M:14, F:0 | 23.27±2.91 | M:15, F:0 | 29 |
| LEUVEN_2 | 13.81±1.06 | M:11, F:2 | 14.22±1.45 | M:14, F:5 | 32 |
| MAX_MUN | 30.44±13.99 | M:15, F:3 | 25.92±8.32 | M:23, F:1 | 42 |
| NYU | 14.92±7.09 | M:64, F:9 | 15.67±6.22 | M:72, F:26 | 171 |
| OHSU | 11.43±2.18 | M:12, F:0 | 10.37±1.10 | M:11, F:0 | 23 |
| OLIN | 16.79±3.77 | M:11, F:3 | 17.55±3.17 | M:9 , F:2 | 25 |
| PITT | 19.35±7.52 | M:18, F:4 | 19.13±6.32 | M:20, F:3 | 45 |
| SBL | 35.29±10.76 | M:14, F:0 | 34.42±6.04 | M:12, F:0 | 26 |
| SDSU | 15.05±1.67 | M:12, F:0 | 14.32±1.89 | M:15, F:6 | 33 |
| STANFORD | 10.15±1.65 | M:13, F:4 | 9.89±1.62 | M:15, F:4 | 36 |
| TRINITY | 17.01±3.12 | M:21, F:0 | 17.48±3.66 | M:23, F:0 | 44 |
| UCLA_1 | 13.62±2.69 | M:26, F:2 | 13.52±1.95 | M:23, F:4 | 55 |
| UCLA_2 | 12.35±2.06 | M:8, F:0 | 12.40±1.03 | M:10, F:2 | 20 |
| UM_1 | 13.44±2.41 | M:28, F:8 | 14.24±3.18 | M:31, F:15 | 82 |
| UM_2 | 15.05±1.49 | M:11, F:1 | 16.94±4.12 | M:18, F:1 | 31 |
| USM | 24.60±8.57 | M:38, F:0 | 22.33±7.87 | M:23, F:0 | 61 |
| YALE | 13.01±3.10 | M:15, F:7 | 12.76±2.84 | M:19, F:7 | 48 |

Table 3.1: Demographic information for the subset of the ABIDE I dataset that is used.

realigned to the first volume to correct for head motion. Intensity normalization is not applied. To remove confounding variation due to physiological noise, 24 parameters of head motion, mean white matter, and cerebrospinal fluid signals are regressed. Motion realignment parameters are also regressed, as are linear and quadratic trends in low-frequency drifts. Bandpass filtering was performed after regression of interfering signals to remove high frequency noise and low frequency drifts. Finally, functional to anatomical registration is performed using rigid body transformation and anatomical to standard space registration using DARTEL [111].

## 3.3 Brain Atlases

Since the dimensionality of the preprocessed data is very high, we perform dimensionality reduction by dividing the brain into a set of parcels or regions with similar properties according to a brain atlas. In this work, we process our data using three different atlases. The first atlas is the Automated Anatomical Labeling

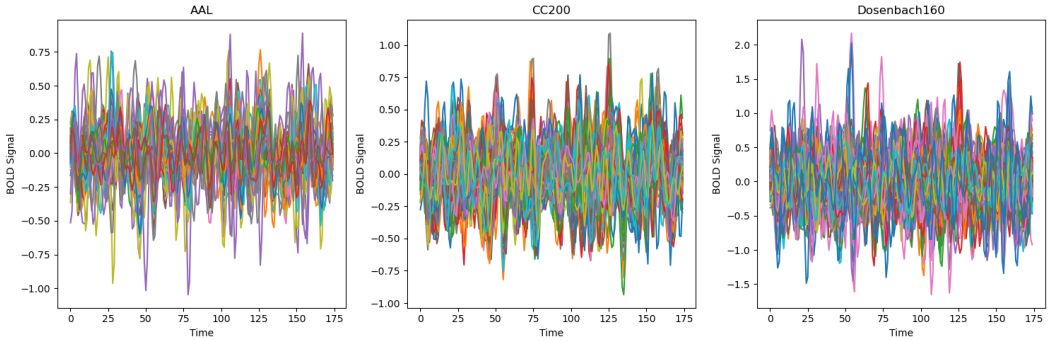



Figure 3.1: Preprocessed and parcellated fMRI time series of the three atlases under consideration for a randomly sampled subject.

(AAL) atlas [112]. This atlas, which is widely used in the literature, divides the brain into 116 ROIs based on anatomical landmarks and has been fractionated to a functional resolution of $3mm^3$ using nearest neighbor interpolation. The second atlas is the Craddock 200 (CC200) atlas [113]. It divides the brain into 200 ROIs based on functional connectivity and has been fractionated to a functional resolution of $3mm^3$ using nearest-neighbor interpolation. The third atlas we considered is the Dosenbach 160 (DOS160) atlas [114], which contains uniform spheres placed at coordinates obtained from meta-analyses of task-related fMRI studies.

### 3.3.1 Automated Anatomical Labeling (AAL)

The Automated Anatomical Labeling (AAL) atlas is a widely used brain parcellation scheme and anatomical labeling template [115]. It divides the brain into multiple ROIs, with each region corresponding to a specific anatomical area.

The development of the AAL atlas involved the following steps:
**Image Acquisition:** The AAL atlas is based on a single subject MNI T1 volume.
**Spatial Normalization:** The 27 scans of the single-subject MNI T1 volumes were spatially normalized using a nine-parameter linear transformation. The average of the 27 scans was then calculated.
**Segmentation:** The atlas provides segmentation into eight classes, including gray matter, white matter, cerebrospinal fluid, fat, muscle/skin, skin, skull, and glial matter.
**Parcellation:** After sulci delineation, 90 ROIs were manually drawn every 2 mm on axial sections. Sulci landmarks played a limiting role in ROI delineation. Each region was filled in 2D using a four-neighbor connectivity algorithm. For each anatomical region, a 3D Anatomical Volume of Interest (AVOI) was created that included all 3D pieces. Each AVOI was assigned a grayscale code.
**Automated Anatomical Labeling:** The AAL atlas provides three methods for automated anatomical labeling of functional studies: Extremum Labeling, Percent of

Voxels Within AVOIs, and Percent of Voxels Within AVOIs for Activated Clusters. Over time, the AAL atlas has been updated twice. The AALv2 [116] added ROIs in the orbitofrontal cortex, while the AALv3 [117] incorporated additional ROIs in the anterior cingulate cortex, thalamus, nucleus accumbens, substantia nigra, ventral tegmental area, red nucleus, locus coeruleus, and raphe nuclei. This development resulted in a total of 116 ROIs that are visualized in Figure 3.2.

For our analysis, after preprocessing and extracting the time series data, we parcellated the data into these 116 ROIs. This was done by calculating the average signal within each ROI. The result of this process is a 2D vector of the form $d_{AAL\,\mathrm{raw}} = (\#\,\text{time steps}, 116)$.

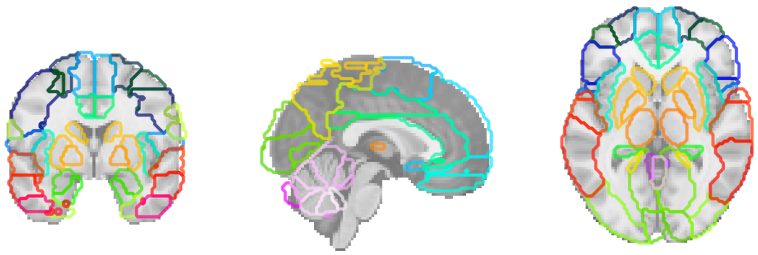


Figure 3.2: ROIs of the Automated Anatomical Labeling (AAL) atlas.

### 3.3.2 Craddock 200 (CC200) Atlas

The use of atlases based on anatomical or cyto-architectonic boundaries for defining ROIs is a common practice in various analyses. The Craddock 200 (CC200) atlas, proposed by Craddock et al. [113], employs a data-driven approach to delineate ROIs by parcellating whole-brain rs-fMRI data into spatially coherent regions with homogeneous functional connectivity.

The CC200 atlas is created in the following steps:
**Image Acquisition:** Data include rs-fMRI and anatomical T1 images acquired from 41 healthy controls.
**Preprocessing:** The preprocessing pipeline includes artifact removal and segmentation of anatomical scans into white matter, gray matter, and cerebrospinal fluid. Standard preprocessing procedures were applied to the rs-fMRI data.
**Spatially Constrained Functional Parcellation:** Spatially-constrained functional parcellation was applied to preprocessed and unfiltered resting-state data from 41

individuals ranging in age from 18 to 55 years (mean: 31.2, standard deviation: 7.8; 19 females). A gray matter mask was created by averaging individual-level gray matter masks obtained from automated segmentation.
**Individual-level connectivity graphs:** Each voxel within the gray matter mask was treated as a node in an individual-level connectivity graph. Edges between voxels were established based on super-threshold temporal correlations within a voxel's 3D neighborhood (27 voxels).
**Graph partitioning:** Individual-level connectivity graphs were partitioned into 200 regions using normalized cut spectral clustering [118, 119]. Association matrices were constructed from the clustering results, with connectivity between voxels set to 1 if they belonged to the same ROI, and 0 otherwise.
**Group-level correspondence matrix:** A group-level correspondence matrix was created by averaging the individual-level association matrices. This group-level matrix was further partitioned into 200 regions using normalized cut clustering.
**Functional resolution and labeling:** The resulting group-level analysis was converted to functional resolution using nearest neighbor interpolation. As the ROIs were not anatomically defined, labels for each resulting ROI were assigned based on their overlap with other anatomical atlases (AAL, EZ [120], HO [121], and TT [122]). The resulting ROIs canbe seen in Figure 3.3

For our specific analysis, we used the CC200 atlas to pre-process and extract time series data from our dataset. The data was then parcellated into the 200 ROIs defined by the CC200 atlas. The time series of each ROI was summarized by the average signal within that ROI, resulting in a 2D vector of the form $d_{\text{CC200 raw}} = (\#\text{ time steps}, 200)$.

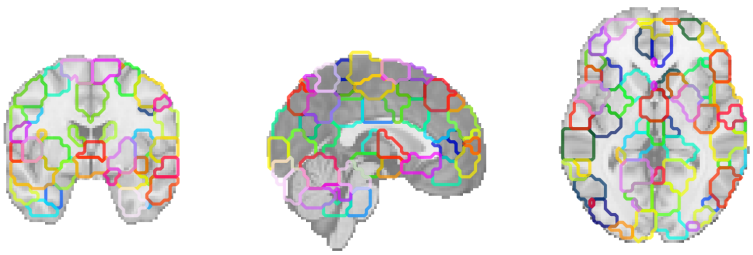


Figure 3.3: ROIs of the Craddock 200 (CC200) atlas.

### 3.3.3 Dosenbach 160 (DOS160) Atlas

The Dosenbach 160 (DOS160) atlas is based on functional connectivity patterns and is designed to capture the underlying functional organization of the brain. This approach to ROI definition emphasizes the functional architecture of the brain, leading to robust and accurate results in network analysis.

The construction of the DOS160 atlas follows these steps:
**Functional ROI Definition:** ROIs are defined functionally to correspond to the inherent functional organization of the brain. Meta-analyses of fMRI activation studies are used to create a comprehensive set of ROIs spanning the cerebral cortex and cerebellum. This approach allows known functions from previous research to be incorporated into the atlas.
**Meta-Analyses:** Five different meta-analyses were performed, each focusing on a specific function: Error processing, default mode (task-induced deactivations), memory, language, and sensorimotor functions. Activation data for these meta-analyses were collected using the same Siemens 1.5 Tesla Vision scanner.
**Thresholding and Analysis:** Thresholding was performed based on the statistical significance observed in the meta-analyses. The resulting thresholded activation images were combined to create a conjunction image, highlighting frequently activated voxels across the meta-analyses. Centroids of these activated voxel groups were determined using peak-finding algorithms within 10-mm diameter spheres.
**ROI Combination:** ROIs from different meta-analyses were combined, with priority given to ROIs from the default mode (task-induced deactivations) and error > correct (control) analyses. Overlapping ROIs across meta-analyses were spatially averaged.
**Total ROIs and Spatial Coordinates:** This construction resulted in a total of 160 non-overlapping ROIs, as can be seen in Figure 3.4 The arrangement of these ROIs ensures a minimum distance of 10 mm between their centers. All three-dimensional coordinates are in standardized MNI space, a common coordinate system for brain mapping.

In our specific analysis, we used the DOS160 atlas to preprocess and extract time series data from our dataset. The data was then parcellated into the 160 ROIs defined by the DOS160 atlas. The time series of each ROI was summarized by calculating the average signal within that ROI, resulting in a 2D vector of the form $d_{DOS160\,\mathrm{raw}} = (\#\,\text{time steps}, 160)$.

## 3.4 Functional Connectivity

As outlined by Stephan and Friston, the study of functional connectivity aims to understand brain function along two fundamental dimensions: Functional specialization and functional integration. The concept of functional specialization and integration seeks to address the following aspects:
**Functional Specialization:** This view postulates that certain aspects of information

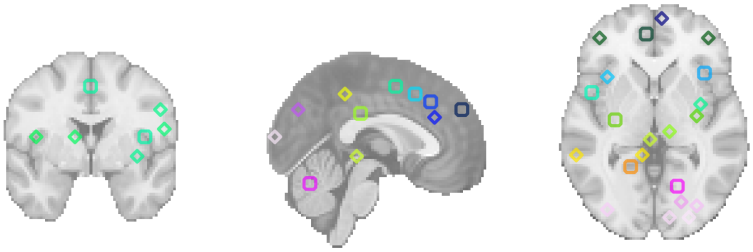


Figure 3.4: ROIs of the Dosenbach 160 (DOS160) atlas.

processing are locally specialized in certain cortical areas. However, it's plausible that specialization may also be distributed across different cortical regions.
**Functional Integration:** This dimension focuses on understanding how these specialized areas interact within a larger system.
Functional specialization assumes that different cortical areas exhibit local specialization. Areas activated by tasks are viewed as distributed elements of a larger system. Although functional specialization provides insights, it has limitations. It does not account for context-dependent interactions between specialized areas, and it does not explain the mechanisms by which these specialized areas are functionally integrated. Functional integration is quantified and explored using a variety of metrics and techniques:
**Functional Connectivity:** This is characterized by the statistical dependencies observed between distant neural data.

Functional connectivity is defined as the temporal correlation between the time series of different brain regions. The relationship between two voxels $x$ and $y$ is often represented by the Pearson correlation coefficient ($r$) of their respective time series.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3.1}$$

After obtaining the ROI time series from the three atlases, as visualized in Figure 3.1, we compute the functional connectivity using the Pearson Correlation Coefficient between each pair of ROIs. The resulting correlation matrix is visualized in Figure 3.5. The upper triangular part of the correlation matrix as well as the diagonal are then dropped and the lower triangular part is vectorized to obtain a feature vector of length $k(k-1)/2$, where $k$ is the number of ROIs, which then serves as input to our models. This results in a feature vector of length $d_{\text{CC200 FC}} = 19900$ for the CC200

atlas, $d_{DOS160\,\mathrm{FC}} = 12880$ for the DOS160 atlas, and $d_{AAL\,\mathrm{FC}} = 6670$ for the AAL atlas.

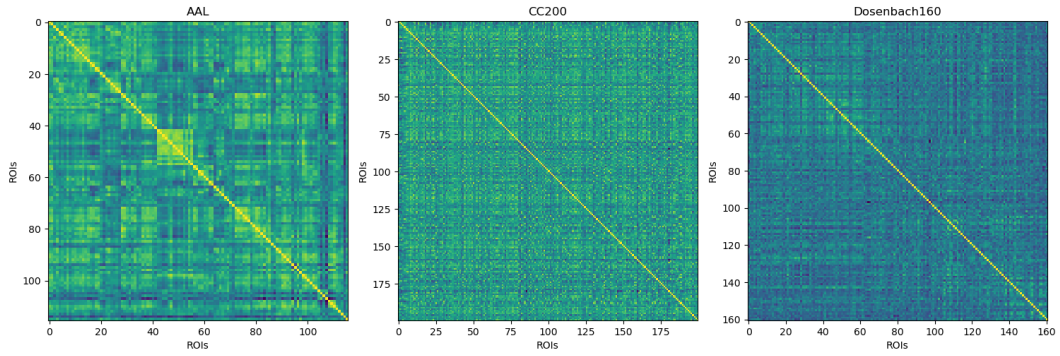



Figure 3.5: Functional connectivity matrices for the three atlases under consideration for a randomly sampled subject.

# 4 Neural Network Architectures

The neural network architectures used in this thesis are introduced in the following sections. First, we introduce a reference architecture, the Multi-Input Single-Output Deep Neural Network (MISODNN) [85] in section 4.1. We then present our proposed Multi-Atlas Transformer (METAFormer) architecture in section 4.2 and describe the self-supervised pretraining task in section 4.2.2.

## 4.1 Reference Architecture

In this section, we present a reference or baseline architecture. The selection of this architecture was based on thorough validation in the original work, i.e., using cross-validation, and on state-of-the-art performance using connectomes obtained from the rs-fMRI data provided by ABIDE I. As our reference architecture also uses connectomes from multiple atlases, it is well suited for comparison with the architecture we propose in section 4.2. Since our reference architecture also uses connectomes from multiple atlases, it is well suited for comparison with the architecture we propose in section 4.2.

### 4.1.1 Multi-Input Single-Output Deep Neural Network (MISODNN)

The Multi-Input Single-Output Deep Neural Network (MISODNN)[85] is designed for binary classification distinguishing Autism Spectrum Disorder (ASD) from typically developing controls (TC). This network architecture accommodates data from different brain atlases, allowing for a comprehensive analysis of brain connectivity.

**Multi-Input Architecture.** MISODNN uses a multi-input architecture to handle data from different brain atlases. In a forward pass, the connectomes obtained from the three atlases for the same subject are fed into the model to obtain a classification decision.

**Atlas Encoders.** MISODNN consists of three encoders, one for each atlas. The input dimension for each encoder corresponds to the number of features in that atlas, which are AAL (6670 features), CC200 (19900 features), and DOS160 (12800 features). Each encoder contains a hidden layer that encodes the input data into a

latent space with 700 neurons. The hidden layer is activated by a sigmoid activation function followed by a dropout with probability $p = 0.5$.

**Fusion.** Once the connectomes have been encoded separately into latent representations (1-D vectors of dimension $dim_{latent} = 700$), they are fused by concatenation. This results in a fused latent representation of dimension $dim_{fusion} = 3 * 700 = 2100$. This fused latent is then passed through a linear layer with 600 output neurons activated by the nonlinearity of the Rectified Linear Unit (ReLU). It is then fed into another linear layer with 400 output neurons, again activated by the ReLU. Finally, the data passes through an output layer of 2 neurons, where each neuron corresponds to one of the classes, ASD or TC. The outputs from this layer are raw class scores, and to obtain class probabilities, a sigmoid activation function is applied.

**Training.** During training, the network aims to minimize the binary cross-entropy loss between its predictions and the target labels. The model is trained for 300 epochs, or until an early termination is triggered, with a patience of 30 epochs.

**Hyperparameter search.** To identify appropriate hyperparameters for the MISODNN, we performed a hyperparameter search. A training and an evaluation set were randomly sampled with a 70/30 ratio. A grid search was then performed to find optimal values for hyperparameters such as learning rate and weight decay. The best model was selected based on achieving the highest accuracy on the evaluation set.

Figure 4.1 shows the architecture of MISODNN.

## 4.2 METAFormer

Here we propose METAFormer, at the core of which is the transformer encoder architecture originally proposed by Vaswani et al.[123] for natural language processing (NLP) tasks. The transformer architecture has since been successfully applied to a variety of tasks and has been shown to outperform convolutional neural networks (CNNs)[124]. Transformers are based on the attention mechanism, which allows the model to learn dependencies between input features. However, since our main goal is to perform classification rather than generation, we do not use the decoder part of the transformer architecture. To accommodate input from multiple different atlases, we use an ensemble of three separate transformers, with each transformer corresponding to a specific atlas. As shown in Figure 4.4, the input to each transformer is a set of flattened functional connectivity matrices.

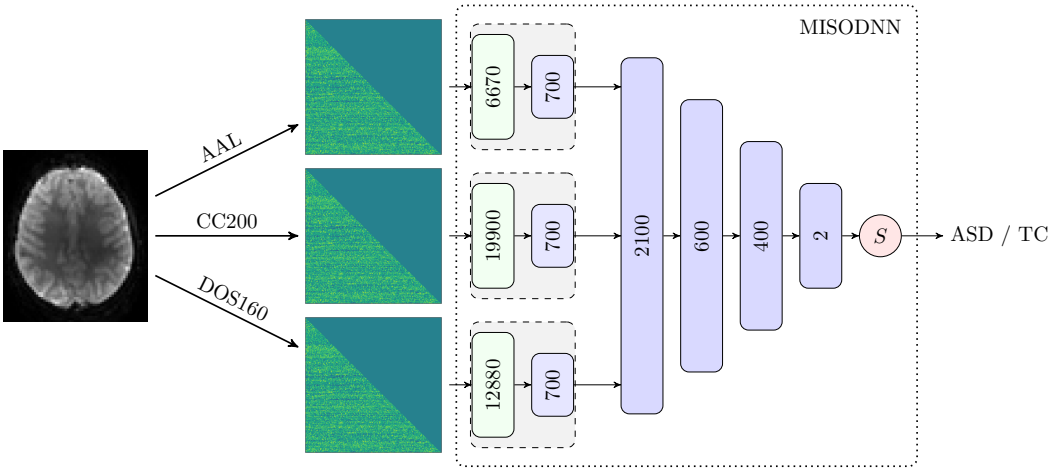



Figure 4.1: Schematic architecture of the MISODNN. Green blocks represent the input to the network and their number of input neurons. Blue blocks represent the hidden layers and their number of output neurons. The red circle represents the output nonlinearity.

### 4.2.1 Encoder Block

The encoder block is the core of our transformer architecture. Each encoder block consists of $N = 2$ identical encoder layers.

**Attention.** Attention, as described in [123], is a mechanism that maps a query and a set of key-value pairs to an output. In this context, the query, keys, and values are all vectors, and the output is a weighted sum of values. The weight assigned to each value is computed using a compatibility function of the query with the corresponding key.

The Scaled Dot-Product Attention mechanism involves the following steps: Queries, keys, and values are vectors of dimensions $d_k$, $d_k$, and $d_v$, respectively. Dot products are computed between queries and all keys, and the results are divided by $\sqrt{d_k}$. The softmax function is then applied to the divided dot products, resulting in weights on values. In practice, a set of queries is computed simultaneously, packed into a matrix $Q$. Keys and values are packed into matrices $K$ and $V$. The output of the scaled dot product attention is computed as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.1}$$

Assuming that $q$ and $k$ are $d_k$-dimensional vectors with components as independent random variables of mean 0 and variance 1, their dot product $q * k = \sum_{i=1}^{d_k} u_i v_i$ has a

mean of 0 and a variance of $d_k$. To ensure unit variance, the dot products are divided by $\sqrt{d_k}$.

**Multi-Head Attention.** Multi-Head Attention[123] is a mechanism that extends the basic attention mechanism by running multiple attention functions in parallel. Instead of using a single attention function with $d_{\text{model}}$-dimensional keys, values, and queries, projections are applied $h$ times using learned linear projections on dimensions $d_k$, $d_k$, and $d_v$. This approach allows the model to attend to different information from different representational subspaces at different locations, enhancing its ability to capture diverse relationships.

In multi-head attention, parallel attention functions are performed on these projected queries, keys, and values, resulting in $d_v$-dimensional output values. These output values from all heads are concatenated and then projected again to produce the final values. The mathematical representation of multi-head attention can be defined as

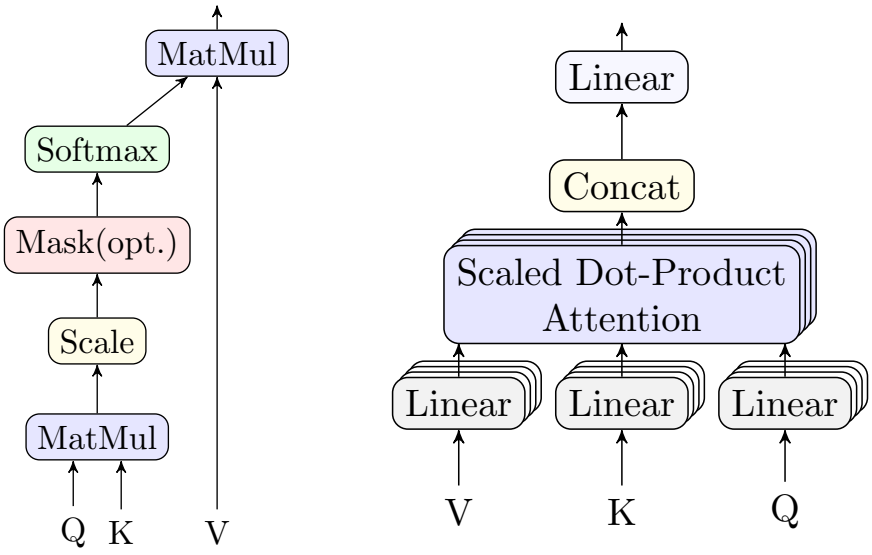
$$\text{MultiHead}(Q,K,V) = \text{Concat}(head_1,...,head_h) \cdot W_O, \tag{4.2}$$

where $head_i$ is computed using the Attention function with projections $QW_{Qi}$, $KW_{Ki}$, and $VW_{Vi}$. The projections are defined by the parameter matrices $W_{Qi}$, $W_{Ki}$, $W_{Vi}$, and the final projection is done by $W_O$.

In the original Transformer model, [123] uses $h = 8$ parallel attention layers (heads) and sets $d_k = d_v = \frac{d_{\text{model}}}{h} = 64$ for each head. Despite the reduced dimensions per head, the overall computational cost is similar to single-head attention with full dimensionality. Multi-head attention improves the model's ability to capture complex relationships by allowing it to focus on different aspects of the input data simultaneously, which is particularly useful for various natural language processing tasks.

**Encoder Blocks.** In our METAFormer architecture, the input to each transformer undergoes an embedding in a latent space using a linear layer with a dimensionality of $d_{model} = 256$. The output of the embedding is then divided by $\sqrt{d_{model}}$ to scale the input features. This scaling operation helps to balance the impact of the input features with the attention mechanism. Since we are not dealing with sequential data, positional encodings are not used.

The embedded input is then passed through two multi-head attention layers with $d_{ff} = 128$ feedforward units and $h = 4$ attention heads. To maintain stability during training, each encoder layer is normalized using layer normalization [125], and GELU [126] is used as the activation function. After the last encoder layer, the output passes through a dropout layer. Then a linear layer with $d_{model}$ hidden units and two output units corresponding to the two classes is applied to obtain the final output. The composition of this single atlas transformer is shown schematically in Figure 4.4.




(a) Scaled Dot-Product Attention

(b) Multi-Head Attention

Figure 4.2: Attention mechanisms used in the encoder block. Figures adapted from [123].

Since we want to use connectomes from three different atlases as input, we use three atlas transformers in parallel. The size of the input layer is adjusted according to the number of features of the given atlas. The final output of the three single-atlas transformers is averaged and passed through a softmax layer to derive the final class probabilities. The architecture of the multi-atlas transformer is shown in Figure 4.4.

**Training Configuration.** To train our Multi-Atlas Transformer model, we follow a series of steps. First, we initialize the model weights using the initialization strategy proposed by He [127], while setting the biases to 0. To optimize the model, we use the AdamW optimizer [128] and minimize the binary cross entropy between predictions and labels. Our training process consists of 750 epochs, using a batch size of 256. To avoid overfitting, we implement early stopping with a patience of 40 epochs. To ensure the robustness of our model, we apply data augmentation. Specifically, we randomly introduce noise into each flattened connectome vector with an augmentation probability of 0.3. The standard deviation of the noise is set to 0.01. We perform hyperparameter tuning using grid search. We optimize hyperparameters related to the optimizer, such as learning rate and weight decay. We also consider the dropout rate.

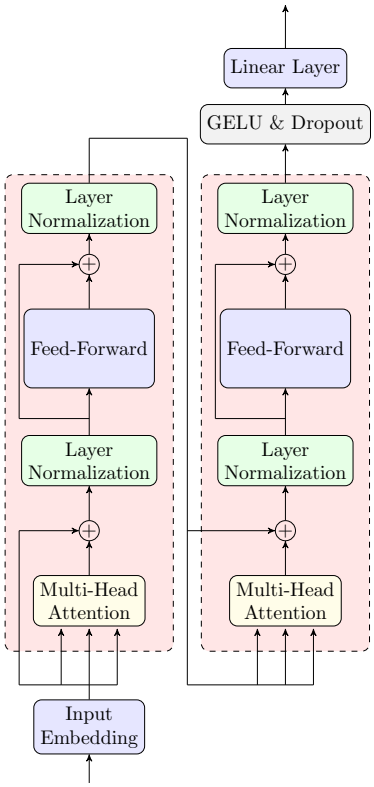



Figure 4.3: Schematic representation of a single atlas transformer. The input is a normalized connectome vector. The output after the final linear layer, or head, is a vector of size 2 representing the class logits.

### 4.2.2 Self-Supervised Pretraining

As popularized by [129], the use of self-supervised generative pretraining followed by task-specific fine-tuning has demonstrated improved performance in transformer architectures. Building on this approach, we propose a self-supervised pretraining task for our model. Our approach involves the imputation of missing elements in the functional connectivity matrices, inspired by the work introduced by [130]. To simulate missing data, we randomly set 10% of the features in each connectome to 0 and train the model to predict the missing values. The corresponding configuration is shown in Figure 4.5. To achieve this, we first randomly sample a binary noise mask $M \in \{0,1\}^{n_i}$ for each training sample, where $n_i$ is the number of features in the $i$-th connectome. Then, the original input $x$ is masked by element-wise multiplication with the noise mask: $x_{masked} = x \odot M$. We then pass the original input and the noise mask to the transformer, and explicitly care only about the unmasked elements.

To estimate the corrupted input, we introduce a linear layer with $n_i$ output neurons on top of the encoder stack, which predicts $\hat{x}_i$. We compute a multi atlas masked

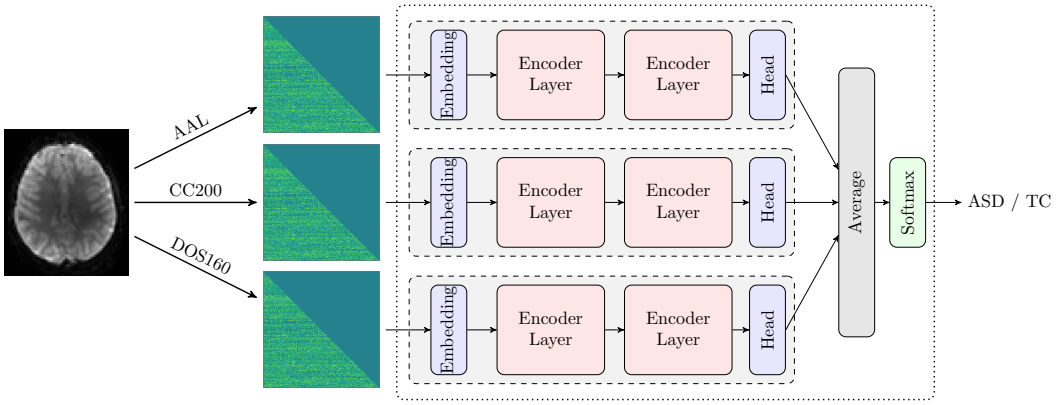



Figure 4.4: Architecture of METAFormer. Our model consists of three separate transformers, each corresponding to a specific atlas. The input to each transformer is a batch of flattened functional connectivity matrices with the diagonal and the upper triangular part of the matrix removed. The output of the transformers is averaged and passed through a softmax layer to derive the final class probabilities.

mean squared error (MAMSE) loss $\mathcal{L}_{multi}$ between the predicted and the original input:

$$\mathcal{L}_{multi} = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{n_i}\sum_{j\in M}^{n_i}\|x_{i,j} - \hat{x}_{i,j}\|^2 \tag{4.3}$$

where $x_{i,j}$ is the original value of the $j$-th masked input from the $i$-th atlas and $\hat{x}_{i,j}$ is the predicted value for the masked input at position $j$ in the $i$-th atlas.

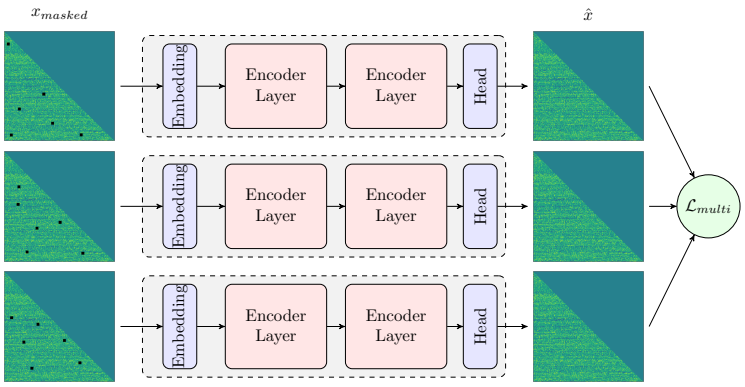



Figure 4.5: Self-supervised pretraining on the imputation task of the METAFormer architecture. The inputs to the model are masked connectomes with 10% of the features randomly set to 0 (shown as black squares). The model is trained to predict the missing values, which means that the output of the model has the same shape as the input.

# 5 ASD Classification

In this chapter we present the setup, evaluation and results of our experiments on ASD classification. Section 5.1 describes the experimental setup, including training and model configurations. Section 5.1.1 describes the evaluation metrics used to assess the performance of our models. Finally, section 5.2 provides a thorough comparison of the results of our ASD classification experiments.

## 5.1 Experimental Setup

To evaluate the classification performance of our models in a robust manner, we used 10-fold stratified cross-validation. For each fold, the model is trained on the remaining nine training folds and evaluated on the withheld test fold. Furthermore, we set aside 30% of each training fold as validation sets, which are then used for hyperparameter tuning and early stopping.

To evaluate the impact of self-supervised pretraining, we compare the performance of our model with and without pretraining. To achieve this, we first pretrain the model using the imputation task on the training folds, and then fine-tune the model using the classification task on the training folds, after which we evaluate it on the withheld test fold.

To verify the effectiveness of using multiple atlases as input, we compared the performance of our METAFormer model with that of single atlas transformer (SAT) models. To do this, we trained three separate transformer models using only one atlas as input. The SAT models are trained using the same architecture and training procedure as the METAFormer model. We also evaluated the performance of the SAT models with and without self-supervised pretraining to assess its impact on the performance of the model. To make the results comparable, we use the same training and validation folds for all model configurations under investigation.

### 5.1.1 Evaluation Metrics

By using cross-validation, we obtained 10 different sets of performance scores per configuration. These scores were then averaged and the standard deviation of each score was obtained, providing reliable estimates of the model's performance on

unseen data. Classification results are reported in terms of accuracy, defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5.1}$$

where TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives, respectively. In addition, we also report precision, defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.2}$$

recall (sensitivity), defined as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.3}$$

F1-score, defined as

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5.4}$$

and AUC-score, defined as the area under the receiver operating characteristic curve:

$$\text{AUC-score} = \int_0^1 \text{TPR}(t) \cdot \text{FPR}(t) \mathrm{d}t \tag{5.5}$$

where TPR and FPR denote the true positive rate and false positive rate, respectively.

## 5.2 ASD Classification Results

Table 5.1 shows the superior performance of our pretrained METAFormer model compared to previously published ASD classifiers trained on atlas-based connectomes. Importantly, our model achieves higher accuracy even when compared to approaches with similar test set sizes that do not employ cross-validation.

To further validate the effectiveness of our proposed multi-atlas transformer model for autism spectrum disorder classification, we compare METAFormer to single atlas transformers. The results, as shown in Table 5.2, demonstrate the superiority of METAFormer over all single atlas models in terms of accuracy, precision, recall, F1-score and AUC-score. It can be observed that the use of multiple atlases as input to the Transformer model improves the performance of the model. All three SAT variants achieve accuracies below 60%, while our multi-atlas model achieves an average accuracy of 62.8% across all 10 test folds. We can also see that using multiple atlases balances precision and recall, as the SAT models achieve lower precision but higher recall compared to the multi-atlas model. Furthermore, the multi-atlas model shows comparable or lower standard deviations in performance metrics compared to the SAT models. This indicates a higher robustness and stability of our multi-atlas METAFormer architecture due to the joint training of the three transformer encoders.

| Study | #ASD | #TC | Model | CV | Acc. |
|---|---|---|---|---|---|
| MAGE[87] | 419 | 513 | Graph CNN | 10-fold | 75.9% |
| AIMAFE[84] | 419 | 513 | MLP | 10-fold | 74.5% |
| 1DCNN-GRU[88] | – | – | 1D CNN | – | 78.0% |
| MISODNN[85] | 506 | 532 | MLP | 10-fold | 78.1% |
| 3D CNN[131] | 539 | 573 | 3D CNN | 5-fold | 74.53% |
| CNNGCN[89] | 403 | 468 | CNN/GRU | – | 72.48% |
| SSRN[90] | 505 | 530 | LSTM | – | 81.1% |
| **METAFormer** | 408 | 476 | Transformer | 10-fold | **83.7%** |

Table 5.1: Overview of state-of-the-art ASD classification methods that use large, heterogenous samples from ABIDE I. Note that our model achieves the highest accuracy while still using 10-fold cross-validation.

### 5.2.1 Impact of Pretraining

We also evaluated the effect of self-supervised pretraining on the classification performance of our models. As Table 5.2 shows pretraining significantly improves the performance of all models in terms of accuracy, precision, recall, F1-score and AUC-score. SAT models gain up to 11% in terms of accuracy. Performance gains are even more stark in the case of METAFormer, where pretraining improves accuracy by 21%. For all metrics considered, except for recall, the pretrained METAFormer model achieves the highest scores on average while also having lower standard deviations compared to the other models.

| Variant | Acc. | Prec. | Rec. | F1 | AUC |
|---|---|---|---|---|---|
| METAFormer PT | **0.837** ±0.030 | **0.819**±0.045 | 0.901 ±0.044 | **0.856** ±0.023 | **0.832** ±0.032 |
| METAFormer | 0.628 ±0.041 | 0.648±0.040 | 0.688 ±0.091 | 0.663 ±0.047 | 0.623 ±0.041 |
| SAT (AAL) | 0.593 ±0.040 | 0.585±0.042 | 0.888 ±0.091 | 0.701 ±0.024 | 0.568 ±0.047 |
| SAT (CC200) | 0.586 ±0.037 | 0.577±0.027 | 0.888 ±0.057 | 0.698 ±0.019 | 0.560 ±0.044 |
| SAT (DOS160) | 0.570 ±0.055 | 0.571±0.038 | 0.816 ±0.101 | 0.670 ±0.051 | 0.550 ±0.056 |
| SAT (AAL) PT | 0.601 ±0.069 | 0.587±0.055 | 0.939 ±0.059 | 0.719 ±0.033 | 0.573 ±0.077 |
| SAT (CC200) PT | 0.632 ±0.071 | 0.622±0.074 | 0.891 ±0.102 | 0.724 ±0.035 | 0.611 ±0.082 |
| SAT (DOS160) PT | 0.683 ±0.094 | 0.652±0.091 | **0.964** ±0.057 | 0.771 ±0.047 | 0.660 ±0.106 |

Table 5.2: Classification results for the different model configurations. Reported values are the mean ± standard deviation over 10 folds. Best results are in bold. SAT=Single Atlas Transformer, PT=Pretrained, atlases are in parentheses. Note that pretraining significantly improves performance across metrics and atlases. Using our multi-atlas METAFormer in combination with pretraining yields impressive performance gains.

# 6 Interpretability Experiments

In this chapter, we want to investigate which features are most important for the classification of ASD and TC subjects. To do this, we explain the experimental setup in section 6.1. We then want to know which feature attribution method provides the best explanations as measured by quantitative metrics. This quantitative evaluation is done in section 6.2. We also want to evaluate our results qualitatively and put them into a broader context. This is done in section 6.3.

## 6.1 Experimental Setup

In our experiments, we used the proposed METAFormer architecture. The dataset was randomly partitioned into different sets for training, validation, and testing purposes. To achieve this, we followed the pretraining and finetuning procedures described in section 4.2.2. We then proceeded to compute feature attributions. This process relied on the methods introduced in Chapter 2, and the computations were performed using the test set, which includes data from 89 subjects ($\#ASD = 41$, $\#TC = 48$). The specific methods used for the feature attribution calculations include Feature Ablation, Integrated Gradients (approximating the integral using the Gauss-Legendre method with $n = 50$ steps), DeepLIFT (with $\epsilon = 1e-10$), Saliency, GradientSHAP, and DeepLIFT SHAP (with $\epsilon = 1e-10$).

### 6.1.1 Baseline Selection

The selection of an appropriate baseline or reference value plays a key role in attribution methods such as Feature Ablation, Integrated Gradients, DeepLIFT, GradientSHAP, and DeepLIFT SHAP. However, this decision is not straightforward due to the unique characteristics of our domain. Unlike the RGB/grayscale images for which these methods were originally intended, we work with correlation matrices. The standardized nature of our connectome data scales the data to the range $[-1, 1]$. In this context, a value of -1 denotes the strongest negative correlation within a given subject, as opposed to absolute anticorrelation. Similarly, a value of 1 indicates the strongest positive correlation for a given subject, while 0 indicates the average correlation rather than a complete lack of correlation. Given these subtleties, a singular choice emerges as appropriate for the baseline/reference value: the mean correlation calculated across the time series.

## 6.2 Quantitative Evaluation

We quantitatively evaluate the quality of each attribution using the max-sensitivity and infidelity metrics introduced in section 2.7. Our focus is on examining the profiles of max-sensitivity and infidelity across the entire test dataset. The goal is not to evaluate individual instances, but rather to identify an attribution technique that consistently produces satisfactory attributions across the dataset. Thus, we seek to capture an aggregate measure of sensitivity and infidelity on average.

The maximum sensitivity of an attribution is computed using a Monte Carlo sampling-based approximation. Specifically, we sample 10 points from a subspace inside an L-infinity sphere with a radius of 0.02. The resulting maximum sensitivities for the entire test set are then averaged for each attribution method.

Similarly, the infidelity is determined by a similar process and averaged over the test set. We use a Gaussian distribution as the perturbation function with a mean of zero and a standard deviation of 0.003.

**Results** From Table 6.1, we can see that our proposed METAFormer model, IntegratedGradients, DeepLIFT, and SHAP have comparable mean maximum sensitivities. Notably, DeepLIFT has a slightly lower mean max-sensitivity and a slightly lower standard deviation, with values of $0.016 \pm 0.003$ for METAFormer and $0.014 \pm 0.002$ for MISODNN. The results for our reference model are strikingly similar, although DeepLIFT SHAP has a slightly lower standard deviation of $0.0014 \pm 0.001$ compared to DeepLIFT alone ($0.014 \pm 0.002$).

Saliency ($0.060 \pm 0.0013$, $0.0046 \pm 0.015$) and Feature Ablation ($0.051 \pm 0.009$, $0.039 \pm 0.011$) show significantly higher mean max-sensitivity for both METAFORMER and MISODNN, respectively. Conversely, GradientSHAP ($0.356 \pm 0.119$, $0.322 \pm 0.102$) differs significantly from the other methods, with mean max-sensitivities an order of magnitude higher for both METAFormer and MISODNN. A visual comparison of the max-sensitivity values is shown in Figure 6.1.

| Methods | METAFormer | MISODNN |
|---|---|---|
| Saliency | 0.060±0.013 | 0.046±0.015 |
| IntegratedGradients | 0.017±0.003 | 0.015±0.002 |
| DeepLIFT | **0.016±0.003** | 0.014±0.002 |
| Feature Ablation | 0.051±0.009 | 0.039±0.011 |
| GradientSHAP | 0.356±0.119 | 0.322±0.102 |
| DeepLIFT-SHAP | 0.017±0.005 | **0.014±0.001** |

Table 6.1: Mean Max-Sensitivity comparison of the different attribution methods for METAFormer and MISODNN. The best results are highlighted in bold. The average±standard deviation is reported for the whole test set.

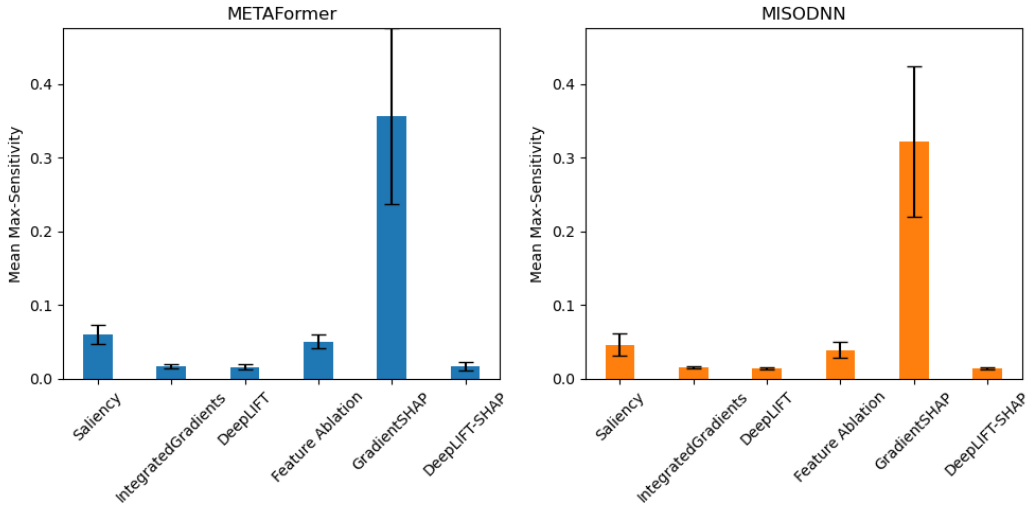



Figure 6.1: Max-Sensitivity comparison of the different attribution methods for both METAFormer and MISODNN. Lower values are better.

Table 6.2 shows the results of the infidelity evaluation. For our METAFormer architecture and the reference MISODNN architecture, Saliency records the lowest mean infidelity ($6.685e-06 \pm 2.867e-06$, $3.081e-06 \pm 1.502e-06$).

Integrated Gradient ($9.497e-06 \pm 4.166e-06$, $4.301e-06 \pm 1.934e-06$) and DeepLIFT ($9.909e-06 \pm 9.909e-06$, $4.425e-06 \pm 4.425e-06$) show closely matched results.

In particular, Feature Ablation provides more faithful attributions for both model architectures ($7.586e-6 \pm 3.243e-6$, $3.119e-6 \pm 1.342e-6$).

GradientSHAP ($10.81e-6 \pm 4.710e-6$, $4.841e-6 \pm 1.893e-6$) and DeepLIFT SHAP ($10.40e-6 \pm 5.297e-6$, $4.628e-6 \pm 2.481e-6$) yield similar but lower infidelity assignments for both METAFormer and MISODNN. A visual representation of these results is provided in Figure 6.2.

Overall, DeepLIFT emerges as an attractive choice, offering a balanced trade-off between mean max-sensitivity and infidelity.

### 6.2.1 Impact of Baseline Choice

Given the critical role of baseline or reference selection for Integrated Gradients, DeepLIFT, and SHAP implementations, we further investigate the influence of different baselines on maximum sensitivity and infidelity of attributions. The evaluation includes baselines in the range $[-1, 1]$ with a step size of 0.1.

Figures 6.3a and 6.3c illustrate that increasing baseline values lead to an increase in mean max-sensitivity for IntegratedGradients, DeepLIFT, and DeepLIFT-SHAP,

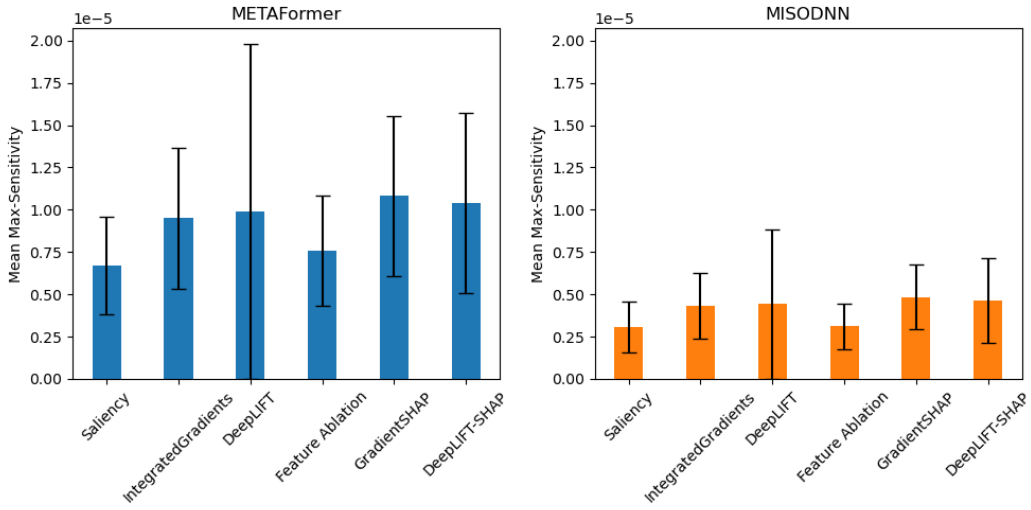


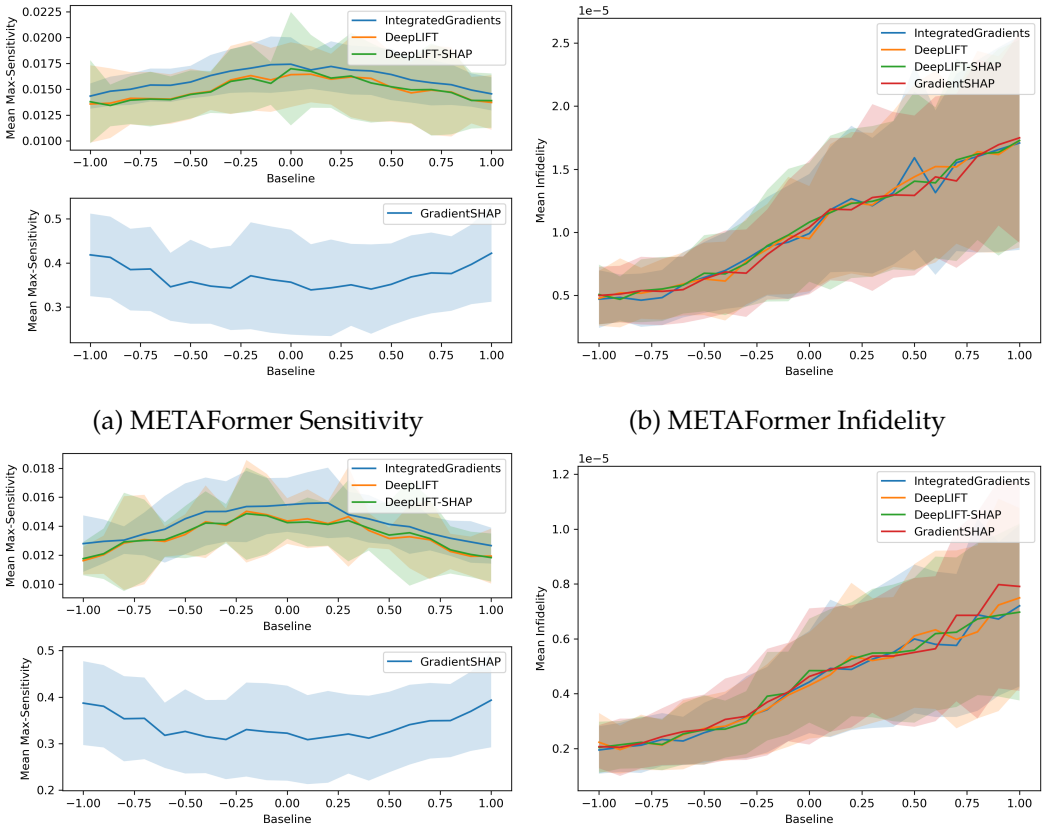

Figure 6.2: Infidelity comparison of the different attribution methods for both METAFormer and MISODNN. Lower values are better.

| **Methods** | **METAFormer** | **MISODNN** |
|---|---|---|
| Saliency | **6.685e-6±2.867e-6** | **3.081e-6±1.502e-6** |
| IntegratedGradients | 9.497e-6±4.166e-6 | 4.301e-6±1.934e-6 |
| DeepLIFT | 9.909e-6±9.909e-6 | 4.425e-6±4.425e-6 |
| Feature Ablation | 7.586e-6±3.243e-6 | 3.119e-6±1.342e-6 |
| GradientSHAP | 10.81e-6±4.710e-6 | 4.841e-6±1.893e-6 |
| DeepLIFT-SHAP | 10.40e-6±5.297e-6 | 4.628e-6±2.481e-6 |

Table 6.2: Infidelity comparison of the different attribution methods for METAFormer and MISODNN. The best results are highlighted in bold. The average±standard deviation is reported for the whole test set.

peaking at 0, after which it decreases, accompanied by an increased standard deviation. Interestingly, DeepLIFT-SHAP experiences a sudden spike in mean max-sensitivity at baseline=0, coinciding with a significant increase in standard deviation. IntegratedGradients consistently maintains a higher mean max-sensitivity compared to DeepLIFT and DeepLIFT-SHAP. GradientSHAP deviates from this trend, with mean max-sensitivity decreasing steadily from -1 to 0, reaching its lowest point near a baseline of 0, and then increasing for values above 0.

Figures 6.3b and 6.3d depict the continuous increase in infidelity with higher baseline values for both METAFormer and MISODNN, along with a corresponding growth in standard deviation. In light of these findings, the assumption of a baseline of 0 may need to be reconsidered. For all methods examined, using -1 as the baseline for the attribution calculation seems to provide more faithful and robust explanations.




(a) METAFormer Sensitivity

(b) METAFormer Infidelity

(c) MISODNN Sensitivity

(d) MISODNN Infidelity

Figure 6.3: Comparison of Sensitivity and Infidelity for METAFormer and MISODNN under different baseline values.

However, DeepLIFT and DeepLIFT-SHAP appear to be relatively stable in terms of max-sensitivity over the range of baselines.

## 6.3 Qualitative Evaluation

Our quantitative evaluation has established DeepLIFT as a robust and faithful method for generating attributions. In this section, we delve into qualitative analysis to gain further insight into the interpretability of the models.

We aim to answer two key questions:

1. What features contribute most, on average, to subjects classified as having ASD?

2. What internal representations most discriminate between ASD and TC?

To address the first question: We focus on the most robust and faithful feature attribution method that we determined based on our quantitative analysis (see section 6.2). Using the baseline value of -1 (as determined in section 6.2.1), we examine the attributions provided by DeepLIFT for both METAFormer and MISODNN. Our interest is to identify the features that have the greatest impact on classifying an individual as having ASD, and to examine whether these features are consistent with the existing literature on ASD. We identify the top 10 traits with the highest absolute attribution scores.

As AAL and DOS160 provide anatomical labels for each ROI, we present the corresponding ROI labels for the top 10 features in METAFormer and MISODNN in Tables 6.3 and 6.4. However, as CC200 lacks anatomical labels, we visualize the locations of the top 10 critical connections between regions in Figures 6.4 and 6.5.

Our observations show the prominence of cerebellar features in both architectures considered. This is consistent with the conclusions of several studies that emphasize cerebellar involvement in ASD [132, 133]. In addition, features associated with the default mode network, such as the cingulate cortex and precuneus, emerge as highly influential, consistent with their reported associations with ASD in the literature [134, 133, 135, 136]. We also find support for the association between weaker functional connectivity and ASD in connections between frontal and occipital regions, consistent with previous research [135, 137, 44]. In addition, brain regions such as the frontal and temporal cortex, amygdala, and hippocampus, which have previously been implicated in ASD [138, 139, 140, 141, 142], emerge as important in the attributions of both architectures. This alignment underscores the consistency between the attributions from the chosen method and the existing neuroscientific literature on neuroanatomical correlates of ASD.

### 6.3.1 Visualizing Learned Class Representations

We use the feature visualization technique described in section 2.6 to visualize the optimal internal representation of a given class. Using METAFormer and MISODNN models with frozen weights, we iteratively optimize a random input to maximize the excitation of the neuron corresponding to the class under consideration. The optimization process uses the Adam optimizer with a learning rate of $5 \times 10^{-5}$ and iterates for 50,000 steps. This process is repeated for each class in both models. Figures 6.6a and 6.6b illustrate the features with the highest absolute value in the difference between the class representations of ASD and TC.

| # ↓ | AAL Connections | DOS160 Connections |
|---|---|---|
| 1. | Fusiform L ↔ Caudate R | Post Occipital ↔ TPJ |
| 2. | Fusiform R ↔ Parietal Sup L | Post Cingulate ↔ Inf Cerebellum |
| 3. | Supp Motor Area L ↔ Precentral L | Precentral Gyrus ↔ Thalamus |
| 4. | Cingulum Mid R ↔ Temporal Mid L | aPFC ↔ Lat Cerebellum |
| 5. | Amygdala R ↔ Vermis 4 5 | Mid Insula ↔ Post Occipital |
| 6. | Vermis 1 2 ↔ Frontal Med Orb R | Temporal ↔ mPFC |
| 7. | Cingulum Mid L ↔ Temporal Mid R | IPL ↔ occipital |
| 8. | Frontal Sup Medial L ↔ Cerebelum 8 R | vFC ↔ Vent aPFC |
| 9. | Olfactory L ↔ Vermis 10 | Occipital ↔ Med Cerebellum |
| 10. | Amygdala R ↔ Vermis 8 | dFC ↔ Parietal |

Table 6.3: Top 10 connections for the AAL and DOS160 atlases as determined by extracting the features with highest absolute attribution scores using DeepLIFT for METAFormer. Connections are sorted descending by their importance. Names of the connections correspond to the labels from the respective atlas.

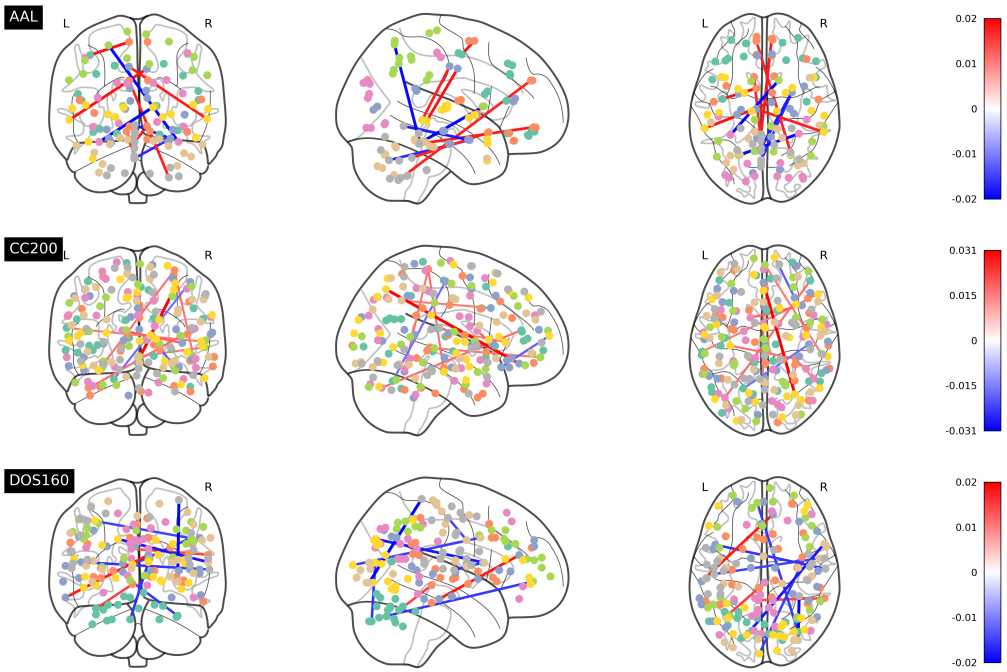



Figure 6.4: Visualization of the top 10 features with the highest absolute attribution scores for the ASD class in METAFormer. Positive attribution scores indicate that the feature's value is contributing to increasing the model's output from the baseline value and vice versa for negative scores.

| # ↓ | AAL Connections | DOS160 Connections |
|---|---|---|
| 1. | Cingulum Mid L ↔ Frontal Inf Oper L | ACC ↔ dACC |
| 2. | Frontal Sup R ↔ Frontal Inf Oper R | Ant Insula ↔ aPFC |
| 3. | Lingual R ↔ Cerebelum Crus2 R | Precuneus ↔ Pre-SMA |
| 4. | Frontal Inf Oper R ↔ Olfactory R | Post Cingulate ↔ vPFC |
| 5. | Temporal Pole Mid L ↔ Cingulum Mid R | Basal Ganglia ↔ Mid Insula |
| 6. | Cerebelum Crus1 R ↔ Cingulum Ant R | Pre-SMA ↔ Mid Insula |
| 7. | Frontal Inf Oper R ↔ Olfactory L | Lat Cerebellum ↔ Post Cingulate |
| 8. | Occipital Sup L ↔ Calcarine L | vPFC ↔ Post Cingulate |
| 9. | Angular R ↔ Amygdala R | Temporal ↔ Vent aPFC |
| 10. | Cerebelum 6 R ↔ Cuneus R | Thalamus ↔ Lat Cerebellum |

Table 6.4: Top 10 connections for the AAL and DOS160 atlases as determined by extracting the features with highest absolute attribution scores using DeepLIFT for MISODNN. Connections are sorted descending by their importance. Names of the connections correspond to the labels from the respective atlas.

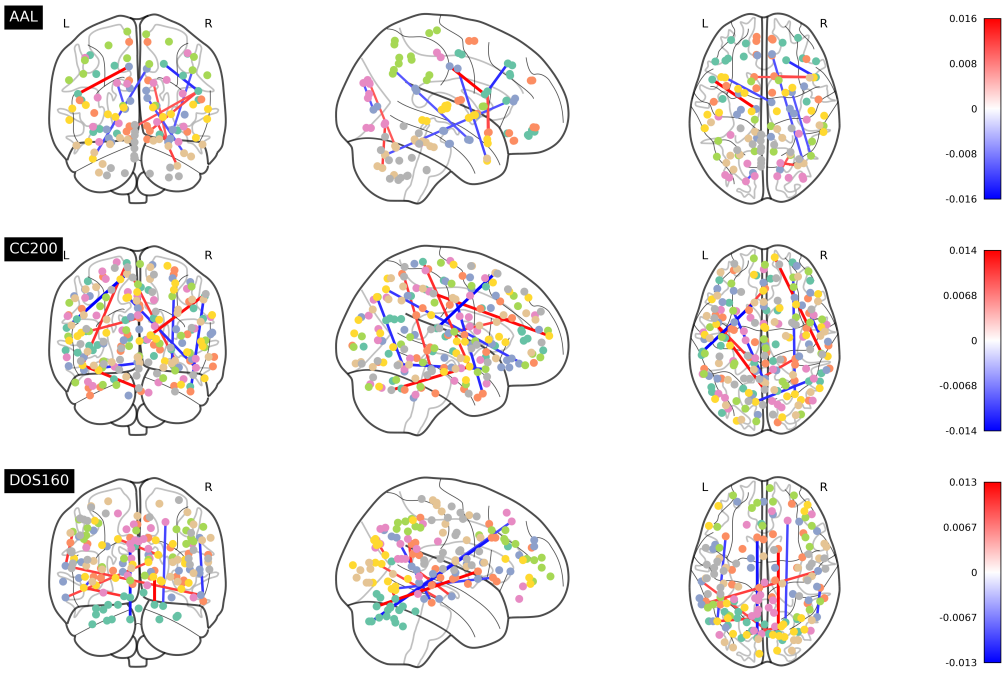
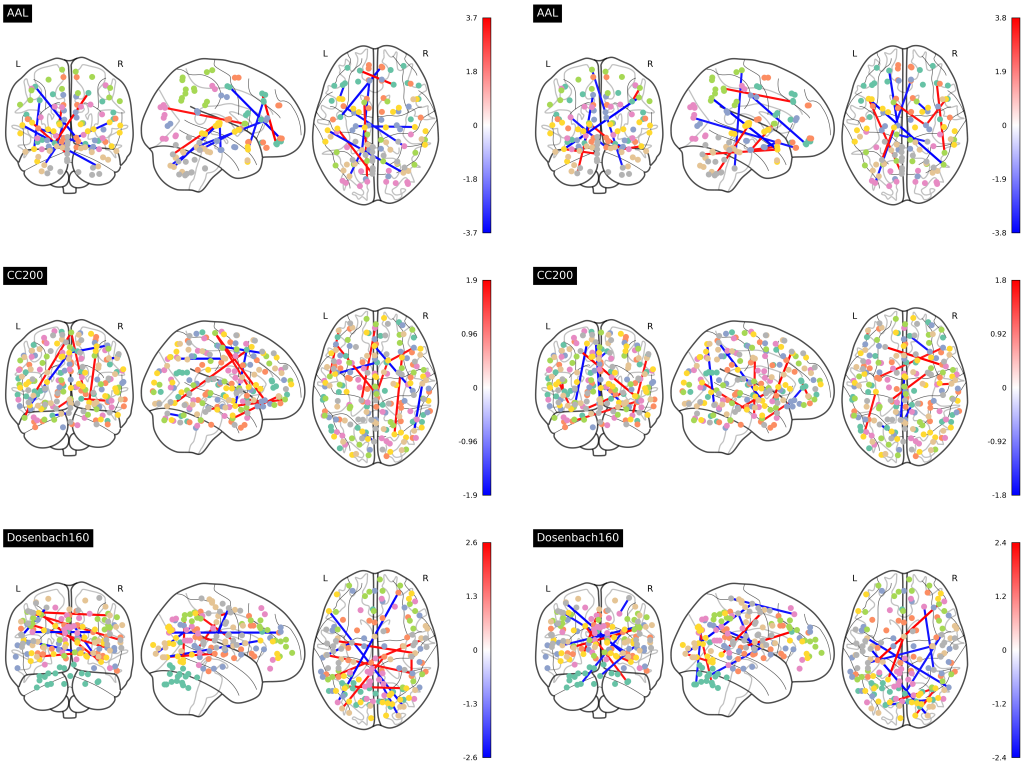



Figure 6.5: Visualization of the top 10 features with the highest absolute attribution scores for the ASD class in MISODNN. Positive attribution scores indicate that the feature's value is contributing to increasing the model's output from the baseline value and vice versa for negative scores.




(a) METAFormer (b) MISODNN

Figure 6.6: Visualization of the difference maps of the top10 features of the learned class representations for the ASD and TC classes. The difference maps are computed by subtracting the feature maps of the ASD class from the TC class

# 7 Discussion & Conclusion

In this work, we sought to determine the feasibility of training a deep learning model to accurately discriminate between ASD and TC subjects using rs-fMRI data, while also striving to extract meaningful and interpretable brain biomarkers from the model. We introduced a novel multi-atlas enhanced transformer architecture, METAFormer, which uses functional connectomes from three different atlases as input to formulate predictions. Our novel METAFormer architecture was pretrained on the same dataset using a self-supervised strategy aimed at imputing missing edges in the input connectomes.

Our cross-validation evaluation showed that our model outperformed previous methods, achieving a mean accuracy of 83.7% across 10 test folds and an AUC score of 0.832. By applying common feature attribution techniques to our METAFormer model and a reference architecture, we quantitatively assessed the quality of the explanations provided by these techniques. Based on measures of max-sensitivity and infidelity, we found that DeepLIFT provided the most robust and faithful attributions. In addition, our qualitative analysis of the features that DeepLIFT identified as critical to ASD classification revealed significant overlap with the neuroscientific literature, strengthening the validity of our findings.

**ASD Classification** To our knowledge, our work marks the first application of transformers to the domain of ASD classification. Our pretrained METAFormer architecture achieves significantly higher accuracies than previous state-of-the-art models that also use functional connectomes. The closest competitor, SSRN [90], uses an LSTM-based architecture and achieves 81.1% accuracy. However, their approach lacks cross-validation, and the exact train-validation-test splits remain undisclosed, preventing a fair comparison.

The reference architecture used in our study, as proposed by [85], comes closest to our reported accuracy, and their original work includes cross-validation results. Nevertheless, direct comparisons with other approaches remain challenging due to the lack of a universally accepted train-validation-test configuration and publicly available held-out test sets, similar to standardized machine learning competitions such as BRATS [143] or the Medical Segmentation Decathlon [144]. This challenge is further compounded by methods that use raw rs-fMRI time series, such as [80, 83, 82]

In addition, as data complexity reguires larger models, training and inference times increase significantly, making cross-validation with numerous folds computationally

demanding. Coupled with the lack of publicly available model code and weights, as well as omitted implementation details in some papers, the task of model comparability is significantly hampered.

We also investigated the effect of using multiple atlases, as multi-atlas or ensemble models are commonly used in ASD classification [85, 94, 84, 87]. Surprisingly, when using Single Atlas Transformer (SAT) models, which mirror the configuration of the multi-atlas METAFormer, their performance was inferior in the single atlas setting. Accuracy remained consistently below 60% for the single atlas variants, although recall showed a significant increase, indicating higher sensitivity but also an increased rate of false positives. This imbalance was corrected when predictions from multiple atlases were combined. Nevertheless, achieving a harmonious trade-off between precision and recall in ASD classification is less straightforward than in tasks such as tumor detection.

The paradigm of pretraining has led to significant advances in the broader field of deep learning [145, 146], with substantial performance gains particularly evident in transformer models [129, 147, 148]. Pretraining is also often used in ASD classification, as seen in [149, 150, 151, 80].

In our context, the self-supervised imputation task we propose for pretraining is conceptually simple and easy to implement. Our experiments show that even for single atlas variants of METAFormer, pretraining leads to remarkable accuracy improvements of up to 11%. Remarkably, when the multi-atlas METAFormer was pretrained and fine-tuned for the classification task, performance gains of over 20% were unexpectedly achieved. This remarkable increase could be attributed to the combined effects of pretraining and ensemble learning.

In addition, the reduced model complexity of single atlas variants likely contributes to the pronounced performance gains of the pretrained METAFormer. Larger transformer models with more trainable parameters generally benefit more from pretraining, a factor that likely underlies the dramatic performance improvement observed with the pretrained METAFormer variant.

**Interpretability** Interpretation of ASD classification results is critical. The use of black-box models in clinical scenarios is challenging because clinicians need to understand the reasoning behind a model in order to trust its results and integrate them effectively into decision-making processes.

For example, [80] used a variant of feature ablation to identify important features, while [94] used feature visualization and variants of saliency maps to identify importance trends. To our knowledge, we have introduced a novel application of Integrated Gradients and DeepLIFT, along with their SHAP counterparts, to provide a comprehensive quantitative assessment of the produced attributions. We demonstrated that the considered feature attribution methods yielded similar attribution

quality for both METAFormer and MISODNN, as measured by quantitative metrics such as max-sensitivity and infidelity.

Unexpectedly, GradientSHAP showed significantly higher max-sensitivity across all configurations examined, while maintaining comparable infidelity to the other attribution methods. Our quantitative evaluation revealed that DeepLIFT provides the most robust and faithful attributions in terms of max-sensitivity and infidelity.

The choice of baseline value has a significant impact on the attribution scores generated by methods such as Integrated Gradients, DeepLIFT, and their SHAP variants. Contrary to our initial intuition of using 0 as a baseline, which represents the mean functional connectivity for a subject, since the input matrices are standardized correlation matrices, this choice did not yield optimal results for most methods.

Integrated Gradients, DeepLIFT, and DeepLIFT-SHAP had the lowest maximum sensitivity and infidelity when using a baseline value of -1. In our context, -1 represents the highest value of anticorrelation within a given subject. These results highlight the need to adapt the choice of baseline to the specific context of the data. However, the adaptation process is not always as straightforward as in other domains, such as image classification, where a background value of 0 could be used.

Quantitative evaluation of the explanations provided is crucial for building trust and transparency. However, qualitative evaluation of the results obtained is equally important for determining neuroscientific validity. For both our METAFormer model and the reference model, we identified the most important connections by selecting features with the highest absolute attribution scores for the ASD class.

As expected, there was considerable overlap with existing autism research. Regions associated with the default mode network emerged as top features in both models, including the cingulate cortex and precuneus, which were consistently prominent across atlases. In both models, default mode network activity appears to be critical for classifying autistic individuals. The default mode network is often associated with specific components of social cognitive disorders, such as self-referential processing and theory of mind [152]. Interestingly, cerebellar connections also consistently appeared to be of high importance, consistent with existing work showing structural and functional differences in the cerebellum in autistic subjects [132]. Similarly, frontal and occipital regions were consistently important features. Looking at the learned class representations that most strongly distinguish the two classes, the above findings are also confirmed.

In general, the attributions provided by DeepLIFT are very consistent with the existing autism literature. The demonstrated influence of the default mode network, cerebellum, frontal and occipital regions in discriminating between ASD and controls could be a significant step towards the development of model-based imaging biomarkers.

Interpretibility is critical when AI-driven decisions impact the lives of patients.

Ensuring model trust, robustness, and transparency is essential prior to integration into the clinical decision-making process. Thus, interpretable and explainable AI is an important component in biomarker discovery, disease understanding, and supporting effective clinical implementation of AI models.

## 7.1 Outlook

Our work suggests several avenues for future research. While parcellation serves as a drastic dimensionality reduction technique, it inevitably results in information loss. In terms of functional connectivity, the technique comes with several caveats: Temporal information is virtually lost, correlation does not necessarily imply causation, and the lack of directionality imposes limitations. In addition, selection bias in the choice of ROIs can significantly affect the results. Variability between sessions, subjects, and sites further complicates the scenario. To mitigate these limitations, alternative measures such as effective connectivity [153], dynamic effective connectivity [154], spectral coherence [155], wavelet coherence [156] or mutual information [157] have been proposed.

In addition, an updated version of the ABIDE dataset, ABIDE II [158], with an additional 1,114 subjects, could potentially lead to improved generalization and provide a more heterogeneous view of the extracted biomarkers. Our ASD classification approach could further benefit from more data, allowing the training of larger transformer models for potentially better generalization capabilities. Exploring alternatives to Pearson correlation for functional connectivity, combining different connectivity metrics, and incorporating raw fMRI or ROI time series data into the model training process could address the limitations of our correlation-based approach. The use of transfer learning, as demonstrated in e.g. [80], could also be explored. In addition, consideration of multimodal data integration could prove fruitful. Expanding the ensemble with more transformers or exploring ways beyond simple averaging to combine representations from individual atlas transformers is worth investigating.

Interpretability of connectome data presents inherent challenges due to the complexity of large correlation matrices and the sheer number of ROIs, necessitating reliance on visualization techniques such as glass brains or correlation plots. Although we have provided a comprehensive analysis of commonly used interpretability methods, a complete investigation of all available methods is beyond the scope of this work. The field of interpretable AI is growing rapidly, with promising applications in the clinical domain. While post-hoc feature attribution methods provide accessible explanations of significant input features, they often lack information on how these regions contribute to the model output. In addition, important input features may overlap across classes, complicating subsequent interpretation.

Emerging advances in language descriptions for deep learning, as seen in [159, 160],

could bridge the gap between the visual and textual explanations provided by attribution methods. Counterfactual explanations, where input images are perturbed to elicit an opposite model prediction [161], could also reveal insights into the underlying data structures.

Careful evaluation of attribution methods is paramount for accurate, reliable, and clinically useful explanations. Integrating human-centered, subjective evaluations with quantitative, objective assessments is necessary for comprehensive evaluation of explanations. Comparison of explanations requires consideration of metrics that are collectively tailored to the task and the user. Limited research has explored the comparison of different explanations based on evaluation metrics [162].

## 7.2 Conclusion

In this thesis, we aimed at two things, first, to develop a robust deep learning model to discriminate between autism spectrum disorder (ASD) and typically developing subjects using rs-fMRI data, and second, to extract meaningful and interpretable biomarkers from the learned representation of the model. Our novel approach, the Multi-Atlas Enhanced Transformer Architecture (METAFormer), integrates functional connectomes from multiple atlases to improve prediction accuracy. Using self-supervised pretraining, we achieved remarkable results, with METAFormer consistently outperforming existing methods. Across ten test folds, the model achieved an average accuracy of 83.7% and an AUC score of 0.832, confirming its effectiveness in ASD classification.

With interpretability in mind, we applied feature attribution techniques to both the METAFormer model and a reference architecture. Our introduction of max-sensitivity and infidelity metrics allowed for quantitative evaluation of these explanations. Remarkably, DeepLIFT emerged as the most robust and faithful attribution method, facilitating better model understanding.

Qualitative analysis of features identified by DeepLIFT as critical to ASD classification revealed significant overlap with established ASD-associated neuroanatomical regions and connections. This alignment between model attributions and existing neuroscientific knowledge strengthens the credibility of our approach and underscores the potential of these highlighted features as potentially meaningful biomarkers.

In conclusion, our study presents METAFormer as a novel tool for accurate and interpretable ASD classification. We highlight its interpretability through advanced feature attribution methods. Beyond diagnostic implications, our research contributes to the understanding of the neurobiological facets of ASD.

# Bibliography


[1] J. Pevsner, "Leonardo da vinci's studies of the brain," *The Lancet*, vol. 393, pp. 1465–1472, Apr. 2019.

[2] H. Berger, "Über das elektroenkephalogramm des menschen," *Archiv für psychiatrie und nervenkrankheiten*, vol. 87, no. 1, pp. 527–570, 1929.

[3] S. K. Mishra and P. Singh, "History of neuroimaging: The legacy of william oldendorf," *Journal of Child Neurology*, vol. 25, pp. 508–517, Apr. 2010.

[4] T. M. Buzug, "Computed tomography," in *Springer Handbook of Medical Technology*, pp. 311–342, Springer Berlin Heidelberg, 2011.

[5] A. S. Brody, D. P. Frush, W. Huda, and R. L. B. and, "Radiation risk to children from computed tomography," *Pediatrics*, vol. 120, pp. 677–682, Sept. 2007.

[6] D. P. Frush, L. F. Donnelly, and N. S. Rosen, "Computed tomography and radiation risks: what pediatric health care providers should know," *Pediatrics*, vol. 112, no. 4, pp. 951–957, 2003.

[7] S. P. Power, F. Moloney, M. Twomey, K. James, O. J. O'Connor, and M. M. Maher, "Computed tomography and patient risk: Facts, perceptions and uncertainties," *World journal of radiology*, vol. 8, no. 12, p. 902, 2016.

[8] R. Webb, "New resonance," *Nature Physics*, vol. 4, pp. S10–S10, Feb. 2008.

[9] R. Damadian, "Tumor detection by nuclear magnetic resonance," *Science*, vol. 171, pp. 1151–1153, Mar. 1971.

[10] I. Young, "Nuclear magnetic resonance imaging," *Electronics and Power*, vol. 30, no. 3, pp. 205–210, 1984.

[11] R. R. Edelman, "The history of MR imaging as seen through the pages of radiology," *Radiology*, vol. 273, pp. S181–S200, Nov. 2014.

[12] A. G. Smith and C. R. Hill, "Imaging assessment of acute ischaemic stroke: a review of radiological methods," *The British Journal of Radiology*, p. 20170573, Dec. 2017.

[13] R. G. González, "Clinical MRI of acute ischemic stroke," *Journal of Magnetic Resonance Imaging*, vol. 36, pp. 259–271, July 2012.

[14] A. Ceccarelli, R. Bakshi, and M. Neema, "MRI in multiple sclerosis," *Current Opinion in Neurology*, vol. 25, pp. 402–409, Aug. 2012.

[15] U. W. Kaunzner and S. A. Gauthier, "MRI in the assessment and monitoring of multiple sclerosis: an update on best practice," *Therapeutic Advances in Neurological Disorders*, vol. 10, pp. 247–261, May 2017.

[16] J. J. Kim and A. D. Gean, "Imaging for the diagnosis and management of traumatic brain injury," *Neurotherapeutics*, vol. 8, pp. 39–53, Jan. 2011.

[17] C. Eierud, R. C. Craddock, S. Fletcher, M. Aulakh, B. King-Casas, D. Kuehl, and S. M. LaConte, "Neuroimaging after mild traumatic brain injury: Review and meta-analysis," *NeuroImage: Clinical*, vol. 4, pp. 283–294, 2014.

[18] B. Levine, N. Kovacevic, E. I. Nica, G. Cheung, F. Gao, M. L. Schwartz, and S. E. Black, "The toronto traumatic brain injury study: Injury severity and quantified MRI," *Neurology*, vol. 70, pp. 771–778, Mar. 2008.

[19] M. K. Abd-Ellah, A. I. Awad, A. A. Khalaf, and H. F. Hamed, "A review on brain tumor diagnosis from MRI images: Practical implications, key achievements, and lessons learned," *Magnetic Resonance Imaging*, vol. 61, pp. 300–318, Sept. 2019.

[20] G. S. Young, "Advanced MRI of adult brain tumors," *Neurologic Clinics*, vol. 25, pp. 947–973, Nov. 2007.

[21] T. Rüber, B. David, and C. E. Elger, "MRI in epilepsy: clinical standard and evolution," *Current Opinion in Neurology*, vol. 31, pp. 223–231, Apr. 2018.

[22] I. Wang, A. Bernasconi, B. Bernhardt, H. Blumenfeld, F. Cendes, Y. Chinvarun, G. Jackson, V. Morgan, S. Rampp, A. E. Vaudano, and P. Federico, "MRI essentials in epileptology: a review from the ILAE imaging taskforce," *Epileptic Disorders*, vol. 22, pp. 421–437, Aug. 2020.

[23] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in alzheimer disease," *Nature Reviews Neurology*, vol. 6, pp. 67–77, Feb. 2010.

[24] P. Vemuri and C. R. Jack, "Role of structural mri in alzheimer's disease," *Alzheimer's research & therapy*, vol. 2, no. 4, pp. 1–10, 2010.

[25] T. R. Stoub, M. Bulgakova, S. Leurgans, D. A. Bennett, D. Fleischman, D. A. Turner, and L. deToledo Morrell, "MRI predictors of risk of incident alzheimer disease: A longitudinal study," *Neurology*, vol. 64, pp. 1520–1524, May 2005.

[26] P. Mahlknecht, A. Hotter, A. Hussl, R. Esterhammer, M. Schocke, and K. Seppi, "Significance of mri in diagnosis and differential diagnosis of parkinson's disease," *Neurodegenerative Diseases*, vol. 7, no. 5, pp. 300–318, 2010.

[27] B. Heim, F. Krismer, R. De Marzi, and K. Seppi, "Magnetic resonance imaging for the diagnosis of parkinson's disease," *Journal of neural transmission*, vol. 124, pp. 915–964, 2017.

[28] M. Stern, B. Braffman, B. Skolnick, H. Hurtig, and R. Grossman, "Magnetic resonance imaging in parkinson's disease and parkinsonian syndromes," *Neurology*, vol. 39, no. 11, pp. 1524–1524, 1989.

[29] N. Georgiou-Karistianis, R. Scahill, S. J. Tabrizi, F. Squitieri, and E. Aylward, "Structural mri in huntington's disease and recommendations for its potential use in clinical trials," *Neuroscience & Biobehavioral Reviews*, vol. 37, no. 3, pp. 480–490, 2013.

[30] S. Dimou, R. Battisti, D. F. Hermens, and J. Lagopoulos, "A systematic review of functional magnetic resonance imaging and diffusion tensor imaging modalities used in presurgical planning of brain tumour resection," *Neurosurgical review*, vol. 36, pp. 205–214, 2013.

[31] S. Sunaert, "Presurgical planning for tumor resectioning," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 23, no. 6, pp. 887–905, 2006.

[32] D. L. Kent, "The clinical efficacy of magnetic resonance imaging in neuroimaging," *Annals of Internal Medicine*, vol. 120, p. 856, May 1994.

[33] P. Hockings, N. Saeed, R. Simms, N. Smith, M. G. Hall, J. C. Waterton, and S. Sourbron, "MRI biomarkers," in *Advances in Magnetic Resonance Technology and Applications*, pp. liii–lxxxvi, Elsevier, 2020.

[34] K. Strimbu and J. A. Tavel, "What are biomarkers?," *Current Opinion in HIV and AIDS*, vol. 5, pp. 463–466, Nov. 2010.

[35] B. C. Dickerson, T. R. Stoub, R. C. Shah, R. A. Sperling, R. J. Killiany, M. S. Albert, B. T. Hyman, D. Blacker, and L. deToledo Morrell, "Alzheimer-signature MRI biomarker predicts AD dementia in cognitively normal adults," *Neurology*, vol. 76, pp. 1395–1402, Apr. 2011.

[36] E. Sarasso, S. Basaia, C. Cividini, T. Stojkovic, I. Stankovic, N. Piramide, A. Tomic, V. Markovic, E. Stefanova, V. S. Kostic, M. Filippi, and F. Agosta, "MRI biomarkers of freezing of gait development in parkinson's disease," *npj Parkinson's Disease*, vol. 8, Nov. 2022.

[37] T. Mitchell, S. Lehéricy, S. Y. Chiu, A. P. Strafella, A. J. Stoessl, and D. E. Vaillancourt, "Emerging neuroimaging biomarkers across disease stage in parkinson disease," *JAMA Neurology*, vol. 78, p. 1262, Oct. 2021.

[38] A. Sampedro, N. Ibarretxe-Bilbao, J. Peña, A. Cabrera-Zubizarreta, P. Sánchez, A. Gómez-Gastiasoro, N. Iriarte-Yoller, C. Pavón, M. Tous-Espelosin, and N. Ojeda, "Analyzing structural and functional brain changes related to an

integrative cognitive remediation program for schizophrenia: A randomized controlled trial," *Schizophrenia Research*, vol. 255, pp. 82–92, May 2023.

[39] S. Ogawa and T.-M. Lee, "Magnetic resonance imaging of blood vessels at high fields:in vivo andin vitro measurements and image simulation," *Magnetic Resonance in Medicine*, vol. 16, pp. 9–18, Oct. 1990.

[40] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature*, vol. 412, pp. 150–157, July 2001.

[41] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar mri," *Magnetic Resonance in Medicine*, vol. 34, pp. 537–541, Oct. 1995.

[42] Global Burden of Disease Collaborative Network, "Global burden of disease study 2019 (gbd 2019) reference life table," 2021.

[43] J. L. Matson and A. M. Kozlowski, "The increasing prevalence of autism spectrum disorders," *Research in Autism Spectrum Disorders*, vol. 5, pp. 418–425, Jan. 2011.

[44] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *The Lancet*, vol. 392, pp. 508–520, Aug. 2018.

[45] B. Zablotsky, M. D. Bramlett, and S. J. Blumberg, "The co-occurrence of autism spectrum disorder in children with ADHD," *Journal of Attention Disorders*, vol. 24, pp. 94–103, June 2017.

[46] A. McCrimmon and K. Rostad, "Test review: Autism diagnostic observation schedule, second edition (ADOS-2) manual (part II): Toddler module," *Journal of Psychoeducational Assessment*, vol. 32, pp. 88–92, Dec. 2013.

[47] C. Lord, M. Rutter, and A. L. Couteur, "Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *Journal of Autism and Developmental Disorders*, vol. 24, pp. 659–685, Oct. 1994.

[48] S. Timimi, D. Milton, V. Bovell, S. Kapp, and G. Russell, "Deconstructing diagnosis: Four commentaries on a diagnostic tool to assess individuals for autism spectrum disorders," *Autonomy (Birmingham, England)*, vol. 1, no. 6, 2019.

[49] A. Hanif, X. Zhang, and S. Wood, "A survey on explainable artificial intelligence techniques and challenges," in *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*, IEEE, Oct. 2021.

[50] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?," 2017.

[51] A. Mohanty and S. Mishra, "A comprehensive study of explainable artificial intelligence in healthcare," in *Augmented Intelligence in Healthcare: A Pragmatic and Integrated Analysis*, pp. 475–502, Springer Nature Singapore, 2022.

[52] A. Adadi and M. Berrada, "Explainable AI for healthcare: From black box to interpretable models," in *Embedded Systems and Artificial Intelligence*, pp. 327–337, Springer Singapore, 2020.

[53] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," 2019.

[54] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," 2017.

[55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, oct 2019.

[56] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems*, vol. 41, pp. 647–665, Aug. 2013.

[57] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[58] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," 2018.

[59] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M. M. Höhne, "Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond," *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023.

[60] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, "On the (in)fidelity and sensitivity for explanations," 2019.

[61] S. Dasgupta, N. Frost, and M. Moshkovitz, "Framework for evaluating faithfulness of local explanations," 2022.

[62] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, feb 2018.

[63] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, June 2006.

[64] J. Theiner, E. Mueller-Budack, and R. Ewerth, "Interpretable semantic photo geolocation," 2021.

[65] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," 2020.

[66] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," 2020.

[67] F. Yan, Y. Chen, Y. Xia, Z. Wang, and R. Xiao, "An explainable brain tumor detection framework for MRI analysis," *Applied Sciences*, vol. 13, p. 3438, Mar. 2023.

[68] S. Hossain, A. Chakrabarty, T. R. Gadekallu, M. Alazab, and M. J. Piran, "Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14, 2023.

[69] D. D. Gunashekar, L. Bielak, L. Hägele, B. Oerther, M. Benndorf, A.-L. Grosu, T. Brox, C. Zamboglou, and M. Bock, "Explainable AI for CNN-based prostate tumor segmentation in multi-parametric MRI correlated to whole mount histopathology," *Radiation Oncology*, vol. 17, Apr. 2022.

[70] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial Intelligence in Medicine*, vol. 94, pp. 42–53, Mar. 2019.

[71] M. R. Hassan, M. F. Islam, M. Z. Uddin, G. Ghoshal, M. M. Hassan, S. Huda, and G. Fortino, "Prostate cancer classification from ultrasound and MRI images using deep learning based explainable artificial intelligence," *Future Generation Computer Systems*, vol. 127, pp. 462–472, Feb. 2022.

[72] Z. Zhang, P. Chen, M. McGough, F. Xing, C. Wang, M. Bui, Y. Xie, M. Sapkota, L. Cui, J. Dhillon, N. Ahmad, F. K. Khalil, S. I. Dickinson, X. Shi, F. Liu, H. Su, J. Cai, and L. Yang, "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nature Machine Intelligence*, vol. 1, pp. 236–245, May 2019.

[73] K. Schutte, O. Moindrot, P. Hérent, J.-B. Schiratti, and S. Jégou, "Using stylegan for visual interpretability of deep learning models on medical images," 2021.

[74] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu, "Visual feature attribution using wasserstein gans," 2018.

[75] W. Silva, A. Poellinger, J. S. Cardoso, and M. Reyes, "Interpretability-guided content-based medical image retrieval," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 305–314, Springer International Publishing, 2020.

[76] B. Hu, B. Vasu, and A. Hoogs, "X-mir: Explainable medical image retrieval," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 440–450, January 2022.

[77] J. D. Arias-Londono, J. A. Gomez-Garcia, L. Moro-Velazquez, and J. I. Godino-Llorente, "Artificial intelligence applied to chest x-ray images for the automatic detection of COVID-19. a thoughtful evaluation approach," *IEEE Access*, vol. 8, pp. 226811–226827, 2020.

[78] A. M. Ismael and A. Şengür, "Deep learning approaches for COVID-19 detection based on chest x-ray images," *Expert Systems with Applications*, vol. 164, p. 114054, Feb. 2021.

[79] Y. Du, Z. Fu, and V. D. Calhoun, "Classification and prediction of brain disorders using functional connectivity: Promising but challenging," *Frontiers in Neuroscience*, vol. 12, Aug. 2018.

[80] X. Li, N. C. Dvornek, J. Zhuang, P. Ventola, and J. S. Duncan, "Brain biomarker interpretation in ASD using deep learning and fMRI," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 206–214, Springer International Publishing, 2018.

[81] Y. Zhao, F. Ge, S. Zhang, and T. Liu, "3d deep convolutional neural network revealed the value of brain network overlap in differentiating autism spectrum disorder from healthy controls," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 172–180, Springer International Publishing, 2018.

[82] R. M. Thomas, S. Gallo, L. Cerliani, P. Zhutovsky, A. El-Gazzar, and G. van Wingen, "Classifying autism spectrum disorder using the temporal statistics of resting-state functional MRI data with 3d convolutional neural networks," *Frontiers in Psychiatry*, vol. 11, May 2020.

[83] M. S. Ahammed, S. Niu, M. R. Ahmed, J. Dong, X. Gao, and Y. Chen, "DarkASDNet: Classification of ASD on functional MRI using deep neural network," *Frontiers in Neuroinformatics*, vol. 15, June 2021.

[84] Y. Wang, J. Wang, F.-X. Wu, R. Hayrat, and J. Liu, "AIMAFE: Autism spectrum disorder identification with multi-atlas deep feature representation and ensemble learning," *Journal of Neuroscience Methods*, vol. 343, p. 108840, Sept. 2020.

[85] T. M. Epalle, Y. Song, Z. Liu, and H. Lu, "Multi-atlas classification of autism spectrum disorder with hinge loss trained deep architectures: Abide i results," *Applied Soft Computing*, vol. 107, p. 107375, 2021.

[86] M. R. Lamani, P. J. Benadit, and K. Vaithinathan, "Multi-atlas graph convolutional networks and convolutional recurrent neural networks-based ensemble learning for classification of autism spectrum disorders," *SN Computer Science*, vol. 4, Feb. 2023.

[87] Y. Wang, J. Liu, Y. Xiang, J. Wang, Q. Chen, and J. Chong, "MAGE: Automatic diagnosis of autism spectrum disorders using multi-atlas graph convolutional

networks and ensemble learning," *Neurocomputing*, vol. 469, pp. 346–353, Jan. 2022.

[88] A. Qayyum, M. K. A. Ahamed Khan, A. Benzinou, M. Mazher, M. Ramasamy, K. Aramugam, C. Deisy, S. Sridevi, and M. Suresh, "An efficient 1dcnn–lstm deep learning model for assessment and classification of fmri-based autism spectrum disorder," in *Innovative Data Communication Technologies and Application* (J. S. Raj, K. Kamel, and P. Lafata, eds.), (Singapore), pp. 1039–1048, Springer Nature Singapore, 2022.

[89] W. Jiang, S. Liu, H. Zhang, X. Sun, S.-H. Wang, J. Zhao, and J. Yan, "CNNG: A convolutional neural networks with gated recurrent units for autism spectrum disorder classification," *Frontiers in Aging Neuroscience*, vol. 14, July 2022.

[90] L. Kang, Z. Gong, J. Huang, and J. Xu, "Autism spectrum disorder recognition based on machine learning with roi time-series," *NeuroImage: Clinical*, 2023.

[91] T. Iidaka, "Resting state functional magnetic resonance imaging and neural network classified autism and control," *Cortex*, vol. 63, pp. 55–67, Feb. 2015.

[92] G. Brihadiswaran, D. Haputhanthri, S. Gunathilaka, D. Meedeniya, and S. Jayarathna, "EEG-based processing and classification methodologies for autism spectrum disorder: A review," *Journal of Computer Science*, vol. 15, pp. 1161–1183, Aug. 2019.

[93] C. Gerloff, K. Konrad, J. Kruppa, M. Schulte-Rüther, and V. Reindl, "Autism spectrum disorder classification based on interpersonal neural synchrony: Can classification be improved by dyadic neural biomarkers using unsupervised graph representation learning?," in *Lecture Notes in Computer Science*, pp. 147–157, Springer Nature Switzerland, 2022.

[94] M. Leming, J. M. Górriz, and J. Suckling, "Ensemble deep learning on large, mixed-site fMRI datasets in autism and other tasks," *International Journal of Neural Systems*, vol. 30, p. 2050012, Apr. 2020.

[95] M. Garcia and C. Kelly, "Towards 3d deep learning for neuropsychiatry: predicting autism diagnosis using an interpretable deep learning pipeline applied to minimally processed structural MRI data," Oct. 2022.

[96] Y. Chen, A. Liu, X. Fu, J. Wen, and X. Chen, "An invertible dynamic graph convolutional network for multi-center ASD classification," *Frontiers in Neuroscience*, vol. 15, Feb. 2022.

[97] X. Li, N. C. Dvornek, X. Papademetris, J. Zhuang, L. H. Staib, P. Ventola, and J. S. Duncan, "2-channel convolutional 3d deep neural network (2cc3d) for fmri analysis: Asd classification and feature learning," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 1252–1255, IEEE, 2018.

[98] K. Supekar, S. Ryali, R. Yuan, D. Kumar, C. de los Angeles, and V. Menon, "Robust, generalizable, and interpretable artificial intelligence–derived brain fingerprints of autism and social communication symptom severity," *Biological Psychiatry*, vol. 92, pp. 643–653, Oct. 2022.

[99] M. Yang, M. Cao, Y. Chen, Y. Chen, G. Fan, C. Li, J. Wang, and T. Liu, "Large-scale brain functional network integration for discrimination of autism using a 3-d deep learning model," *Frontiers in Human Neuroscience*, vol. 15, June 2021.

[100] G. Wen, P. Cao, H. Bao, W. Yang, T. Zheng, and O. Zaiane, "MVS-GCN: A prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis," *Computers in Biology and Medicine*, vol. 142, p. 105239, Mar. 2022.

[101] X. Li, N. C. Dvornek, Y. Zhou, J. Zhuang, P. Ventola, and J. S. Duncan, "Graph neural network for interpreting task-fMRI biomarkers," in *Lecture Notes in Computer Science*, pp. 485–493, Springer International Publishing, 2019.

[102] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," 2013.

[103] L. S. Shapley, *A Value for N-Person Games*. Santa Monica, CA: RAND Corporation, 1952.

[104] C. Molnar, *Interpretable Machine Learning*. 2 ed., 2022.

[105] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.

[106] D. Erhan, Y. Bengio, A. C. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," 2009.

[107] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," 2018.

[108] A. D. Martino, C.-G. Yan, *et al.*, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular Psychiatry*, vol. 19, pp. 659–667, June 2013.

[109] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. S. Khundrakpam, J. D. Lewis, Q. Li, M. Milham, *et al.*, "The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives," *Frontiers in Neuroinformatics*, vol. 7, p. 27, 2013.

[110] C. Yan and Y. Zang, "Dparsf: a matlab toolbox for" pipeline" data analysis of resting-state fmri," *Frontiers in systems neuroscience*, p. 13, 2010.

[111] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, no. 1, pp. 95–113, 2007.

[112] E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot, "Automated anatomical labelling atlas 3," *NeuroImage*, vol. 206, p. 116189, 2020.

[113] R. C. Craddock, G. James, P. E. Holtzheimer III, X. P. Hu, and H. S. Mayberg, "A whole brain fmri atlas generated via spatially constrained spectral clustering," *Human Brain Mapping*, vol. 33, no. 8, pp. 1914–1928, 2012.

[114] N. U. F. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar, K. A. Barnes, J. W. Dubis, E. Feczko, R. S. Coalson, J. R. Pruett, D. M. Barch, S. E. Petersen, and B. L. Schlaggar, "Prediction of individual brain maturity using fmri," *Science*, vol. 329, no. 5997, pp. 1358–1361, 2010.

[115] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.

[116] E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot, "Automated anatomical labelling atlas 3," *NeuroImage*, vol. 206, p. 116189, 2020.

[117] E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot, "Automated anatomical labelling atlas 3," *NeuroImage*, vol. 206, p. 116189, 2020.

[118] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 731–737, 1997.

[119] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, Aug. 2007.

[120] S. B. Eickhoff, K. E. Stephan, H. Mohlberg, C. Grefkes, G. R. Fink, K. Amunts, and K. Zilles, "A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data," *NeuroImage*, vol. 25, pp. 1325–1335, May 2005.

[121] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, pp. 968–980, July 2006.

[122] J. L. Lancaster, M. G. Woldorff, L. M. Parsons, M. Liotti, C. S. Freitas, L. Rainey, P. V. Kochunov, D. Nickerson, S. A. Mikiten, and P. T. Fox, "Automated talairach atlas labels for functional brain mapping," *Human Brain Mapping*, vol. 10, no. 3, pp. 120–131, 2000.

[123] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,

Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[124] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[125] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[126] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2023.

[127] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015.

[128] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

[129] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[130] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, Aug. 2021.

[131] J. Deng, M. R. Hasan, M. Mahmud, M. M. Hasan, K. A. Ahmed, and M. Z. Hossain, "Diagnosing autism spectrum disorder using ensemble 3d-CNN: A preliminary study," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, Oct. 2022.

[132] E. B. Becker and C. J. Stoodley, "Autism spectrum disorder and the cerebellum," in *International Review of Neurobiology*, pp. 1–34, Elsevier, 2013.

[133] Z. Long, X. Duan, D. Mantini, and H. Chen, "Alteration of functional connectivity in autism spectrum disorder: effect of age and anatomical distance," *Scientific Reports*, vol. 6, May 2016.

[134] M. Assaf, K. Jagannathan, V. D. Calhoun, L. Miller, M. C. Stevens, R. Sahl, J. G. O'Boyle, R. T. Schultz, and G. D. Pearlson, "Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients," *NeuroImage*, vol. 53, pp. 247–256, Oct. 2010.

[135] C. S. Monk, S. J. Peltier, J. L. Wiggins, S.-J. Weng, M. Carrasco, S. Risi, and C. Lord, "Abnormalities of intrinsic functional connectivity in autism spectrum disorders,," *NeuroImage*, vol. 47, pp. 764–772, Aug. 2009.

[136] C. L. Keown, P. Shih, A. Nair, N. Peterson, M. E. Mulvey, and R.-A. Müller, "Local functional overconnectivity in posterior brain regions is associated with symptom severity in autism spectrum disorders," *Cell Reports*, vol. 5, pp. 567–572, Nov. 2013.

[137] S. J. Ebisch, V. Gallese, R. M. Willems, D. Mantini, W. B. Groen, G. L. Romani, J. K. Buitelaar, and H. Bekkering, "Altered intrinsic functional connectivity of anterior and posterior insula regions in high-functioning participants with autism spectrum disorder," *Human Brain Mapping*, vol. 32, pp. 1013–1028, July 2010.

[138] S. Baron-Cohen, H. Ring, S. Wheelwright, E. Bullmore, M. Brammer, A. Simmons, and S. Williams, "R.(1999). social intelligence in the normal autistic brain: An fmri study," *European Journal of Neuroscience*, vol. 11, pp. 1981–1898.

[139] K. Pierce, R.-A. Müller, J. Ambrose, G. Allen, and E. Courchesne, "Face processing occurs outside the fusiformface area'in autism: evidence from functional mri," *Brain*, vol. 124, no. 10, pp. 2059–2073, 2001.

[140] M. A. Just, V. L. Cherkassky, T. A. Keller, and N. J. Minshew, "Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity," *Brain*, vol. 127, no. 8, pp. 1811–1821, 2004.

[141] C. Ashwin, S. Baron-Cohen, S. Wheelwright, M. O'Riordan, and E. T. Bullmore, "Differential activation of the amygdala and the 'social brain'during fearful face-processing in asperger syndrome," *Neuropsychologia*, vol. 45, no. 1, pp. 2–14, 2007.

[142] R. C. Philip, M. R. Dauvermann, H. C. Whalley, K. Baynham, S. M. Lawrie, and A. C. Stanfield, "A systematic review and meta-analysis of the fmri investigation of autism spectrum disorders," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 2, pp. 901–942, 2012.

[143] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. V. Leemput, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1993–2024, Oct. 2015.

[144] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, M. Bilello, P. Bilic, P. F. Christ, R. K. G. Do, M. J. Gollub, S. H. Heckers, H. Huisman, W. R. Jarnagin, M. K. McHugo, S. Napel, J. S. G.

Pernicka, K. Rhode, C. Tobon-Gomez, E. Vorontsov, J. A. Meakin, S. Ourselin, M. Wiesenfarth, P. Arbeláez, B. Bae, S. Chen, L. Daza, J. Feng, B. He, F. Isensee, Y. Ji, F. Jia, I. Kim, K. Maier-Hein, D. Merhof, A. Pai, B. Park, M. Perslev, R. Rezaiifar, O. Rippel, I. Sarasua, W. Shen, J. Son, C. Wachinger, L. Wang, Y. Wang, Y. Xia, D. Xu, Z. Xu, Y. Zheng, A. L. Simpson, L. Maier-Hein, and M. J. Cardoso, "The medical segmentation decathlon," *Nature Communications*, vol. 13, July 2022.

[145] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," 2019.

[146] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J.-R. Wen, J. Yuan, W. X. Zhao, and J. Zhu, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[147] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[148] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," 2022.

[149] W. Yin, S. Mostafa, and F. xiang Wu, "Diagnosis of autism spectrum disorder based on functional brain networks with deep learning," *Journal of Computational Biology*, vol. 28, pp. 146–165, Feb. 2021.

[150] F. Zhang, Y. Wei, J. Liu, Y. Wang, W. Xi, and Y. Pan, "Identification of autism spectrum disorder based on a novel feature selection method and variational autoencoder," *Computers in Biology and Medicine*, vol. 148, p. 105854, Sept. 2022.

[151] P. K. C. Prasad, Y. Khare, K. Dadi, P. K. Vinod, and B. R. Surampudi, "Deep learning approach for classification and interpretation of autism spectrum disorder," in *2022 International Joint Conference on Neural Networks (IJCNN)*, IEEE, July 2022.

[152] A. Padmanabhan, C. J. Lynch, M. Schaer, and V. Menon, "The default mode network in autism," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 2, pp. 476–486, Sept. 2017.

[153] L. Harrison and K. J. Friston, "Effective connectivity," *Statistical parametric mapping: the analysis of functional brain images*, pp. 508–521, 2007.

[154] T. S. Zarghami and K. J. Friston, "Dynamic effective connectivity," *NeuroImage*, vol. 207, p. 116453, Feb. 2020.

[155] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical

analysis of structural and functional systems," *Nature reviews neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[156] J.-P. Lachaux, A. Lutz, D. Rudrauf, D. Cosmelli, M. L. V. Quyen, J. Martinerie, and F. Varela, "Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 32, pp. 157–174, June 2002.

[157] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[158] A. D. Martino, D. O'Connor, B. Chen, K. Alaerts, J. S. Anderson, M. Assaf, J. H. Balsters, L. Baxter, A. Beggiato, S. Bernaerts, L. M. E. Blanken, S. Y. Bookheimer, B. B. Braden, L. Byrge, F. X. Castellanos, M. Dapretto, R. Delorme, D. A. Fair, I. Fishman, J. Fitzgerald, L. Gallagher, R. J. J. Keehn, D. P. Kennedy, J. E. Lainhart, B. Luna, S. H. Mostofsky, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, M. Solomon, R. Toro, C. J. Vaidya, N. Wenderoth, T. White, R. C. Craddock, C. Lord, B. Leventhal, and M. P. Milham, "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II," *Scientific Data*, vol. 4, Mar. 2017.

[159] H. Lee, S. T. Kim, and Y. M. Ro, "Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9*, pp. 21–29, Springer, 2019.

[160] A. Chowdhury, A. Santamaria-Pang, J. R. Kubricht, and P. Tu, "Emergent symbolic language based deep medical image classification," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE, Apr. 2021.

[161] Y. Tang, Y. Tang, Y. Zhu, J. Xiao, and R. M. Summers, "A disentangled generative model for disease decomposition in chest x-rays via normal image synthesis," *Medical Image Analysis*, vol. 67, p. 101839, Jan. 2021.

[162] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, p. 593, Mar. 2021.