Exploratory ABIDE Dataset Analysis for Autism Detection

Department of Computer Science

Master's in Computer Networks and Embedded Systems

Author :
Imane Chkifa

Supervisors :
Mrs. Aicha Majda
Mr. Moncef Garouani

Jury : Pr. Aicha Majda , FSJES Meknes
Pr. Moncef Garouani , IRIT- Université Toulouse Capitole
Pr. Abdelbaki El Belrhiti El Alaoui , FS Meknes
Pr. El Mehdi Ismaili Alaoui , FS Meknes

September 23, 2024

Abstract

Mental disorders are a major public health concern, affecting millions of people worldwide. The quest for reliable biomarkers for diagnosis and treatment remains a complex challenge. Autism Spectrum Disorder (ASD), a common psychiatric condition, is characterized by atypical patterns in cognitive, emotional, and social domains. Early and accurate diagnosis is crucial to enable effective interventions and improve outcomes for affected individuals.

In this study, we propose an innovative model based on the use of brain atlases (CC200, HO, and AAL) and feature selection, utilizing resting-state functional MRI data from the ABIDE I dataset. Our approach involves feature selection methods to the vectors obtained from flattened connectivity matrices derived from the atlases. These vectors are then used as input for a classifier.

Through cross-validation, we demonstrate that our model outperforms the most advanced methods available for this dataset, achieving an average accuracy of 99.66% in 10-fold cross-validation with the CC200 atlas, using the RFE method and the logistic regression classifier. We also perform feature selection using Lasso and information gain analysis for comparison.

Keywords : ASD detection, fMRI, ABIDE, rs-fMRI, Machine Learning , Deep Learning , BOLD , RFE , LASSO

Résumé

Les troubles mentaux représentent un problème majeur de santé publique, affectant des millions de personnes à travers le monde. La recherche de biomarqueurs fiables pour le diagnostic et le traitement de ces troubles reste un défi complexe. Le trouble du spectre de l'autisme (TSA), une affection psychiatrique courante, qui se caractérise par des schémas atypiques dans les domaines cognitif, émotionnel et social. Un diagnostic précoce et précis est crucial pour permettre des interventions efficaces et améliorer la vie des personnes touchées.

Dans cette étude, nous proposons un modèle qui repose sur l'utilisation d'atlas cérébraux (CC200, HO et AAL) et sur la sélection de caractéristiques extraites des données d'IRM fonctionnelle au repos issues du jeu de données ABIDE I. Notre approche consiste à appliquer des méthodes de sélection de caractéristiques aux vecteurs dérivés des matrices de connectivité aplaties obtenues à partir des atlas. Ces vecteurs sont ensuite utilisés comme entrée pour un classificateur ou un modèle de réseau neuronal profond (DNN).

Nous avons montré, grâce à la validation croisée, que notre modèle dépasse les méthodes les plus avancées disponibles pour ce jeu de données. Avec l'atlas CC200, en utilisant la méthode RFE pour sélectionner les caractéristiques, et un classificateur de régression logistique pour les classer, nous avons atteint une précision moyenne de 99,66% en validation croisée à 10 folds . Pour comparer les résultats, nous avons également sélectionné les caractéristiques en utilisant les méthodes Lasso et l'analyse du gain d'information

mot clés : détection TSA , IRMf, ABIDE, IRMf-rs, Apprentissage Automatique , Apprentissage profond , BOLD , ROI , RFE , LASSO

Acknowledgements

First and foremost, I thank **ALLAH**, the Almighty, for granting me the strength, patience, and wisdom to complete this work. It is through His guidance and grace that I overcame the challenges I faced during this project.

I am truly grateful to **my family** for their incredible support. I especially want to thank **my mother** for her love, patience, and optimism. Her encouragement and unwavering belief in me were essential in helping me complete this project.

I am deeply grateful to my supervisors, **Mrs. Aicha Majda** and **Mr. Moncef Garouani** , for their exceptional guidance, availability, and unwavering support. Their strictness, valuable advice, and kind guidance were essential throughout this project. In moments of doubt or when I was stuck, their assistance helped me find solutions and keep moving forward. Without them, I certainly would not have had the courage to undertake a project of this scale.

I also want to thank **Mr. Hicham El Akhal** for his generous time, invaluable insights, and thoughtful advice, which greatly enriched my work.

I am thankful to **my friends** for their kindness, constant encouragement, and moral support. Their friendship provided comfort during times of fatigue and discouragement, helping me navigate this period with ease.

Finally, I wish to thank all the people who, in one way or another, contributed to this work. Every act of support and word of encouragement mattered, and without each person's help, this project would not have come to life in the same way. To everyone who believed in me and this work, I offer my sincerest thanks.

Table of Contents

List of Acronyms

AAL	Automated Anatomical Labeling
ABIDE	Autism Brain Imaging Data Exchange
ADHD	Attention Deficit Hyperactivity Disorder
ADOS	Autism Diagnostic Observation Schedule
ADI-R	Autism Diagnostic Interview- Revised
AE	Autoencoder
ANOVA	Analysis Of Variance
aMRI	anatomical Magnetic Resonance Imaging
BOLD	Blood Oxygen Level Dependent
Caltech	California Institute of Technology
CARS-2	Childhood Autism Rating Scale- Second Edition
CCS	Connectome Computation System
CC200	Craddock 200
CC400	Craddock 400
CMU	Carnegie Mellon University
CNN	Convolutional Neural Network
CPAC	Configurable Pipeline for the Analysis of Connectomes
CSF	Connectome Computation System
DL	Deep Learning
DMN	Default Mode Network
DISCO	Diagnostic Interview for Social and Communication Disorders
DPARSF	Data Processing Assistant for Resting-State fMRI
DSM	Diagnostic and Statistical Mannual
DTI	Diffusion Tensor Imaging
EL	Ensemble Learning
EZ	Eickhoff-Zilles
FC	Functional Connectivity
FCNN	Fully Connected Neural Network
fMRI	Functional Magnetic Resonance Imaging
GCN	Graph Convolutional Networks
GNN	Graph Neural Networks
GAT	Graph Attention Network
HIPAA	Health Insurance Portability and Accountability Act
HC	Healthy Controls
HO	Harvard-Oxford
ICD	International Classification of Diseases
ICA	Independent Component Analysis

INDI	International Neuroimaging Data-sharing Initiative
INP	International Neuroimaging Data-sharing Initiative
KKI	Kennedy Krieger Institute
KNN	K-Nearest Neighbors
Leuven	University of Leuven
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MaxMun	Ludwig Maximilians University Munich
MEG	Magnetoencephalography
ML	Machine Learning
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NIAK	Neuroimaging Analysis Kit
NYU	NYU Langone Medical Center
OHSU	Oregon Health and Science University
Olin	Institute of Living at Hartford Hospital
PET	Positron Emission Tomography
PCA	Principal Component Analysis
PDD-NOS	Pervasive Developmental Disorder-Not Otherwise Specified
Pitt	University of Pittsburgh School of Medicine
RF	Random Forest
RFE	Recursive Feature Elimination
RS-fMRI	Resting-State Functional Magnetic Resonance Imaging
ReLU	Rectified Linear Unit
ROI	Region of Interest
SASL	Self-weighted Adaptive Structure Learning
SC	Structural Connectivity
SGD	Stochastic Gradient Descent
SBL	Social Brain Lab
sMRI	structural Magnetic Resonance Imaging
SPECT	Single-Photon Emission Computed Tomography
Stanford	Stanford University
SVM	Support Vector Machine
TC	Typical Controls
TD	Typically Developing
Trinity	Trinity Centre for Health Sciences
TT	Talaraich and Tournoux
UM	University of Michigan
USM	University of Utah School of Medicine
UCLA	University of California, Los Angeles
Yale	Yale Child Study Center
3Di	Developmental, Dimensional and Diagnostic Interview

List of Figures

List of Tables

Introduction

The brain, the central organ of our existence, governs all the vital processes that ensure our survival. It orchestrates functions as diverse as breathing, blood circulation, digestion, and temperature regulation. In addition to these physiological functions, the brain is the seat of our cognitive and emotional capacities. It is what allows us to perceive the world, reason, make decisions, and store memories. Every thought, every emotion, every movement we make is the result of a complex and perfectly orchestrated brain activity. However, despite this apparent perfection, the brain is also vulnerable. A simple disruption in this delicate system can lead to profound disorders, affecting our behavior, perception, and even our ability to interact with others.

Among the many disorders that can affect the brain, autism spectrum disorders are particularly complex and puzzling. Autism is not a disease that one "catches" or "cures". It is a neurodevelopmental disorder that manifests itself from childhood and persists throughout life. However, the way it manifests varies greatly from one person to another, making diagnosis difficult and often delayed. According to the Diagnostic and Statistical Manual of mental disorders (DSM), ASD includes several subtypes or categories, including Autistic Disorder, Asperger's Syndrome, and Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS). Each type presents a unique set of symptoms and degrees of severity, adding an extra layer of complexity to assessment and diagnosis.

The challenge of diagnosing autism lies not only in the diversity of symptoms but also in the ambiguity of underlying causes. Genetic, environmental, and neurobiological factors combine in complex ways to give rise to this spectrum of disorders. Due to this complexity, it is often difficult to pinpoint the exact causes of ASD and, consequently, to develop effective treatment strategies. However, recent advances in neuroscience offer new opportunities to explore the brain mechanisms involved in autism.

Functional Magnetic Resonance Imaging (fMRI) is one of the most promising technologies in this field. By allowing real-time observation of brain activity, fMRI provides valuable information on how the brains of individuals with ASD function differently from those of others. Through this data, researchers can identify patterns of brain activity associated with autism, which could lead to earlier and more accurate diagnostic methods.

In this research work, we explore the development of an early diagnosis system for autism based on fMRI data, specifically time series extracted using brain atlases. The objective of this work is to prepare a diagnostic support system capable of detecting autism well before the appearance of its first clinical symptoms, which would allow intervention at a

stage where treatments can be more effective.

To achieve this ambitious goal, it is essential to adapt diagnostic methods to make them accessible and safe for young children, who are the target population for this type of early detection. This adaptation requires a rigorous evaluation of the safety and effectiveness of the tools used to ensure that they pose no risk to young patients.

Once these safety aspects are ensured, we will implement advanced data analysis methods, combining signal processing techniques and machine learning. These methods will enable the extraction and highlighting of key information from fMRI data, that is often hidden within complex time series structures.

Ultimately, the goal is to create a diagnostic system that not only classifies the tested subjects but does so with a level of precision that is clinically useful. Such a system could play a crucial role in improving patient care by enabling faster and more targeted intervention, thereby offering better prospects for children at risk of developing autism spectrum disorders.

This report is structured as follows : Chapter 1 provides a general overview of the autism spectrum, highlighting the crucial importance of neuro-imaging techniques for the early detection of neuro-developmental disorders. This section not only outlines the main characteristics of autism but also addresses the challenges associated with its diagnosis, particularly in young children, where rapid identification can significantly improve clinical outcomes. Chapter 2 offers a detailed description of the ABIDE (Autism Brain Imaging Data Exchange) dataset and its components. This dataset, which serves as an essential resource for researchers in the field, includes a vast array of neuro-imaging data collected from multiple sites, allowing for an in-depth analysis of individual variations in the brains of people with autism spectrum disorders. We also explore the different categories of available data, such as structural and functional MRI images, and how these can be utilized to enhance the understanding of ASD. Chapter 3 discusses the various data representations that can be used in machine learning and deep learning models for autism detection. We examine several approaches, such as functional connectivity matrices, extracted time series, and representations based on specific brain features. Additionally, this section reviews some models recently employed for the diagnosis of autism. Chapter 4 introduces the machine learning and deep learning models that we propose for autism detection. We also discuss the importance of feature selection to improve the accuracy of classification models.

Chapter 1

Autism Spectrum Disorder

1.1 Definition

Autism Spectrum Disorder (ASD) encompasses a range of neurodevelopmental conditions that affect communication, social interaction, and behavior. The term "spectrum" reflects the diversity of symptoms and severity levels seen in individuals with autism. Some may have high intellectual abilities, while others may face significant challenges in communication and daily living.

Key features of ASD include difficulties with communication, repetitive behaviors, and altered social interactions. Autistic individuals might struggle with both verbal and non-verbal communication, such as speech delays or limited understanding of social cues. They may also exhibit repetitive movements or an intense focus on specific interests. Social interactions can be challenging, as autism can affect one's ability to understand social norms, build relationships, or respond appropriately in social situations.

Additionally, individuals with autism might have heightened or reduced sensitivity to sensory stimuli, such as sounds, lights, textures, or tastes, which can make certain experiences, like loud noises, particularly distressing. In general, the characteristics of autism can be observed before the age of three. Abnormalities in social interaction, language used in social communication, or symbolic and imaginative play can be noted early.

However, due to factors such as limited parental awareness and the availability of diagnostic centers, the average age at which autism is diagnosed is often around five years old (Daniels and Mandell, 2013 [1]).

The precise causes of autism remain unclear, but it is believed to result from a combination of genetic , prenatal and environmental factors. Diagnosis usually occurs in childhood, often around ages two to three, but can be made later. Evaluations typically involve behavioral observations and interviews with parents or caregivers.

There is no cure for autism, but various interventions can enhance the quality of life for individuals with autism. These may include behavioral therapy, speech therapy, specialized education, and sometimes medication to address specific symptoms like

anxiety or hyperactivity. Because each person with autism is unique, personalized care is crucial, as different approaches may work for different individuals.

In adulthood, some individuals with autism may experience a reduction in the severity of core symptoms compared to their early development. This phenomenon, often referred to as "amelioration," suggests that certain symptoms may become milder as individuals grow older. The researches (Billstedt, Gillberg, and Gillberg, 2007 [2] ; Taylor and Seltzer, 2010) [3] indicates that this reduction might be due to the maturation and stabilization of brain processes, as well as the normalization of brain growth that often occurs after childhood overgrowth .

1.2 Key dates related to autism

- **1911** : For the first time, the term “autistic” appears in a psychiatry treatise and is associated with schizophrenia.
- **1943** : American child psychiatrist Leo Kanner describes the clinical picture of early infantile autism.
- **1944** : Austrian child psychiatrist Hans Asperger uses the term “autism” to describe patients with social interaction difficulties but without impairments in language and cognitive functions.
- **1950-1960** : American psychoanalyst Bruno Bettelheim hypothesizes that the behavior of certain “refrigerator” mothers is the cause of their child’s autism.
- **1960-1970** : For most experts, autism is considered a psychosis, a mental illness acquired through psychoanalysis.
- **1970-1980** : American and French researchers use encephalography and hypothesize that autism is related to early sensory perception disorders. Increasing studies focus on autism in twins or related children, revealing the significant contribution of genetics.
- **2001** : An American report concludes that individualized education programs are the best approach for managing autism. In France, psychotherapeutic treatments are still predominantly favored.
- **2003** : For the first time, genetic mutations are precisely identified in autistic children by Thomas Bourgeron’s team at the Pasteur Institute in collaboration with French psychiatrists (Pr Marion Leboyer) and Swedish (Pr Christopher Gillberg). Other mutations will be discovered later.
- **2005-2007** : First autism plan in France.
- **2008** : Creation of the first animal models carrying mutations associated with autism spectrum disorder (ASD).
- **2008-2010** : Second autism plan.
- **2012-2018** : EU-AIMS project.
- **2013** : New version of DSM-5 (Diagnostic and Statistical Manual of Mental Disorders) with the use of the autism spectrum disorder diagnosis.

- **2013-2017** : Third autism plan (early screening and schooling of children with ASD).
- **2018-2024** : AIMS-2-Trials project.
- **2018-2022** : Fourth autism plan.
- **2019** : The official SPARK list enumerates 141 genes associated with autism.

1.3 Factors Affecting the Risk of Autism

Risk factors for the development of autism include genetic, prenatal, and gender differences. Genetics play a crucial role, with increased risk for children born to older parents. If one child is autistic, there is between a 2% and 18% chance that their sibling will also be affected. For identical twins, if one is autistic, the other is affected 36% to 95% of the time, while for non-identical twins, the risk is about 31%.

Since Leo Kanner first described autism in 1943, research has revealed that genetic mutations play an important role. In 2003, Thomas Bourgeron and his team at the Pasteur Institute discovered genetic mutations on the X chromosome linked to autism, affecting synapses, the connections between neurons. In 2008, they found that some mutations affect melatonin production, a hormone regulating sleep, which is often deficient in autistic children [4].

Autism is diagnosed much more frequently in boys than in girls, with a generally accepted ratio of 4 boys to 1 girl, although studies show variations from 2 :1 to 7 :1. This might indicate diagnostic bias or a protective effect in girls.

Biological theories, such as the fetal testosterone theory, propose that elevated levels of testosterone in the amniotic fluid are associated with certain autistic traits. These traits can include larger brain volume, increased connectivity within the same brain hemisphere, an enlarged amygdala, and reduced connectivity between the hemispheres due to a smaller corpus callosum. The theory suggests that higher testosterone levels can enhance abilities in analyzing and understanding complex systems and patterns, but may also lead to a reduction in empathetic traits (Dooley et al., 2022) [5].

Furthermore, prenatal factors play a significant role in the risk of developing autism. Maternal drug use, including alcohol and medications like selective serotonin reuptake inhibitors and sodium valproate, has been linked to an increased risk of autism spectrum disorder (Bastaki, Alwan, and Zahir, 2020) [6]. Illnesses experienced by the mother during pregnancy, such as hypertension, obesity, asthma, diabetes, and autoimmune disorders, have also been associated with a higher likelihood of autistic symptoms in children.

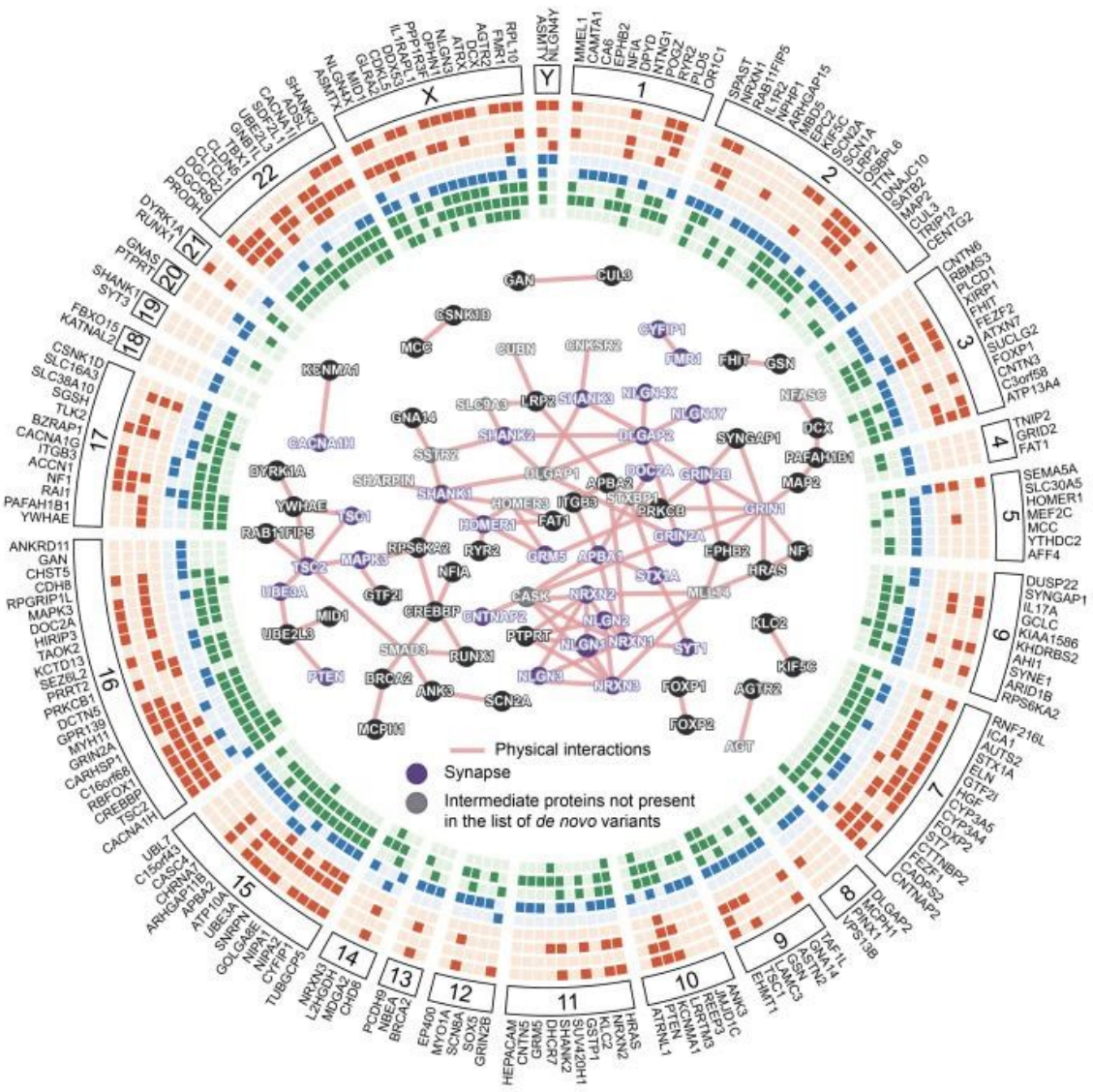

FIGURE 1.3.1 – Identification of Genes Linked to Autism Across the Whole Genome. [4]

Additionally, maternal stress and exposure to environmental pollutants have been found to contribute to the development of autism both during pregnancy and after birth. Moreover, several studies have shown that sociodemographic factors such as race and education level are associated with ASD diagnosis rates. An American study from 2011 found that white non-Hispanic mothers and those with higher education levels were more likely to have a child diagnosed with ASD. This trend may be explained by greater awareness of developmental issues in these populations and better access to healthcare services and appropriate evaluations, leading to higher diagnosis rates compared to other racial or socioeconomic groups [7].

In summary, the development of autism is influenced by a complex interaction of biological, genetic, and environmental factors. Understanding these factors, including hormonal influences such as elevated fetal testosterone levels and prenatal conditions such as drug use, illnesses, and exposure to pollutants, is crucial for grasping the full

range of elements that contribute to autism [8, 9].

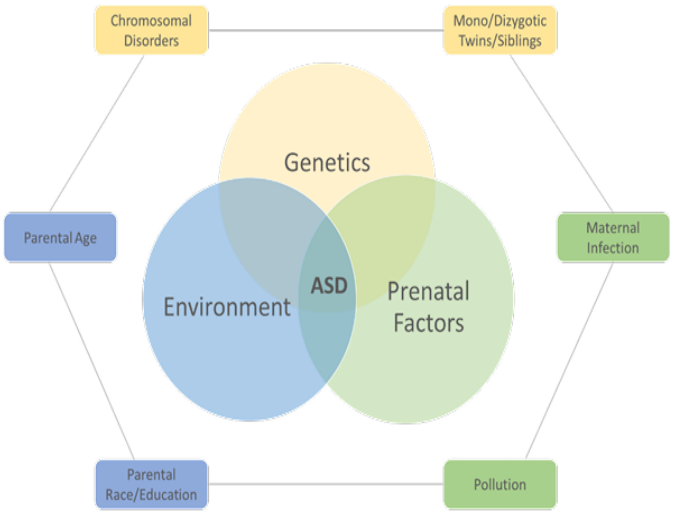

FIGURE 1.3.2 – Summary of main risk factors in ASD development. [7]

1.4 Importance of Early Diagnosis

Early detection of autism is crucial due to the ongoing ambiguity surrounding its causes. While there are theories suggesting a hereditary basis (Havdahl et al., 2021 [10]; Xie et al., 2020 [11]), others propose environmental factors as contributors [12, 13].

Research led by John Lewis at the Montreal Neurological Institute has highlighted a link between network inefficiencies and the severity of autism symptoms, emphasizing the value of early diagnosis. Identifying autism signs before behavioral changes become evident can facilitate earlier intervention, which is essential for improving outcomes.

Early intervention can be transformative for children with autism, offering benefits from various treatments, including speech and occupational therapy. It can enhance learning, communication, and social skills, and promote better brain development. Detecting autism early can help prevent some children from losing skills they've already learned. This loss, called developmental regression, happens in about 20% of kids with autism, usually between 1 and 3 years old. By addressing autism early, we provide the best chance for supporting healthy development and reducing the disorder's impact throughout life.

1.5 Diagnosis Methods and Challenges

Diagnosing autism is one of the most challenging tasks because it is so hard to identify. Right now, autism is mainly diagnosed using behavior tests and observations by family or people close to the person. Most of these evaluations are designed to catch

autism early, as it usually shows up in the first few years of life.

Some of the tools used to diagnose Autism Spectrum Disorder are semi-structured, while others are computerized. "Semi-structured" means the evaluation has a basic set of questions but allows the examiner to ask more questions based on the answers they get. These tests have a set structure but can be adjusted based on how the person responds. Here are some examples of semi-structured tests : ADI-R, ADOS, and DISCO (Noens and Nygren, 2012) [14].

The Autism Diagnostic Interview – Revised (ADI-R) uses information from parents or caregivers to assess the individual. It's suitable for people aged 18 months and older and follows the criteria of DSM-5 and ICD-11. It looks at social interaction, communication, and repetitive behaviors to help diagnose autism.

The Autism Diagnostic Observation Schedule (ADOS) assesses social communication, interaction, and play through direct observation. It has different modules based on the person's verbal skills, whether they are children or adults. Each module includes specific activities to help identify ASD behaviors.

DISCO, the Diagnostic Interview for Social and Communication Disorders, is used to evaluate someone suspected of autism over a longer period. It requires an evaluator who knows the subject well and can track their development over several years. Its goal is to understand the behavior patterns that develop as the person grows.

The main computerized tool for detecting ASD is **the Developmental, Dimensional, and Diagnostic Interview (3Di)** (Slappendel et al., 2016 [15]). It's designed for parents of autistic children of any age and evaluates the intensity of autistic symptoms as well as potential comorbidities. The 3Di utilizes a digital platform to administer the interview, collect and analyze responses, and generate detailed reports. This computerized approach ensures a precise and standardized evaluation of symptoms and associated conditions.

Another important test is **the Childhood Autism Rating Scale – Second Edition (CARS-2)** (Dawkins, Meyer, and Bourgondien, 2016 [16]). It helps differentiate between children with moderate or severe ASD and those with other cognitive disorders, including developmental delays. This tool is often used in research and helps with clinical diagnoses. It can be given by caregivers or clinicians.

However, these methods have some downsides. They mainly rely on subjective observations, which can introduce biases. Different people might see things differently based on their experience or stress level, which can lead to inaccurate diagnoses.
Also, these tests often need specialized training to be used and interpreted correctly. Evaluators need to be very skilled, which can limit how widely these tests are used.
Lastly, behavior-based evaluations might not capture all the details of symptoms because behaviors can change based on context, emotions, or the relationship with the evaluator. This means a diagnosis based only on behavior might not be complete or accurate. That's

why researchers are looking for new ways to diagnose autism that are earlier and more reliable.

1.6 Brain Development and Autism

Autism Spectrum Disorder presents a wide range of symptoms, which can vary greatly in severity and combination. This variability makes it challenging to diagnose ASD accurately and to identify its origins. Each symptom might be associated with different areas of the brain, complicating the task of pinpointing specific brain deficits.
In early childhood, the brain undergoes a crucial process called neural pruning. During this process, unnecessary and non-functional neurons are removed to make the remaining neurons work more efficiently. In typical brain development, this leads to a reduction in the number of neurons while the remaining ones form complex connections. However, in children with ASD, this pruning process does not work correctly, resulting in an excess of neurons.

Stiles and Jernigan (2010) [17] , describe two key aspects of brain development. The first is natural cell death, where the brain eliminates more than half of its neurons. The second aspect is synaptic exuberance and pruning : initially, the brain forms many connections between neurons, but then it systematically removes up to half of these connections to fine-tune brain function.

In contrast, Courchesne et al. (2011) [18], found that individuals with autism, particularly males, often have a significantly higher number of neurons in the prefrontal cortex—about 67% more than in typical brains. This excess of neurons adds complexity to understanding brain development in autism.

Additionally, research on infants at high risk for ASD or those diagnosed later reveals that these children often experience unusually rapid brain growth during their first year. This accelerated growth is usually followed by a plateau or slowing down in the second year. These changes have been observed in several brain regions, including the frontal cortex, temporal cortex, and amygdala [19, 20, 21]

1.7 Brain Imaging Techniques and ASD Detection

Research on Autism Spectrum Disorder aims not only to find a cure but also to accelerate the detection process using non-invasive methods, especially since the affected individuals are children. This requires careful selection of diagnostic techniques and methods.

Brain imaging techniques hold promise for diagnosing autism because they are not limited by age, allowing for earlier detection. We have access to various anatomical and functional neuroimaging methods, such as X-ray, MRI, MEG, fMRI, PET, and SPECT.

Anatomical imaging provides insights into the brain's structure and its components, whereas functional neuroimaging focuses on understanding the brain's cognitive organization and dynamic processes.

In young children, developmental rates can vary significantly, which can make anatomical imaging unreliable for detecting autism [22]. Instead, studying brain function, particularly network communication, might offer more meaningful insights into the differences between autistic individuals and controls. In this context, fMRI is particularly valuable. It is widely used due to its accessibility and safety for children. This non-invasive technique measures neural connections by tracking fluctuations in the Blood Oxygen Level Dependent (BOLD) signal across different brain regions [23]. The resulting data are four-dimensional images that reveal both structural and functional aspects of the brain.

fMRI can be conducted in two main contexts : resting state and task-based paradigms. Although task-based fMRI may be better for exploring specific autism-related theories, such as the weak coherence theory. The weak coherence theory is a cognitive theory that suggests individuals with ASD tend to focus on details rather than the overall picture. According to this theory, autistic individuals exhibit "weak central coherence," meaning they have difficulty integrating global or contextual information to make sense of a situation or stimulus. Instead, they concentrate on specific, isolated details. This theory is often used to explain certain characteristics of autistic behavior, such as high proficiency in tasks requiring attention to detail, but also challenges in social situations where a broader understanding is necessary [24] , this method may be challenging for individuals with ASD, especially those who are young or have lower functioning levels. On the other hand, resting state fMRI does not require an experimental setup or active participation, making it suitable for capturing brain region interactions and potentially identifying diagnostic biomarkers for neurological conditions [25]. Thus, resting state fMRI is considered well-suited for studying ASD.

1.7.1 Magnetic resonance imaging

Magnetic Resonance Imaging is a medical imaging method that lets doctors see inside the body without using radiation. Instead, it relies on magnetic fields and radio waves. MRI uses a big machine, called an MRI scanner, which has a strong circular magnet. This magnet creates a powerful magnetic field that aligns the protons in hydrogen atoms in the body.

When the scanner sends out a radio wave pulse at the right frequency, the hydrogen protons absorb this energy and then send out a weak signal when they return to their normal state. The MRI system's radio wave coils pick up this weak signal, and a computer turns it into a detailed image of the body's tissues.

Early MRI machines were basic, but improvements in technology have made them much better over time. Today, MRI is a standard tool in radiology because it provides clear, high-resolution images with great contrast between different tissues, making it very

useful for diagnosing and monitoring many health conditions.

1.7.2 Functional Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging is a technique used to see how the brain functions. Unlike other methods, fMRI does not use X-rays, making it safe and suitable for all ages, including children. It is widely used in hospitals because it is relatively affordable and can produce very precise images quickly.

When a part of the brain is more active, it uses more energy. This energy is supplied by the blood, so when a brain region is active, there is more blood flowing to it. fMRI can detect these changes in blood flow and create images that show which parts of the brain are most active.
The most common technique in fMRI is BOLD contrast, which measures the differences between oxygenated and deoxygenated hemoglobin (a protein in the blood). When a brain area is active, it uses more oxygen, changing the amount of oxygenated and deoxygenated hemoglobin in that area. fMRI detects these changes and produces images that show the active parts of the brain. This helps researchers and doctors understand how different brain regions work together and study the effects of various diseases or conditions.

1.7.2.1 Functional Magnetic Resonance Imaging in Resting State

Resting state fMRI is a type of functional MRI used to study brain activity when a person is not performing any specific task. Instead of focusing on how the brain responds to external stimuli or tasks, resting state fMRI looks at the brain's activity while at rest, often with the eyes closed or focused on a neutral point.

This method primarily examines functional connectivity, revealing how different brain regions interact and communicate when not engaged in tasks. The technique utilizes the Blood Oxygen Level Dependent (BOLD) signal. When certain brain areas become more active, they require more oxygen, and the BOLD signal reflects these changes. By analyzing the correlations in the BOLD signal across various brain regions, researchers can identify networks of brain areas with synchronized activity.

Resting state fMRI is advantageous because it is non-invasive and does not require the subject to perform specific tasks, making it suitable for a wide range of populations, including those who may struggle with task-based fMRI. It provides baseline information about the brain's default activity patterns, which is useful for understanding normal brain function and various disorders.

Applications of resting state fMRI include studying brain disorders like ADHD, schizophrenia,Bipolar disorder, autism and others by identifying changes in connectivity associated with these conditions [26]. It is also used to investigate how brain connectivity evolves throughout development and aging, offering insights into both typical and atypical brain maturation.

1.7.2.2 Default Mode Network and autism

The Default Mode Network (DMN) is a group of brain regions that become active when we are at rest, meaning when we're not focused on a specific task. This network, like the brain's autopilot, switches off when we start doing something.

Unlike what was previously thought, the brain doesn't stop working when it's resting. In fact, the DMN is very active and plays a key role in personal memory, understanding others, planning for the future, and even creativity.

The main brain areas involved in the DMN are the medial prefrontal cortex (which helps with decision-making and emotions), the posterior cingulate cortex (which is involved in attention and self-awareness), the precuneus (linked to self-reflection), and the inferior parietal cortex (which integrates sensory information to build our perception of the world).

Studies have shown that in people with autism, the DMN works differently [27]. For example, there is less connection between certain parts of the brain, which is related to difficulties in social interactions. At the same time, some connections are stronger than they should be, which seems to be linked to the repetitive and restricted behaviors often seen in autism. This overconnection could either cause these behaviors or be a reaction to them.

In short, the DMN is essential for many mental functions, and studying it could help us better understand and treat certain disorders, like autism.

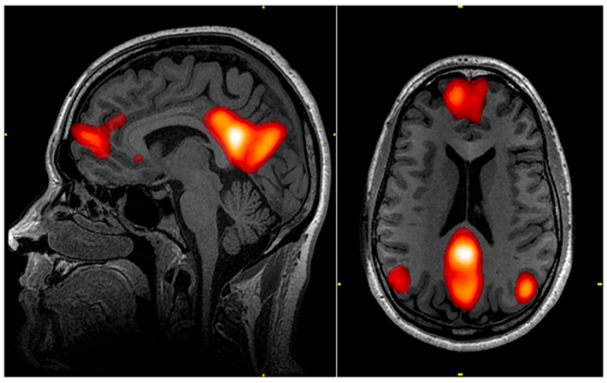

FIGURE 1.7.1 – Default mode network regions. [28]

1.7.3 Overview of 3D and 4D MRI Data.

MRI data is 3D for each subject and captures a static volumetric image of the brain. This data consists of three dimensions : height (H), width (W), and depth (D), representing the vertical dimension, horizontal dimension, and the number of slices in the volume, respectively. These provide a detailed "picture" in three dimensions, useful

for observing the brain's structure.

In contrast, fMRI data is 4D for each person, adding a temporal dimension to the 3D data. This allows the capture of a series of 3D brain volumes over time, with dimensions of height (H), width (W), depth (D), and time (T). Thanks to this fourth dimension, it's possible to extract time series that represent brain activity in different regions of the brain, called Regions of Interest (ROIs). These time series show how each region activates or deactivates over time, providing valuable information about the dynamic functions of the brain

1.8 Conclusion

This chapter has established a solid foundation for understanding Autism Spectrum Disorder by exploring its definitions, risk factors, and the importance of early diagnosis. We highlighted the challenges associated with current diagnostic methods, emphasizing the complexity of the condition and the variability of symptoms.

By delving into the aspects of brain development related to ASD, we examined how brain imaging techniques, particularly fMRI , can play a crucial role in detecting and understanding the brain connectivity abnormalities associated with autism. fMRI, being a non-invasive method, is particularly suitable for use with children, offering valuable insights into the altered brain networks in individuals with ASD.

In conclusion, this chapter lays the groundwork for further investigations by underscoring the importance of integrating different imaging techniques and analytical tools to enhance our understanding of ASD. This sets the stage for new perspectives in achieving earlier and more accurate diagnoses, which are essential for improving intervention and treatment strategies

Chapter 2

Autism Brain Imaging Data Exchange

Data availability can greatly contribute to the advancement of research. Therefore, with the aim of facilitating discoveries and comparisons in Autism research, the Autism Brain Imaging Data Exchange (ABIDE) initiative has compiled functional and structural brain imaging data from laboratories worldwide into a repository known as ABIDE.

2.1 ABIDE dataset

ABIDE encompasses two major collections : ABIDE I and ABIDE II. Each collection was assembled by aggregating datasets independently collected across more than 24 international brain imaging laboratories and made accessible to researchers worldwide, in line with open science principles, such as those central to the International Neuroimaging Data-sharing Initiative.

ABIDE I was the first initiative, starting in August 2012, involving 17 international sites. It includes **resting-state functional MRI (RS-fMRI), anatomical, and phenotypic data**, all shared on various platforms. The database contains 1112 datasets, with 539 from people with ASD and 573 from typical controls, ages ranging from 7 to 64. Since it began, ABIDE I has been used in many research projects worldwide, showing that brain imaging can reveal brain properties related to autism. The shared data has been very helpful for autism research. All datasets are anonymized, meaning no personal health information is included, following HIPAA and 1000 Functional Connectomes Project / INDI protocols.
ABIDE II was created to further support research on the brain connectome in autism. It started in June 2016, providing new data to researchers. ABIDE II adds over 1000 new datasets with more detailed information about core ASD and related symptoms. It also includes longitudinal data from 38 people collected at two different times, 1 to 4 years apart. ABIDE II involves 19 sites, including 10 founding institutions and 7 new ones. It has 1114 datasets from 521 people with ASD and 593 controls, ages ranging from 5 to 64. Like ABIDE I, all datasets are anonymized according to HIPAA guidelines and 1000 Functional Connectomes Project / INDI protocols.
ABIDE dataset is publicly available at [29] .

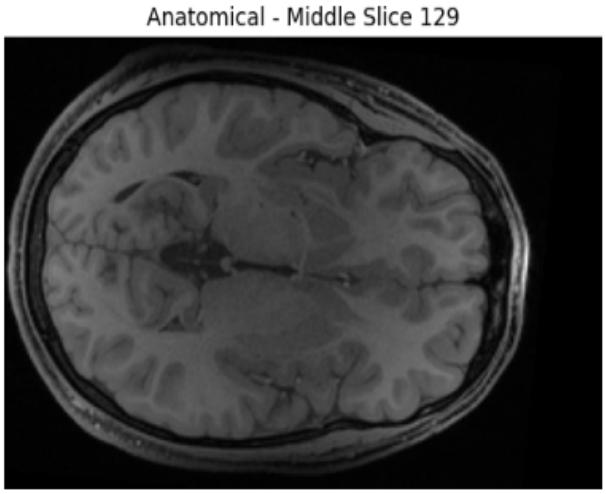

FIGURE 2.1.1 – Anatomical Slice of Subject 51464. [29]

2.2 ABIDE Preprocessed

Preprocessing in the medical field is crucial, especially with varied imaging data. Different brain imaging machines produce images with different characteristics [30]. Preprocessing reduces these differences by standardizing the data format and correcting abnormalities. There are many preprocessing techniques for functional images, from timing correction to standardizing formats from different acquisitions.
Preprocessing reduces noise in the data, which is essential because even a small amount of noise can lead to inaccurate results.
Combining various preprocessing methods creates different preprocessing pipelines. The choice of pipeline significantly impacts analysis results [31]. The same data analyzed with different preprocessing pipelines can yield different results, highlighting the importance of pipeline selection.
To address the challenge of choosing a preprocessing pipeline, preprocessed neuroimaging data from ABIDE I was shared. Five teams used their preferred tools to preprocess this data. Functional preprocessing was performed using the Connectome Computation System (CCS), the Configurable Pipeline for the Analysis of Connectomes (CPAC), the Data Processing Assistant for Resting-State fMRI (DPARSF), and the NeuroImaging Analysis Kit (NIAK) [32].
Due to debates about bandpass filtering and global signal regression, four preprocessing strategies were used with each pipeline. These strategies involved filtering with and without global signal correction and no filtering, also with and without global signal correction [32] .
The strategies are named :

- TrueTrue : Includes band-pass filtering and global signal regression.
- TrueFalse : Includes only band-pass filtering.
- FalseFalse : Excludes both band-pass filtering and global signal regression.
- FalseTrue : Includes only global signal regression.

Most research papers use CPAC, providing a comparison base for autism research. Additionally, various studies have employed software tools for preprocessing MRI modalities. These tools include brain extraction tools (BET) , the deformable medical image registration toolbox (DRAMMS) , FMRIB software libraries (FSL) , FreeSurfer , and statistical parametric mapping (SPM). MATLAB has also been used for preprocessing fMRI data.

2.3 Dataset description

The ABIDE dataset is a valuable resource for research on Autism Spectrum Disorder. It compiles data from various institutions worldwide, each contributing a specific number of participants with ASD as well as control subjects. These participants span a wide range of ages, allowing for the exploration of variability and common characteristics of ASD across different populations. The data are collected from renowned sites such as Caltech, Carnegie Mellon University, and New York University, among others. The table below provides a detailed overview of the distribution of participants by site, including the number of ASD participants, control subjects, age range, and the total number of participants per site.

Site	**ASD**	**Control**	**Age Range**	**Total**
Caltech	19	19	17.0-56.2	38
CMU	14	13	19-40	27
KKI	22	33	8.0-12.8	55
MaxMun	24	33	7-58	57
NYU	79	105	6.5-39.1	184
Olin	20	16	10-24	36
OHSU	13	15	8.0-15.2	28
SDSU	14	22	8.7-17.2	36
SBL	15	15	20-64	30
Stanford	20	20	7.5-12.9	40
Trinity	24	25	12.0-25.9	49
UCLA 1	49	33	8.4-17.9	82
UCLA 2	13	14	9.8-16.5	27
Leuven 1	14	15	18-32	29
Leuven 2	15	20	12.1-16.9	35
UM 1	55	55	8.2-19.2	110
UM 2	13	22	12.8-28.8	35
Pitt	30	27	9.3-35.2	57
USM	58	43	8.8-50.2	101
Yale	28	28	7.0-17.8	56

TABLE 2.1 – Demographic characteristics of the participants by site in the ABIDE dataset. [29]

The table below provides an overview of the different methods used for data acquisition at each participating site. It presents key information such as the type of acquisition used

(e.g., T2-weighted), voxel size (which determines the spatial resolution of the images), matrix size (which describes the dimensions of the obtained images), repetition time (TR, i.e., the interval between each successive image acquisition), as well as the total number of image volumes collected at each site.

Site	AcquisitionType	VoxelSize (mm)	MatrixSize	TR(msec)	volumes
Caltech	T2-weighted*	3.5 × 3.5 × 3.5	64 × 64	2000	146
CMU a	T2-weighted*	3.0 × 3.0 × 3.0	64 × 64	2000	236
CMU b	T2-weighted*	3.0 × 3.0 × 3.0	64 × 64	2000	316
KKI	T2-weighted*	3.0 × 3.0 × 3.0	84 × 84	2500	152
MaxMun	T2-weighted*	3.0 × 3.0 × 4.0	64 × 64	3000	116
NYU	T2-weighted*	3.0 × 3.0 × 4.0	80 × 80	2000	176
Olin	T2-weighted*	3.4 × 3.4 × 4.0	64 × 64	1500	206
OHSU	T2-weighted*	3.8 × 3.8 × 3.8	64 × 64	2500	78
SDSU	T2-weighted*	3.4 × 3.4 × 3.4	64 × 64	2000	176
SBL	T2-weighted*	2.75 × 2.75 × 2.72	80 × 80	2200	196
Stanford	T2-weighted*	3.125 × 3.125 × 4.5	64 × 64	2000	236
Trinity	T2-weighted*	3.0 × 3.0 × 3.5	80 × 80	2000	146
UCLA 1	T2-weighted*	3.0 × 3.0 × 4.0	64 × 64	3000	116
UCLA 2	T2-weighted*	3.0 × 3.0 × 4.0	64 × 64	3000	116
Leuven 1	T2-weighted*	3.59 × 3.59 × 4.0	64 × 64	1667	246
Leuven 2	T2-weighted*	3.59 × 3.59 × 4.0	64 × 64	1667	246
UM 1	T2-weighted*	3.438 × 3.438 × 3.0	64 × 64	2000	296
UM 2	T2-weighted*	3.438 × 3.438 × 3.0	64 × 64	2000	296
Pitt	T2-weighted*	3.1 × 3.1 × 4.0	64 × 64	1500	196
USM	T2-weighted*	3.4 × 3.4 × 3.0	64 × 64	2000	236
Yale	T2-weighted*	3.4 × 3.4 × 4.0	64 × 64	2000	19

TABLE 2.2 – Overview of data acquisition methods at each site in the ABIDE dataset where TR=Time of Repetition, volumes = Number of slices. [33]

2.4 Quality Assessment

The phenotypic file contains information about data quality, categorized into two types : automated quality assessment and manual quality assessment [34]. Automated quality assessment involves evaluations performed by algorithms or computer tools that measure objective aspects of the data. These assessments are typically precise and reproducible because they are based on automated calculations. In contrast, manual quality assessment is carried out by human reviewers who examine the data and provide ratings or comments based on their expertise and subjective judgment. This human evaluation include observations on clarity, the presence of artifacts, or other aspects that automated tools might not detect. Therefore, the file combines both objective and subjective evaluations to provide a comprehensive view of data quality.

2.5 Brain parcellation

Due to the high dimensionality of the preprocessed data, we reduce dimensions by dividing the brain into regions with similar properties based on a brain atlas.

2.5.1 Automated Anatomical Labeling

The Automated Anatomical Labeling (AAL) atlas, developed by Tzourio-Mazoyer et al. (2002) [35], is a popular brain parcellation and labeling tool. It splits the brain into different regions of interest (ROIs) based on a single-subject MNI T1 volume. This atlas is spatially normalized and segmented into eight classes like gray matter and white matter. Initially, 90 ROIs were manually drawn and filled using a specific algorithm. Updates have increased the number of ROIs to 116.

For functional resolution (3x3x3 mm^3), the atlas uses nearest-neighbor interpolation. In our analysis, we divided the preprocessed time series data into these 116 ROIs by averaging the signal within each region, resulting in a 2D vector with time steps and 116 ROIs.

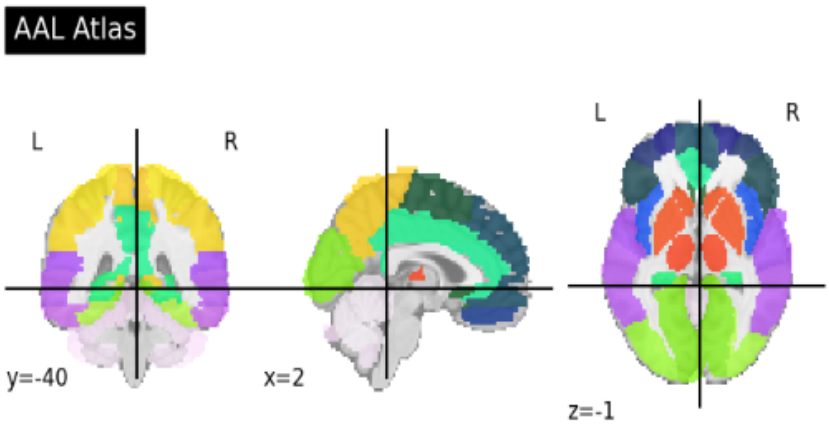

FIGURE 2.5.1 – Automated Anatomical Labelling atlas.

2.5.2 Harvard-Oxford

The Harvard-Oxford (HO) atlas provided by the FMRIB Software Library (FSL) was divided into cortical and subcortical probabilistic atlases, and then further separated into left and right hemispheres. Regions of interest (ROIs) corresponding to the left/right white matter, left/right gray matter, left/right cerebrospinal fluid, and brainstem were excluded from the subcortical atlas, resulting in a total of 111 ROIs.

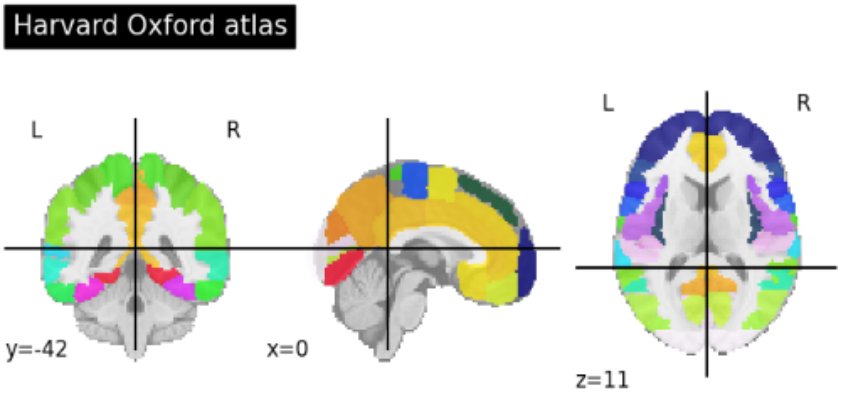

FIGURE 2.5.2 – Harvard-Oxford atlas.

2.5.3 Craddock 200 (CC200)

Functional parcellation was achieved through a two-stage spatially-constrained functional process on resting state data from 41 individuals. A grey matter mask was formed by averaging individual masks. Connectivity graphs were created for each individual, treating each voxel as a node and significant temporal correlations as edges. These graphs were partitioned into 200 regions using normalized cut spectral clustering. Association matrices were constructed from the clustering results. A group-level correspondence matrix was formed by averaging individual matrices and partitioned into 200 regions. The final group-level analysis was refined using nearest-neighbor interpolation, and labels for each ROI were assigned based on overlap with AAL, EZ, HO, and TT atlases using the `pyClusterROI` toolbox.

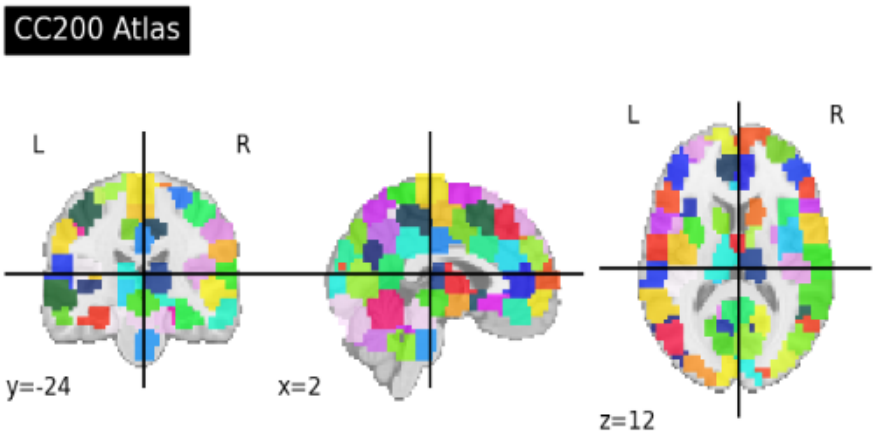

FIGURE 2.5.3 – Craddock 200 atlas.

2.5.4 Craddock 400 (CC400)

The procedure described for the CC200 atlas was repeated for 400 regions to create the CC400 atlas.

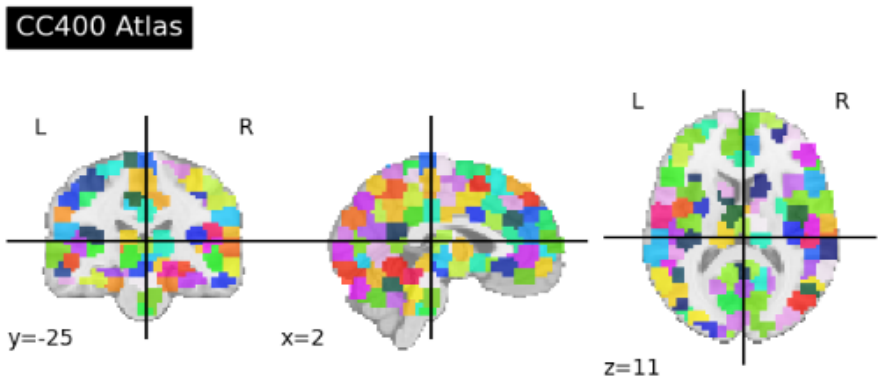

FIGURE 2.5.4 – Craddock 400 atlas.

2.5.5 Dosenbach 160

The Dosenbach 160 atlas consists of 160 virtual spheres, each with a radius of 4.5 mm, placed at specific locations in the brain. These locations, listed in Table S6 of the Dosenbach et al. (2010) paper [36], were determined through meta-analyses of task-related fMRI studies. These spheres represent regions of interest (ROIs) identified as being most active during these cognitive tasks.

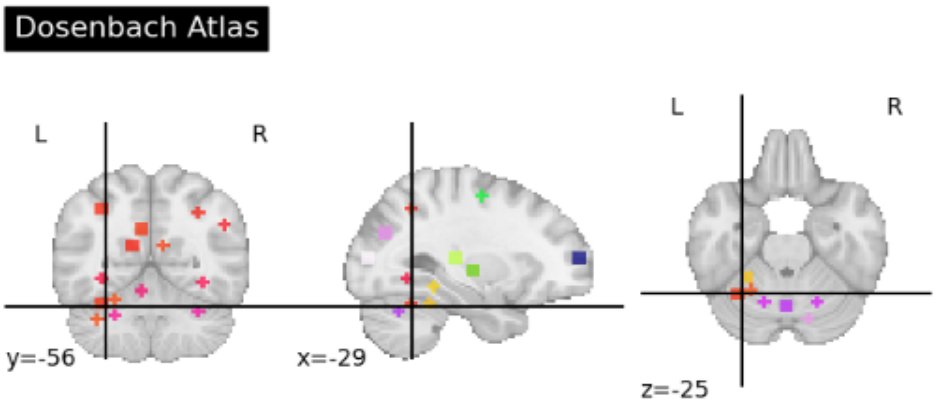

FIGURE 2.5.5 – Dosenbach atlas.

2.6 Time series

A time series is a sequence of data points collected and recorded over time, ordered chronologically, with each data point associated with a specific time or time interval. This type of data is used to analyze changes over time, identifying patterns, trends, and relationships (Shumway and Stoffer, 2017 [37]). Time series data can be recorded at regular intervals (e.g., hourly, daily) or irregular intervals, and can be univariate or multivariate, depending on the number of variables observed simultaneously.
Time series data originates from various fields, including economics, finance, meteorology, healthcare, and engineering [38] . It can pertain to any variable that changes over time. Time series analysis considers the internal structure of the data, such as autocorrelation, trends, or seasonal variations, to understand the factors influencing variables over time.
In the human body, every action and reaction involves fluctuations in brain activity. Activities such as looking, hearing, talking, and feeling induce different patterns of brain activity, captured in functional brain images. The brain, segmented into specialized regions, shows different activity patterns even at rest, revealing networks of areas that often act together.
Brain functional imaging time series capture neural oscillations, reflecting fluctuations in neuronal excitability in regions of interest. These oscillations can vary in speed and duration. Time series allow for capturing interactions of brain regions in a one-dimensional signal.
Techniques applied in brain research for time series analysis include spectral analysis (Fourier analysis, wavelet analysis), time-frequency analysis (spectrograms, wavelet transforms), connectivity analysis (coherence, phase synchronization), and advanced machine learning algorithms to extract temporal patterns or predict brain states . By utilizing these techniques, scientists gain insights into brain function, connectivity, cognitive processes, and neurological disorders, aiding in understanding brain dynamics, identifying biomarkers, and developing diagnostic and therapeutic tools for neurological and psychiatric conditions.

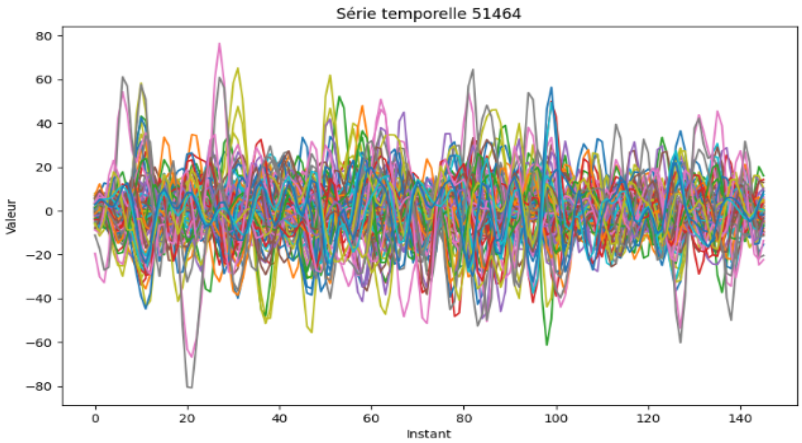

FIGURE 2.6.1 – Time series of subject 51464 extracted with the HO atlas.

2.7 Connectivity matrices

Connectivity matrices are widely used to study normal and pathological brain connectivity in various neurological and psychiatric states. For example, in the context of autism spectrum disorders, researchers use connectivity matrices to identify distinctive patterns of functional or structural connectivity that may be associated with ASD compared to neurotypical individuals. They are mathematical representations that allow quantifying and studying the functional or structural relationships between brain regions. They are essential for understanding how different brain circuits contribute to normal or disrupted brain function in various neurological and psychiatric conditions.

2.7.1 Functional connectivity

In functional neuroimaging, such as with fMRI, functional connectivity refers to the measurement of the correlation or temporal synchronization between neuronal activity in different brain regions. To calculate functional connectivity, fluctuations in brain activity over time are typically examined, and the degrees of similarity or coherence between signals from different regions are quantified. This measurement is often represented as a matrix, where each element of the matrix represents a measure of connectivity (e.g., Pearson correlation) between two specific regions.

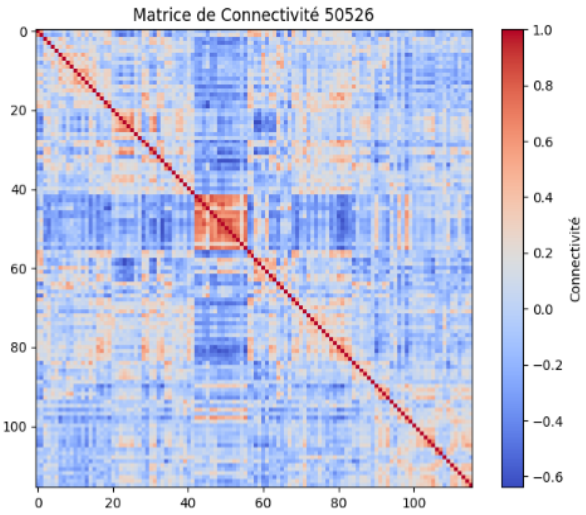

FIGURE 2.7.1 – connectivity matrix of subject 50526.

2.7.1.1 Correlation matrices

Correlation analysis is a widely-used method in the examination of resting-state functional magnetic resonance imaging (rs-fMRI). It determines the correlation coefficient between two regions of interest to indicate the degree of connectivity between them. A higher coefficient signifies stronger connectivity. This technique is employed to identify functional connections and resting-state networks within the human brain [39, 40, 41].
Consider X and Y as two time series representing two ROIs, with X = x1, x2, · · · , xt

and Y = y1, y2, $\cdots$, yt.
The Pearson correlation between these two time series is calculated as follows :

$$r = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{X})^2 \sum(y_i - \bar{Y})^2}}$$

Where :

$\bar{X}$ – the mean (average) of the X-variable
$\bar{Y}$ – the mean (average) of the Y-variable

2.7.1.2 Partial correlation

Partial correlation estimates the direct connection between two regions of interest (ROI X and ROI Y) by removing the influence of a third region (ROI Z), helping to uncover a causal relationship between the first two. The partial correlation between two ROIs, X and Y, after accounting for the influence of ROI Z, is calculated as follows :

$$r = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

Where :

r_{AB} : The simple correlation between ROIs A and B without considering other ROIs.

2.7.1.3 Covariance matrices

Covariance is a measure that indicates how two variables move together. If the covariance is positive, it means that when one of the variables increases, the other tends to increase as well. If it's negative, it means that when one increases, the other tends to decrease.

In short, covariance tells you if two things are related and how they move together, either in the same direction or in opposite directions.
The covariance between X and Y is expressed by :

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Where :
X_i – the values of the X-variable
Y_i – the values of the Y-variable
$\bar{X}$ – the mean (average) of the X-variable
$\bar{Y}$ – the mean (average) of the Y-variable
n – the number of data points

2.7.1.4 Tangent space matrices

Tangent space connectivity matrices offer a more sophisticated approach by using mathematical concepts to analyze how directions and changes in the brain connect at different points on a kind of geometric 'surface.'
These matrices are particularly useful in studies such as autism detection because they allow for the combination of information from multiple brain connections into a single view. This provides a more precise and detailed understanding of relationships within the brain.

To compute the tangent space matrix, we start by determining the covariance matrix Σ_i for each subject. Next, we calculate the geometric mean of these covariance matrices to obtain the reference matrix Σ_{mean} :

$$\Sigma_{\text{mean}} = \text{gmean}\left(\{\Sigma_1, \Sigma_2, \ldots, \Sigma_N\}\right)$$

Whiten each covariance matrix using the inverse square root of Σ_{mean}, which gives :

$$W_i = \Sigma_{\text{mean}}^{-1/2} \cdot \Sigma_i \cdot \Sigma_{\text{mean}}^{-1/2}$$

Finally, we apply the logarithm to each whitened matrix to obtain the tangent space representation :

$$L_i = \log(W_i)$$

The logarithm is applied element-wise to get the tangent matrix [42].

2.7.2 Structural connectivity

In structural neuroimaging, particularly with techniques like anatomical MRI (aMRI), structural connectivity focuses on understanding the physical, anatomical connections between different regions of the brain. Unlike functional connectivity, which examines the correlations of brain activity over time, structural connectivity is concerned with the actual physical pathways that link various brain regions. These connections are primarily made up of white matter fibers, which are bundles of axons that facilitate communication between different parts of the brain [43].

One of the most commonly used methods to assess structural connectivity is Diffusion Tensor Imaging (DTI). DTI is a type of MRI technique that measures the diffusion of water molecules in the brain's white matter. Because water molecules tend to diffuse more rapidly along the direction of axon fibers, DTI can be used to infer the orientation, directionality, and strength of these white matter pathways. The data collected from DTI allows researchers to map out the brain's structural network, often represented in the form of structural connectivity matrices.

These matrices are mathematical representations where each element corresponds to the strength or density of connections between a specific pair of brain regions. For example, if a particular matrix entry has a high value, it indicates a strong or dense connection between two regions, suggesting that these areas are anatomically well-linked. Conversely,

a lower value would suggest a weaker or less dense connection. Structural connectivity matrices are crucial in studying how brain regions are physically interconnected and can provide insights into how these connections relate to various cognitive functions or neurological conditions.

2.8 Conclusion

In this chapter, we thoroughly explored the data and methodologies essential for investigating Autism Spectrum Disorder through neuroimaging. Beginning with the ABIDE dataset and its preprocessed version, we provided a detailed description of the dataset, emphasizing the importance of quality assessment to ensure the reliability of the data used in our analyses.

We then examined the different brain parcellation techniques, including the AAL, HO , CC200, CC400, and Dosenbach 160 atlases. Each of these parcellation methods offers unique advantages for segmenting the brain into regions of interest, providing a foundation for the extraction of meaningful time series data.

The time series derived from these parcellations were used to construct connectivity matrices, which are pivotal in understanding the functional and structural connectivity patterns in the brain. We discussed various types of connectivity matrices, including correlation, partial correlation, covariance, and tangent space matrices for functional connectivity, as well as structural connectivity matrices. Each type of matrix presents a different perspective on the brain's connectivity, contributing to a comprehensive analysis of the neural networks involved in ASD.

This chapter thus provides a critical overview of the datasets, tools, and techniques that form the backbone of our study.

Chapter 3

MRI Data Representation and Techniques For Autism Diagnosis

3.1 MRI data representation

Raw MRI images are often too large and complex to be used directly in diagnosing autism through artificial intelligence. While a radiologist can examine these images to identify signs of diseases, the massive amount of data makes this task difficult to perform manually and quickly. Additionally, raw data lacks the specific details necessary for accurate diagnosis.
This is why researchers prefer to extract key information from these raw images instead of using them directly, as this helps AI models diagnose autism more effectively.
However, machine and deep learning models often require specific data formats, such as 2D images 1D vectors or particular shapes, making raw 3D/4D MRI data unsuitable.
To address this issue, researchers have developed alternative methods to represent MRI data, such as transforming it into 2D images, 3D brain volumes, or connectivity matrices. These methods help better prepare the data for AI models, thereby improving the diagnosis of autism.

3.1.1 Connectivity matrices

Connectivity matrices obtained from MRI scans are frequently used to identify biomarkers for autism spectrum disorder. These matrices provide insights into brain structure and activity by quantifying interactions between different brain regions. Researchers generate these matrices by segmenting the brain into regions of interest using a brain atlas, which reduces data dimensionality. Statistical methods, such as correlation coefficients, are then applied to evaluate co-activation levels between ROIs, resulting in weighted connectivity matrices that represent the strength of connections between these regions. These matrices are particularly useful for classification models requiring two-dimensional input data. Functional connectivity (FC) matrices are derived from fMRI scans, while structural connectivity (SC) matrices come from sMRI data. Various statistical methods, including Pearson's correlation, are used to create both FC and SC matrices.
Several studies have applied connectivity matrices in different ways for ASD classification. For example, [44] transformed a 200x200 connectivity matrix from the CC200 atlas

into a vector for use with an autoencoder, which extracted features for a deep learning model. [45] used a Pearson correlation matrix, transformed it into a vector, and selected features based on p-values before classifying autism subtypes with a machine learning model. [46] applied the CC400 atlas to extract time series data, created a correlation matrix, and used it in a CNN model for classifying autistic individuals. Similarly, [47] employed vectors from a correlation matrix in a fully connected neural network (FCNN) for classification.
Recent advances in neuroimaging data fusion have allowed for the integration of multiple datasets from the same subject, improving insights into brain function and activation changes related to neurological disorders. Combining data from different brain atlases helps extract more informative features for ASD diagnosis. Studies have explored combining functional connectivity matrices from various atlases using deep learning models and evaluating both spatial and functional connectivity information.
Structural connectivity can be estimated using the morphological covariance network approach, which captures interregional structural connectivity and individual-level variations. Some research integrates both structural and functional MRI information by combining FC and SC matrices. Approaches include concatenating feature vectors or training separate classifiers for each type of connectivity.
The use of connectivity matrices in models for detecting autism and other neurodevelopmental disorders is extensive and continually expanding.

3.1.2 Graph-based representations

Graph-based representations help researchers extract valuable information for tasks like brain network classification and visualization. The main goal is to create a meaningful representation of the brain's complex connections by reducing them to a lower-dimensional space, which improves accuracy and performance in diagnosing diseases such as autism (ASD). This involves constructing a graph where nodes represent specific brain regions and links represent the connectivity between these regions, often derived from FC analysis of fMRI data. There are different approaches to modeling these networks : one represents brain regions as nodes with functional connectivity as links, while another represents subjects as nodes with links based on similarities in their fMRI and phenotypic features.
Author [48] has designed a model using Graph Neural Networks (GNN) to learn effective representations for graph classification in an end-to-end manner. Specifically, it involves a method called "hi-GCN" (hierarchical Graph Convolutional Network), a variant of the GCN model. This model aims to encode the features of a graph while considering both the network's structure (the connections between nodes) and the relationships between subjects. Thus, hi-GCN enhances graph representation, making classification easier by integrating both structural information and the relationships between entities within the graph.
Author [49] uses the GAT-LI method, which combines a graph attention network (GAT2) and an interpretation technique (GNNExplainer) to classify functional brain networks between autistic individuals (ASD) and healthy controls (HC).

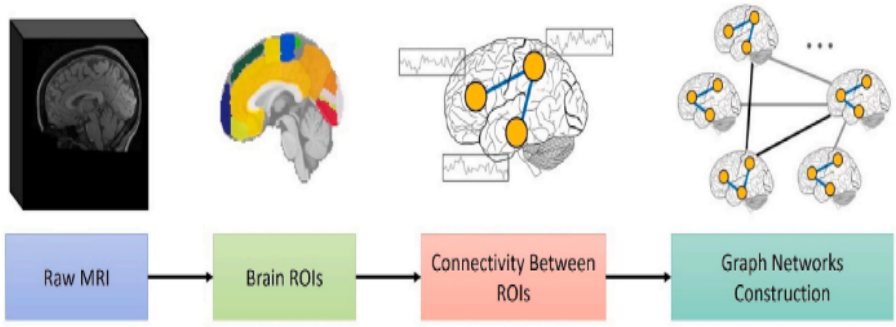

FIGURE 3.1.1 – Process for Constructing Graph-Based Representations from MRI. [50]

3.1.3 2D-slice images

Researchers use 2D images, known as "slices," extracted from 3D or 4D MRI data to analyze the brain. These 2D images provide a snapshot of brain activity or structure and are particularly useful for AI models that require 2D input. For instance, the author [51] used 3D fMRI data and saved it as 2D images to use with DarkASDNet for classifying autism brain images between ASD and TC.
In fMRI studies, 2D slices are taken from different anatomical planes (such as sagittal, coronal, or axial) at specific moments to visualize brain activity. These images are used to enhance understanding and aid in the diagnosis of autism. Similarly, the author [52] used 2D neuroimages and transfer learning with InceptionResNetV2 for autism spectrum disorder classification.
Similarly, in structural MRI studies, 2D slices are used to provide detailed views of the brain's anatomy. These images are then fed into deep learning models to assist in diagnosing autism, offering visual data on brain structure and activity.

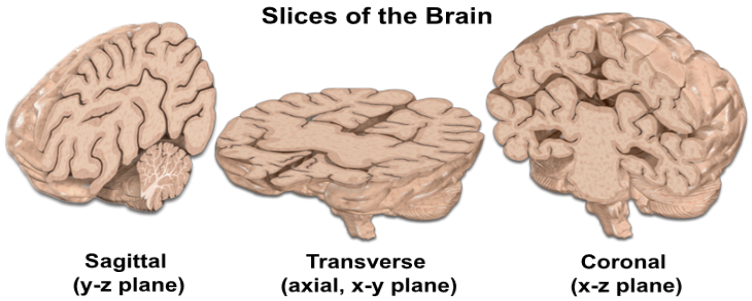

FIGURE 3.1.2 – Slices of the brain : sagittal, coronal, and axial. [53]

3.1.4 3D-volume

A 3D brain volume represents the entire brain in three dimensions from MRI scans. Unlike 2D slices, which provide only flat images, using the full 3D MRI data allows for a detailed and comprehensive view of the brain's structure.
This complete representation can be especially useful for deep learning (DL) models

or pretrained models in diagnosing autism spectrum disorder. By preserving all spatial details, these models can better identify complex patterns across the brain, potentially leading to more accurate ASD diagnoses. Research supports this approach : for instance, study [54] achieved 70% accuracy in ASD diagnosis using a 3D convolutional neural network (CNN) to analyze MRI scans. Another study [55] employed a pretrained 3D-ResNet model to extract key features from MRI data for diagnosing ASD.

3.1.5 Time series analysis

Studies on diagnosing ASD have focused a lot on looking at the temporal aspects of BOLD data from fMRI scans. The goal is to capture how s1ignals change over time and understand how brain anatomy affects these changes. fMRI signals are initially represented in two ways :

- Temporal-level : Signals are collected from different regions of interest at each time point.
- Spatial-level : Signals are gathered from each ROI at various time points.

To understand the overall temporal characteristics of brain networks, the time series data from each region within the network are either averaged or analyzed using different statistical measures. Author [56] uses time series extracted from a 4D fMRI scan with a brain atlas. The representative time series were obtained by averaging the responses of voxels within each ROI and then used as input for a transformer-based model to diagnose autism. Additionally, studies [57] and [58] used brain atlases to extract ROIs and averaged the time series data from each ROI to use as input for transformer-based models.

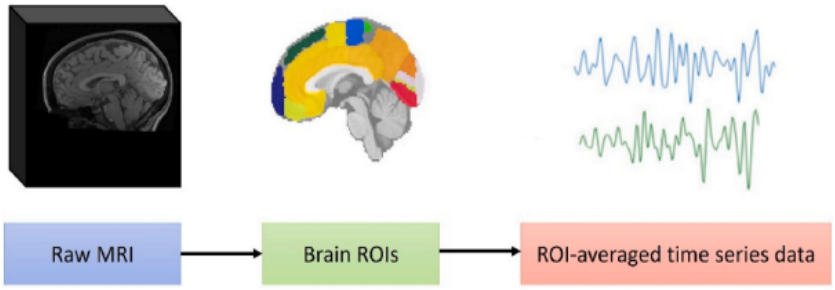

FIGURE 3.1.3 – Process of Time Series Data Extraction from MRI. [50]

3.1.6 Region-based representations

Finding brain imaging markers for autism is crucial for understanding the disorder. However, directly using raw MRI data from the entire brain makes it challenging to extract useful features for ASD diagnosis. A more effective strategy involves focusing on specific brain regions, known as regions of interest. This often means segmenting or outlining particular brain regions from the 3D brain volume to obtain more relevant information. For instance, in one study [54], raw MRI data was processed using a 3D convolutional neural network to identify key features. These features were then refined by applying masks from a brain atlas to specific ROIs, which were subsequently used in a Genetic

Algorithm for diagnosing ASD, resulting in a 73% accuracy rate. In another study [59], the 3D cerebral cortex was transformed into 2D images through a geometric mapping technique that preserved the spatial relationships of brain regions. These 2D images were then utilized for ASD diagnosis. Additionally, the author [60] used four brain areas in the diagnosis of ASD subjects, based on the CC400 functional parcellation atlas of the brain and a LeNet-5-based deep neural network for ASD detection.

3.2 Autism Spectrum Disorder Diagnosis Techniques

3.2.1 Machine Learning Algorithms

Traditional research has explored various machine learning (ML) models for diagnosing Autism Spectrum Disorder. Typically, ASD diagnosis is approached as a binary classification problem, where the dataset includes two labels : ASD and typically developing (TD/TC) subjects. However, ASD can also be framed as a multi-class classification problem, with labels corresponding to different ASD subtypes, such as autistic disorder, Asperger's disorder, pervasive developmental disorder, and TC.
An essential step in applying ML models is feature selection and extraction. In ASD diagnosis, researchers extract relevant data from raw MRI scans to use as input for ML algorithms. This often leads to high-dimensional data, which can impact the accuracy of ASD diagnosis. To address this, feature reduction methods are employed to simplify the data while retaining key information [61]. These methods enhance model performance, prevent overfitting, and reduce computation time in ASD diagnosis. Techniques include t-test [62, 63, 64, 65, 66], recursive feature elimination (RFE) [67, 68, 69, 70], analysis of variance (ANOVA) [71], principal component analysis (PCA) [72], F-score [73], Fisher score [69], and autoencoder (AE) [74, 75, 76, 77, 78].

The Support Vector Machine (SVM) algorithm has been extensively used in ASD diagnosis with promising results [79, 80, 81, 68, 62, 82, 83, 71, 64, 84, 72, 85].
For instance, one study achieved a 92.9% accuracy using ASD-related functional connectivity data from fMRI [86]. Another study used a cubic kernel function SVM for classifying ASD with non-oscillatory FC data, reaching an accuracy of 88.9% [80]. A multi-view high order-FC network from fMRI reached 86% accuracy [62]. A limitation of these studies was their small sample sizes : 98 subjects [86], 144 subjects [80], and 92 subjects [62].

Random Forest (RF), another ML algorithm, has been used for ASD classification with MRI data [79, 76, 78, 87, 88]. For example, combining FC and volumetric features from fMRI and sMRI with an RF classifier achieved an accuracy of 80% [87]. Evaluating FC from fMRI with a conditional RF method for feature reduction resulted in a 73% accuracy [88]. K-Nearest Neighbors (KNN) is also used in ASD classification. Combining features from sMRI and fMRI, KNN achieved accuracies of 75% and 79% on larger datasets [89]. With 11 neighbors and cosine metric, KNN reached 78% [80]and 58% [79] accuracy on smaller datasets.

Other ML models used in ASD diagnosis include Decision Tree (DT) [90], Logistic Regression (LR) [71], Ridge Classifier [82], Sparse Representation Classifier (SRC) [91], Ensemble Learning (EL) [80],[63], and Linear Discriminant Analysis (LDA) [79]. Ensemble classifiers improve precision by combining multiple models [92]. One study used a self-weighted adaptive structure learning (SASL) method for ASD diagnosis, leveraging FC from fMRI as weights and achieving an accuracy of 89% [63]. The table below summarizes the ML models, brain atlases, preprocessing techniques, and feature selection methods used in traditional ASD diagnosis research based on MRI data.

Model	**Data**	**Atlas**	**Prepro**	**feature-Sel**	**Val**	**MaxAcc**	**Nb.subj**
SVM	fMRI	HO	CPAC	t-tests, LASSO	10-CV	67.28%	403 ASD, 468 TC [64]
	fMRI	AAL	SPM8	t-test, LASSO	6-CV	86.2%	45 ASD, 47 TC [62]
	fMRI	CC200	CPAC	Extra-trees	10-CV	72.2%	506 ASD, 548 TC [85]
	sMRI	HO	FSL	Gaussian distribution	LOO-CV	AUC 0.77	151 ASD, 151 TC [84]
	fMRI	Glasser multimodal	CPAC	ANOVA	CV	60.89%	403 ASD, 468 TC [71]
	fMRI, Phenotypique	CC400	-	-	5-CV	71.1%	505 ASD, 530 TC [82]
	fMRI, sMRI	AAL	-	PCA	LOO-CV	78.89%	49 ASD, 41 TC [72]
	fMRI	Dosenbach	DPABI, SPM, MATLAB	Boruta method	LOO-CV	92.9%	48 ASD, 50 TC [86]
	fMRI, Genetic	AAL	MATLAB, SPIM12	t-test, SVM-RFE	LOO-CV	83.6%	47 ASD, 24 TC [66]
RF	fMRI	CC200	DPARSF	AE	10-CV	60.63%	432 ASD, 556 TC [76]
	fMRI	HO	FSL	FC	-	73.75%	306 ASD, 350 TC [88]
LR	fMRI, Phenotypique	CC200	-	-	5-CV	71.1%	505 ASD, 530 TC [82]
	fMRI	Glasser multimodal	CPAC	ANOVA	CV	60.89%	403 ASD, 468 TC [71]
DT	fMRI	AAL	CPAC	t-test, LASSO	LOO-CV	74.8%	201 ASD, 251 TC [90]
KNN	sMRI	-	-	Entropy, chi-square	5-CV	54.79%	592 ASD, 571 TC [79]

TABLE 3.1 – ASD diagnosis using MRI and ML models : Support Vector Machine (SVM),Random Forest (RF), and K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT).

3.2.2 Multi-layer perceptron

A multilayer perceptron, or MLP, is one of the most used ML algorithms. It is an artificial neural network composed of several layers. The first layer, known as the input

layer, receives the raw data, with each neuron representing a specific feature of the dataset. These data are then transmitted through one or more hidden layers, which apply nonlinear activation functions to model complex relationships within the data. Finally, the information reaches the output layer, which generates the model's final predictions. The MLP operates in two main phases : forward propagation, where the data is processed through the layers to produce an output, and backpropagation, where the error between the prediction and reality is used to adjust the weights of the neural connections. This learning process enables the MLP to solve complex classification and regression problems by learning nonlinear relationships within the data.

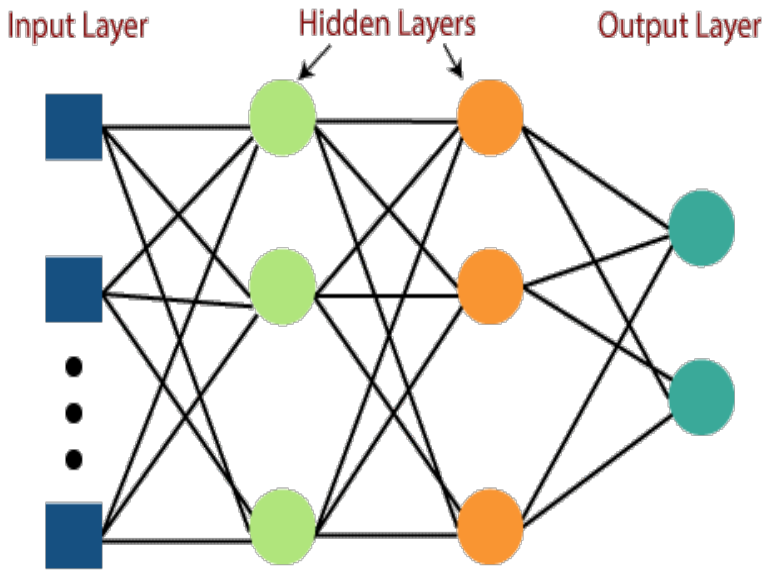

FIGURE 3.2.1 – MLP Architecture. [100]

The table below lists some papers that used Multi-Layer Perceptrons for classification. It gives a brief overview of how they selected features and subjects, and what their best results were.

Model	Dataset	No. of cases	Brain Atlas	Preprocessing	Feature selection	Max Accuracy	Ref.
MLP	fMRI	493 ASD, 530 TC	CC200, AAL	CPAC, FSL	–	84%	[94]
	fMRI	432 ASD, 556 TC	CC200	DPARSF	AE	71.35%	[76]
	fMRI	419 ASD, 530 TC	CC200, AAL, Dosenbach	CPAC	AE	74.52%	[75]
	fMRI	505 ASD, 530 TC	CC200	CPAC	Sparse AE	70.8%	[74]
	fMRI	403 ASD, 468 TC	AAL	CPAC	–	69.81%	[95]
	fMRI	539 ASD, 573 TC	AAL	CPAC, FMRIB, SPM, Artifact detection tools	Sparse AE	81.5%	[96]
	fMRI	402 ASD, 464 TC	BASC	CPAC	–	88%	[97]
	fMRI	505 ASD, 530 TC	CC200	–	F-score, AE	70.9%	[73]
	fMRI	505 ASD, 530 TC	CC400	CPAC	–	75.27%	[98]
	fMRI	79 ASD, 105 TC	–	CCS	–	94.70%	[99]
	sMRI, fMRI	368 ASD, 449 TC	AAL, CC200, Destrieux	CPAC	Fisher score, AE	85.06%	[69]
	fMRI	506 ASD, 532 TC	CC200, AAL90, DOS160	DPARSF	–	79.13%	[100]

TABLE 3.2 – References Using MLP for ASD Diagnosis.

3.2.3 Convolutional neural network

A Convolutional Neural Network is a Deep Learning algorithm and a type of artificial neural network that is particularly effective for processing grid-like data structures, such as images. CNNs are widely used in fields such as computer vision, image recognition, and video analysis.

CNNs operate through several types of layers that work together to extract and interpret features from images. Convolutional layers are at the heart of CNNs.

They use filters or kernels that slide across the input image to detect local features such as edges, textures, or specific patterns. Each filter generates a feature map that represents the presence of these patterns at different locations in the image. The main advantage of convolutional layers is that they preserve the spatial relationships between pixels in the image. In other words, they maintain the layout and structure of objects within the image. For example, recognizing a face requires understanding where the eyes, nose, and mouth are relative to each other. Convolutional layers preserve this spatial structure, allowing the network to detect patterns and objects more accurately.

After convolutional layers, CNNs often include pooling layers, such as max pooling. These layers reduce the dimensionality of the feature maps by taking the maximum or average value within a small window of each map. This helps to reduce the number of parameters and computations in the network while retaining the most important information. Pooling layers also make the network more robust to variations in the position of features within the image.

Once features are extracted by convolutional and pooling layers, CNNs typically move on to fully connected layers, similar to those in MLP. These layers interpret the extracted features to perform the final task, such as image classification or object detection.
Like multilayer perceptrons, CNNs use nonlinear activation functions, such as ReLU (Rectified Linear Unit), to introduce non-linearity into the model.

This allows the network to capture complex relationships in the data.
Learning in a CNN is done through backpropagation, where the error between the network's prediction and the actual result is used to adjust the weights of the connections across all layers of the network, including the convolutional layers. Through this process, the CNN gradually improves its performance, becoming capable of solving increasingly complex problems.

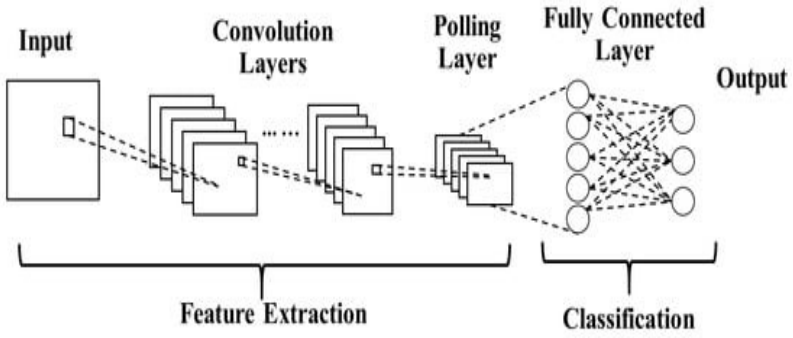

FIGURE 3.2.2 – Process of Convolutional Neural Networks (CNNs). [101]

The table below lists some papers that used CNN for classification.

Model	Dataset	No. of cases	Brain Atlas	Preprocessing	Feature selection	Max Accuracy	Ref.
CNN	fMRI	539 ASD, 573 TC	–	CPAC	AE	84.05%	[78]
	fMRI	539 ASD, 573 TC	CC200	CPAC	–	80%	[92]
	sMRI	592 ASD, 571 TC	–	–	Entropy, chi-square, symmetric uncertainty	66%	[79]
	fMRI	79 ASD, 105 TC	–	CCS	–	94.70%	[99]
	fMRI	491 ASD, 528 TC	AAL	DPARSF	–	82.12%	[60]

TABLE 3.3 – References Using CNN for ASD Diagnosis.

3.2.4 Autoencoder (AE)

An autoencoder is a Deep Learning algorithm and a type of neural network primarily used for unsupervised learning. Its main goal is to encode input data into a lower-dimensional space and then reconstruct this data from the compressed representation. Autoencoders are often used for tasks such as dimensionality reduction, data denoising, and data generation.

An autoencoder consists of two main parts : the encoder and the decoder. The encoder takes the input data and transforms it into a lower-dimensional representation called the encoding or bottleneck. This part of the network is composed of several layers, often densely connected layers or convolutional layers, which compress the information while attempting to preserve the essential features of the original data. The code or encoding is the compressed representation of the data obtained after passing through the encoder. It captures the most significant information in a more compact form, such as patterns or textures in the case of images.
The decoder, the second part, takes the encoding and attempts to reconstruct the original data from this compressed representation. It is generally symmetric to the encoder, using similar layers but in reverse order to generate a reconstruction of the data.
The primary goal of an autoencoder is to minimize the difference between the input data and the generated reconstruction. This difference is often measured by a loss function, such as mean squared error (MSE). By training the network to minimize this loss, the autoencoder learns to encode and decode data effectively.
Autoencoders can also be used to reduce the dimensionality of data while retaining their main features. This reduction is useful for preprocessing before applying other machine learning algorithms. Some autoencoders, known as denoising autoencoders, are designed to remove noise from data by learning to reconstruct the input data from noisy versions.
Autoencoders can also generate new data. Variational autoencoders (VAEs) allow for generating samples by sampling from the learned encoding distribution.
Regarding their role in classification, autoencoders can be used in several ways. They are often employed to preprocess data by reducing its dimensionality while preserving essential features. The encoding produced by the autoencoder can serve as input to a traditional classifier such as a support vector machine (SVM) or logistic regression, thereby improving the classifier's performance by simplifying the data and reducing noise.
Additionally, autoencoders can be extended to include a classification layer directly in the model. This approach involves adding fully connected layers on top of the decoder to predict classes from the encodings. By training the model with a combined loss function, which includes both reconstruction loss and classification loss, the autoencoder learns both to effectively reconstruct the data and to perform classification tasks.

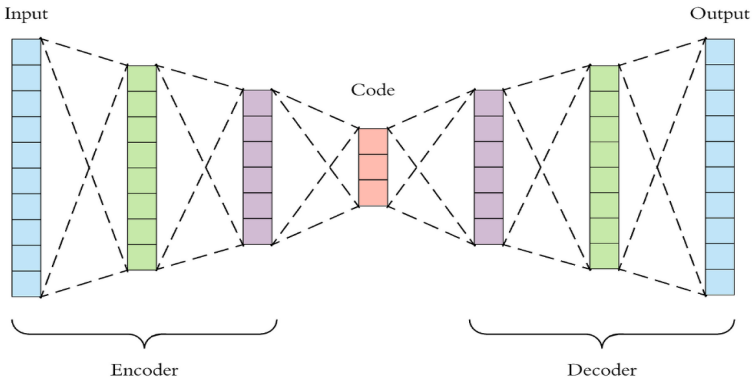

FIGURE 3.2.3 – Autoencoder Architecture. [102]

The table below lists some papers that used autoencoder for classification.

Model	Dataset	No. of cases	Brain Atlas	Preprocessing	Feature selection	Max Accuracy	Ref.
AE	fMRI	432 ASD, 556 TC	CC200	DPARSF	–	71.35%	[76]
	fMRI	322 ASD, 352 TC	CC200	CPAC	–	71.3%	[77]
	fMRI	419 ASD, 530 TC	CC200, AAL, Dosenbach	CPAC	–	74.52%	[75]
	fMRI	539 ASD, 573 TC	–	CPAC	–	84.05%	[78]
	fMRI	505 ASD, 530 TC	CC200	CPAC	Sparse AE	70.8%	[74]
	fMRI	539 ASD, 573 TC	AAL	CPAC, FMRIB, SPM, Artifact detection tools	Sparse AE	81.5%	[96]
	fMRI	505 ASD, 530 TC	CC200	–	F-score	70.9%	[73]
	sMRI, fMRI	368 ASD, 449 TC	AAL, CC200, Destrieux	CPAC	Fisher score	85.06%	[69]

TABLE 3.4 – References Using AE for ASD Diagnosis.

3.2.5 Graph convolutional networks

Graph Convolutional Networks (GCN) are a class of neural networks specifically designed to handle data structured as graphs, such as social networks or molecular structures. Inspired by CNN used for image processing, GCNs generalize the convolution operation to graphs. Their goal is to capture local graph features by aggregating information from the immediate neighbors of each node across multiple network layers. This aggregation allows each node to update its representation based on the graph structure and the characteristics of its neighbors. GCNs are used for tasks such as node classification, link prediction, and graph generation.
The table below lists some papers that used GCN for ASD diagnosis.

Model	Dataset	No. of cases	Brain Atlas	Preprocessing	Feature selection	Max Accuracy	Ref.
GCN	fMRI	403 ASD, 468 TC	HO	CPAC, MNI152, FSL	Deep feature selection	79.5%	[103]
	fMRI	403 ASD, 468 TC	HO	CPAC	RFE	73.71%	[67]

TABLE 3.5 – References Using Graph convolutional networks for ASD Diagnosis.

3.2.6 Graph attention network

Graph Attention Networks (GAT) are an advancement of GCNs that incorporate an attention mechanism. This mechanism allows the network to assign different importance to the neighbors of a node during the aggregation of information. Unlike GCNs, which treat all neighbors equally, GATs automatically determine the weights for each neighbor, enabling them to better capture complex relationships between nodes. This attention mechanism makes GATs particularly suited to contexts where relationships between nodes are heterogeneous or complex, often resulting in better performance than GCNs in such scenarios.

The table below lists some papers that used GAT for ASD classification.

Model	Dataset	No. of cases	Brain Atlas	Preprocessing	Feature selection	Max Accuracy	Ref.
GAT	fMRI	403 ASD, 468 TC	HO	CPAC	Two sample t-test, LASSO	72.40%	[64]
	fMRI	505 ASD, 530 TC	HO	CPAC	–	95%	[49]

TABLE 3.6 – References Using GAT for ASD Diagnosis.

3.3 Conclusion

In this chapter, we delved into the various techniques and models employed for autism diagnosis, particularly focusing on how MRI data can be represented and utilized within different classification models.

We began by exploring the different representations of MRI data, including connectivity matrices, graph-based representations, 2D-slice images, 3D-volume reconstructions, time series analysis, and region-based representations. Each of these methods offers a unique way to capture and interpret the intricate patterns in brain data, which are crucial for identifying the neural signatures associated with Autism Spectrum Disorder.
Following the discussion on data representation, we introduced a variety of models used for autism detection. We examined traditional machine learning algorithms and advanced neural network architectures, such as CNN, AE, GCN, and GAT.

To provide a comprehensive overview, we also created tables summarizing the performance of these models, highlighting their effectiveness in the state of the art. This comparison offers valuable insights into the strengths and limitations of each approach, helping guide the selection of the most appropriate techniques for specific diagnostic tasks.

Chapter 4

Machine learning and Deep Learning models for ASD detection

As previously mentioned, it is impossible for the human eye to detect whether a person is autistic or neurotypical based solely on their MRI, whether anatomical or functional. Even by plotting the time series or displaying the connectivity matrices, it is difficult to perceive a significant difference between autistic and neurotypical individuals, as these representations appear almost identical to the naked eye. This is why it is essential to use a computerized system capable of detecting subtle differences, often imperceptible to humans, and analyzing complex data. These systems, generally based on machine learning or artificial intelligence algorithms, allow for the identification of hidden patterns and correlations that the human eye cannot discern. Thanks to these tools, it becomes possible to classify individuals as autistic or neurotypical, based on the fine characteristics of the collected data, such as neural connections or variations in the brain's time series.

In this work, we attempted to create models using CPAC pipeline, applying one of the strategies TrueFalse or TrueTrue each time. We used several brain atlases, such as CC200, CC400, AAL, HO, BASC, POWER, and DOSENBACH, as well as different connectivity matrices, including correlation, tangent, and partial correlation.

Initially, We applied these approaches to machine learning classifiers using fMRI data from 1,102 subjects available for each preprocessing pipeline. However, the results did not exceed a mean accuracy of 71.14% using the CC400 atlas in a 5-fold stratified cross-validation, and 67% for a 30% test set . We then selected a subset of 866 subjects, then 871 subjects based on a Quality Assessment evaluation. Despite these efforts, the performance of the results remained low.

We also attempted to replicate deep learning architectures using both single-input models (based on a single atlas) and multi-input models (integrating multiple atlases) as described in papers [97, 98], and [100]. We performed fine-tuning of the hyperparameters, but we still encountered issues with low accuracy or overfitting of the model.

Subsequently, We turned to feature selection, relying on the concept of Deep Feature Selection as outlined in paper [103]. This method involves selecting features with the highest weights in a deep learning model. However, for this method to be reliable, the model must be able to classify the data well, which was not the case due to the persistent instability of the models used.

Finally, We explored various feature selection methods to enhance the performance of models. Initially, We experimented with Information Gain, a filter-based approach. Information Gain is a measure that assesses the importance of features by evaluating the reduction in uncertainty they provide. However, despite its utility in certain applications, the results obtained with this method did not exceed an accuracy of 69.35% ± 3.39% after cross-validation using the CC200 atlas. While this performance was respectable, it fell short of the requirements of our project.

To advance further, We decided to explore more sophisticated feature selection methods. First, we employed the LASSO (Least Absolute Shrinkage and Selection Operator) method, an embedded approach. This method integrates feature selection directly into the model training process by applying a regularization penalty that forces some regression coefficients to zero. This regularization helps to simplify the model by eliminating less relevant features while retaining the most important ones.

Subsequently, we applied the RFE (Recursive Feature Elimination) method, which is a wrapper-based approach. RFE is an iterative method that selects the most relevant features by gradually eliminating those that contribute the least to the model's performance. At each iteration, the least significant feature is removed, and the model is retrained on the remaining features until an optimal subset is identified.

The application of these two methods, LASSO and RFE, resulted in significantly improved performance levels. After cross-validation using the CC200 atlas, we achieved an impressive accuracy of 99.66% ± 0.52%. This substantial increase in performance demonstrated the effectiveness of these approaches in addressing the complex challenges of data classification.

Through this approach, we was finally able to achieve results that not only exceeded our expectations but also surpassed the state of the art, marking a significant advancement in data classification. This experience highlights the importance of advanced feature selection methods in improving model performance and ensuring robustness when dealing with complex datasets.

4.1 Machine learning

Machine learning is a branch of artificial intelligence (AI) that involves creating algorithms and statistical models that allow computers to learn from data and improve their performance on a specific task without being explicitly programmed. The main aim is for computers to recognize patterns, make predictions, or take actions based on

the data they have encountered.

Machine learning can be divided into three main types :

- Supervised Learning : This type involves training the algorithm with a labeled dataset, where both the input data and the correct output are provided. The goal is for the algorithm to learn the relationship between inputs and outputs so that it can make accurate predictions on new, unseen data. Common tasks include classification (where inputs are assigned to predefined categories) and regression (where continuous values are predicted).
- Unsupervised Learning : Here, the algorithm is trained on an unlabeled dataset, meaning the input data doesn't have corresponding correct outputs. The algorithm's goal is to uncover patterns, structures, or representations within the data on its own. Examples include clustering, where the algorithm groups similar data points, and dimensionality reduction, where the complexity of the data is reduced while keeping essential information intact.

- Reinforcement Learning : This type focuses on training agents to interact with an environment and learn from the feedback they receive in the form of rewards or penalties. The agent's objective is to maximize cumulative rewards over time, improving its decision-making and actions within the environment. Reinforcement learning is commonly applied in areas like gaming, robotics, and autonomous vehicles.

The typical machine learning process involves several steps :

- Data Collection : Gathering relevant and representative data for training and evaluating the model. The success of a machine learning model greatly depends on the quality and quantity of the data.
- Data Preprocessing : Preparing the data by cleaning, transforming, and formatting it to be suitable for the learning algorithm. This may involve handling missing values, normalizing features, or encoding categorical variables.
- Model Selection : Choosing the right algorithm or model architecture that best fits the problem and the nature of the data. Different algorithms are suitable for different types of tasks.
- Training : Feeding the model with training data so it can learn the underlying patterns and relationships. The learning process usually involves adjusting the model's parameters iteratively to reduce errors or enhance performance.
- Evaluation : Testing the trained model on a separate dataset (the test set) to evaluate its performance and generalization ability. Various metrics are used to measure how well the model performs on new, unseen data.
- Deployment : Once the model has been successfully trained and evaluated, it is deployed in a real-world setting to make predictions or take actions.

4.1.1 Autism Detection Using Machine Learning

In this work, we used the preprocessed version of the ABIDE dataset, which collects resting-state fMRI data and phenotypic information from subjects at 17 international sites, as mentioned earlier. The preprocessed dataset included 1102 scans, with 531 from individuals with ASD and 571 from control individuals. However, not all functional data met the Quality Assessment Protocol (QAP) metrics set by the PCP community [104]. Consequently, we tested 1102, 871, and 866 subjects.

4.1.1.1 Data preprocessing

After selecting the brain atlas used to extract our time series, for example, the AAL atlas, we calculated connectivity matrices that represent the correlation between different regions of interest defined by the atlas. The resulting connectivity matrices for each subject have dimensions of (116, 116) for the AAL atlas. Since the connectivity matrix is symmetrical, the values in the upper triangular part are a repetition of those in the lower triangular part. To reduce dimensionality, the upper triangular values, including the principal diagonal, were removed, and the lower triangular values were retained. The lower triangular part was then flattened to a 1D feature vector of size :

$S = \frac{N(N-1)}{2}$, Where N is the number of ROIs. For the AAL atlas, this resulted in a feature vector of size 6670 for each subject.

The diagram below summarizes the preprocessing steps to obtain the input vector.

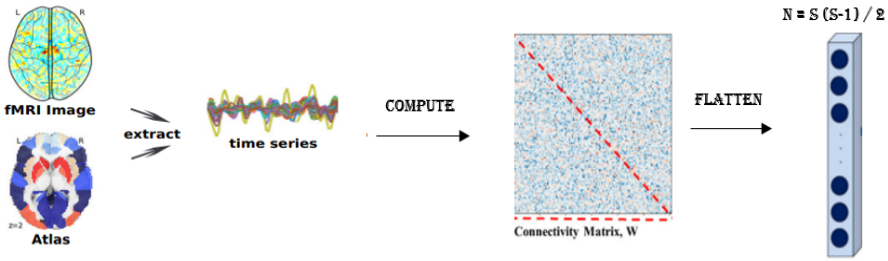

FIGURE 4.1.1 – Preprocessing pipeline to obtain the input vector.

4.1.1.2 Approach 1 : without feature selection

In this study, we tested several atlases for time series extraction, including HO , CC200, CC400, Dosenbach, AAL ,and two rarely atlases BASC and Power used in paper [97].

After extracting the time series, we explored three types of connectivity matrices - tangent space, partial correlation, and correlation - to represent functional relationships between brain regions. These matrices were used as the basis for training several machine learning classifiers.

We evaluated various classification algorithms, including Logistic Regression (LR) , Support Vector Classifier (SVC), Decision Tree, k-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), and Linear Discriminant Analysis (LDA). Each algorithm was tested using a 5 stratified cross-validation to ensure rigorous and balanced performance evaluation.

A one-dimensional vector derived from the preprocessed data served as input for these algorithms. Stratified 5-fold cross-validation involved dividing the dataset into 5 folds, using four for training and one for testing in each iteration. This process was repeated five times, rotating the test fold, to ensure each sample was used for both training and testing.

The final performance results for each model were calculated as the average score across the five folds. This information reveals the model's stability or instability based on data variation.

The table below summarizes the results, highlighting the models that delivered the best performance for each tested atlas and preprocessing pipeline.

Model	**Atlas**	**Strategy**	**Preprocessing**	**Connectivity**	**SubjNum**	**MeanAcc**
LR	BASC	TrueTrue	CPAC	Tangent	866	65.94%
LR	POWER	TrueTrue	CPAC	Tangent	866	66.75%
LR	CC400	TrueFalse	CPAC	Correlation	1102	71.14%
LR	CC400	TrueFalse	CPAC	Correlation	871	67.74%
SVC	HO	TrueTrue	CPAC	Correlation	871	67.63%
SVC	AAL	TrueTrue	CPAC	Correlation	871	66.37%
LR	CC200	TrueTrue	CPAC	Tangent	871	68.89%
LR	DOSENBACH	TrueFalse	CPAC	Tangent	871	64.41%

TABLE 4.1 – Results of each the models with different atlases, strategies, preprocessing, and connectivity methods.

To evaluate the performance of various machine learning models on the most commonly used atlases (HO, CC200, and AAL) using correlation and 871 subjects (see the table below for the description) , we conducted a stratified 5-fold cross-validation. The accuracy scores of different classifiers were measured and are summarized below with their corresponding standard deviations. The plot illustrates the performance comparison of these models, providing a visual representation of their effectiveness in classifying the data.

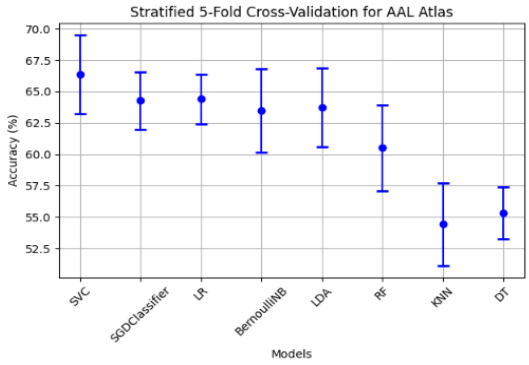

FIGURE 4.1.2 – Stratified 5-fold cross-validation for AAL atlas.

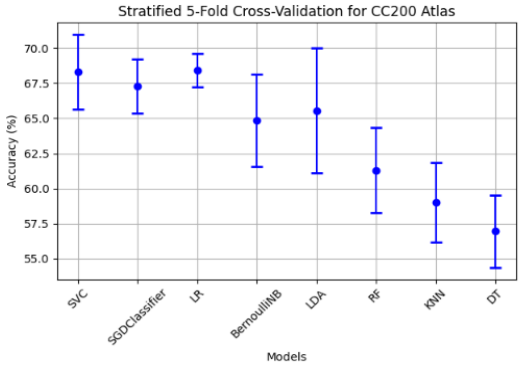

FIGURE 4.1.3 – Stratified 5-fold cross-validation for CC200 atlas.

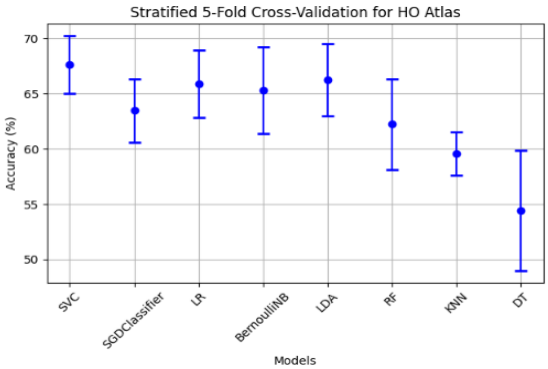

FIGURE 4.1.4 – Stratified 5-fold cross-validation for HO atlas.

Site	ASD Age (year)	ASD Sex (M/F)	TD Age (year)	TD Sex (M/F)
Caltech	24.0 ± 7.6	4/1	28.2 ± 12.2	6/4
CMU	26.0 ± 5.4	4/2	27.8 ± 4.4	3/2
KKI	10.7 ± 1.3	9/3	10.1 ± 1.2	15/6
Leuven_1	21.9 ± 4.1	14/0	23.0 ± 2.8	14/0
Leuven_2	13.9 ± 1.5	9/3	14.4 ± 1.5	12/4
Max_Mun	28.4 ± 13.2	16/3	25.2 ± 8.4	26/1
NYU	14.8 ± 7.1	64/10	15.8 ± 6.2	72/26
OHSU	11.4 ± 2.2	12/0	10.2 ± 1.0	13/0
Olin	17.1 ± 3.3	11/3	16.9 ± 3.6	12/2
Pitt	18.3 ± 7.0	21/3	18.7 ± 6.7	22/4
SBL	34.0 ± 6.6	12/0	33.6 ± 6.8	14/0
SDSU	15.3 ± 1.8	8/0	14.0 ± 1.9	13/6
Stanford	10.2 ± 1.6	9/3	9.8 ± 1.7	9/4
Trinity	17.0 ± 3.2	19/0	17.1 ± 3.8	25/0
UCLA_1	13.3 ± 2.6	31/6	13.4 ± 2.1	24/3
UCLA_2	12.8 ± 2.0	11/0	12.1 ± 1.2	8/2
UM_1	13.3 ± 2.5	26/8	14.1 ± 3.2	35/17
UM_2	14.9 ± 1.6	12/1	16.7 ± 4.0	20/1
USM	23.6 ± 8.4	43/0	20.9 ± 8.3	24/0
Yale	13.1 ± 3.0	14/8	13.6 ± 2.1	11/8
Total	**17.1 ± 8.0**	**349/54**	**16.8 ± 7.2**	**378/90**

TABLE 4.2 – Summary of Age and Sex Data Across Sites for 871 Subjects.

4.1.1.3 Discussion

As you can see, using this method, we achieved a maximum accuracy of 71.14% with the CC400 atlas. However, this accuracy was obtained using the complete dataset without excluding subjects whose scans had issues, which limits the reliability of this accuracy. On the other hand, by selecting subjects with high-quality scans, we achieved a maximum accuracy of 68.89% with the CC200 atlas. Although this result might initially seem encouraging, it remains insufficient for a model intended for medical diagnostic support. In a medical context, even a small error can have serious consequences.

The importance of accuracy in such a model cannot be understated. An accuracy rate of 68.89% means that there remains a significant margin of error, where nearly one-third of predictions could be incorrect. In a diagnostic support tool, this margin of error is unacceptable, as it could lead to incorrect diagnoses, thereby endangering the health and safety of patients.
To improve the model's accuracy, we decided to reduce the dimensionality of the input vector by selecting only the most relevant features. To do this, we used different feature selection techniques : Information Gain, RFE, and Lasso.
By testing these different selection techniques, our goal was not only to improve

the model's accuracy but also to enhance its robustness and ability to generalize to new data. By reducing the model's complexity while retaining the most informative features, we aimed to increase the chances of obtaining more reliable predictions, which is crucial for clinical use.

4.1.1.4 Approach 2 : with feature selection

Feature selection is a crucial process in developing classification models. It helps reduce the dimensionality of data by keeping only the most relevant features, which can improve performance, reduce computation time, and avoid overfitting.

Supervised Methods
Supervised methods use class label information to guide feature selection.

a) Filtering Approach
Filtering methods evaluate each feature independently of the others based on its relationship with the target variable.

Chi-Square Test : This method calculates a score based on the statistical dependence between each feature and the class. For example, in medical data, the Chi-square test might be used to identify which features (like blood test results) are most strongly associated with a specific disease.

Fisher Score : This method measures how well a feature separates different classes. It is often used in classification problems to select features that maximize the distance between classes. For example, to classify flowers by species, the Fisher score might help identify which measurements (petal length, petal width, etc.) are most discriminative.

Information Gain : Information Gain measures how much a feature helps in predicting the target variable by reducing uncertainty. It looks at how much knowing the value of a feature improves our ability to make accurate predictions. The higher the Information Gain, the more useful the feature is for making predictions. For example, in a decision tree, features with the highest Information Gain are chosen to split the data, leading to better decision-making and more accurate models.

b) Wrapping Approach

Wrapping methods evaluate features based on the performance of the classification model. They are more computationally expensive because they require building and evaluating multiple models.

Recursive Feature Elimination (RFE) : RFE is an iterative method that starts by training a model on all features. It then evaluates the importance of

each feature and removes the least important ones. This process is repeated until an optimal number of features is selected. For example, in an SVM model used to classify emails as spam or not spam, RFE might be used to select the most relevant words by gradually removing those that contribute least to classification.

c) Embedded Approach

Embedded methods perform feature selection during model training.
Regularization (Lasso) : Lasso (Least Absolute Shrinkage and Selection Operator) adds a penalty to the model's cost function, forcing some feature coefficients to become zero. This results in automatic elimination of the least important features. For example, in a logistic regression model used to classify customers based on their purchase likelihood, Lasso might automatically exclude variables that do not add predictive value.

Random Forest Feature Importance : In a random forest, feature importance is determined by how well a feature reduces impurity in nodes during splits. The most important features are those that contribute the most to creating effective decision rules. For example, to classify tree species based on environmental characteristics (like soil type or precipitation), a random forest might identify which features are most crucial for this task.

Unsupervised Methods

Unsupervised feature selection methods do not use class labels. They are often used to explore data structure or to reduce dimensionality before using the data in a classification model.

a) Principal Component Analysis (PCA)

PCA is a dimensionality reduction method that transforms the original data into a new set of uncorrelated variables (called principal components). These components are ordered so that the first one retains the most variance present in the data. For example, in image recognition problems, PCA can be used to reduce the number of pixels while keeping the essential variations needed to classify images into different categories.

b) Independent Component Analysis (ICA)

ICA aims to decompose data into statistically independent components, often used to separate mixed sources in the data. For example, in audio classification tasks, ICA might be used to separate different sound signals before classifying them into categories (music, speech, noise, etc).

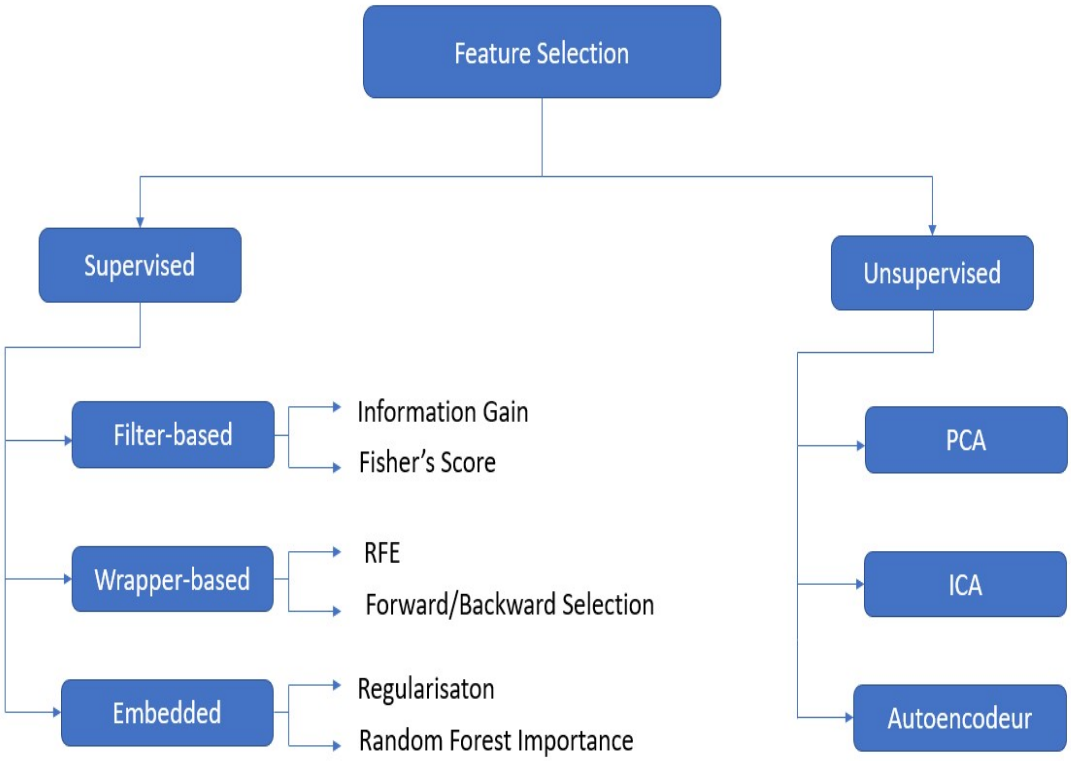

FIGURE 4.1.5 – Advanced Feature Selection Techniques.

As mentioned before, the most commonly used pipeline for autism detection is CPAC, which is also the one we used for this work, considering the TrueTrue strategy with the most commonly used atlases (HO, CC200, and AAL) using correlation and 871 subjects
In this work, we used the same vector obtained after preprocessing in the context of Approach 1. This vector then underwent feature selection to reduce its dimensionality and improve the performance of the classification models.

Feature selection based on Lasso method :
The figures below display the results obtained by using LASSO for feature selection and different cross-validation methods to assess model performance on the AAL, CC200, and HO atlases. Three types of cross-validation were applied : 5-fold Stratified Cross-Validation, 10-fold Stratified Cross-Validation, and 10-fold standard Cross-Validation. Additionally, a separate test set, representing 30% of the data, was used to evaluate the final model performance .
The accompanying table summarizes the best performances achieved for each atlas, expressed as the average accuracy, along with the corresponding standard deviation.

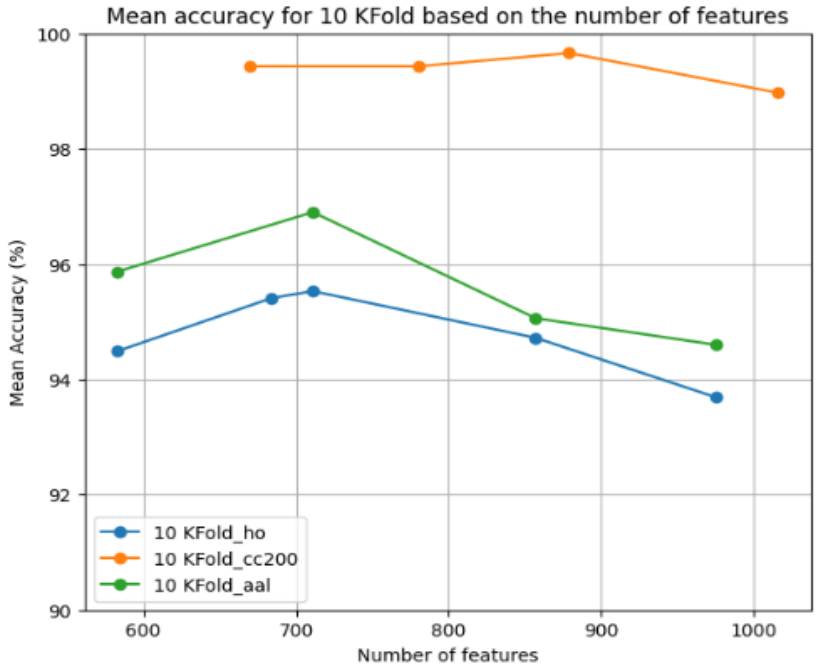

FIGURE 4.1.6 – Mean Accuracy for 10-Fold Cross-Validation Based on the Number of Features Selected by Lasso for AAL ,HO and CC200 atlases.

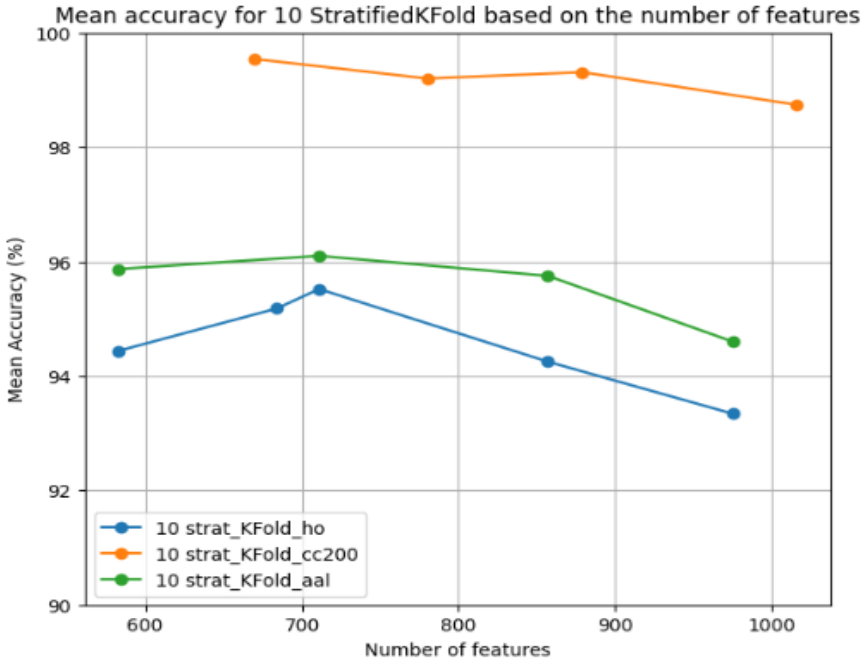

FIGURE 4.1.7 – Mean Accuracy for 10 StratifiedKFold Cross-Validation Based on the Number of Features Selected by Lasso for AAL ,HO and CC200 atlases.

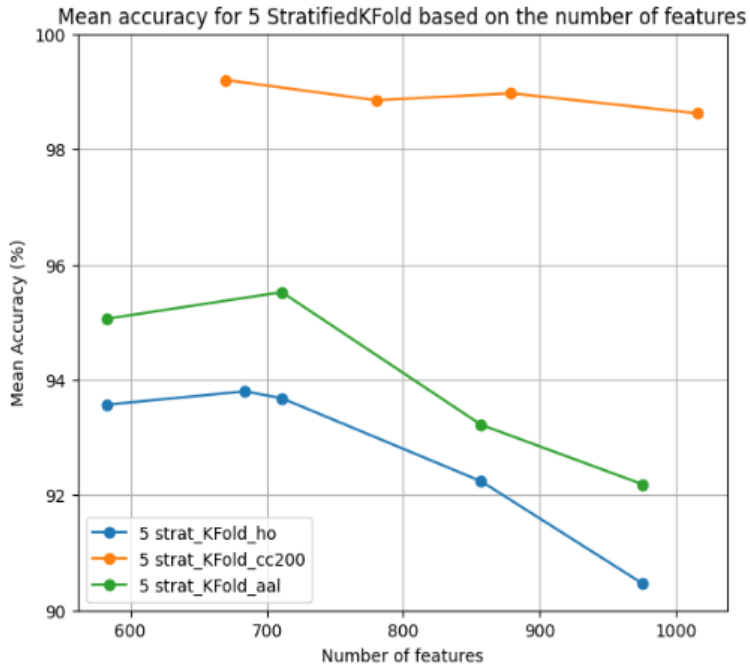

FIGURE 4.1.8 – Mean Accuracy for 5 StratifiedKFold Cross-Validation Based on the Number of Features Selected by Lassofor AAL ,HO and CC200 atlases.

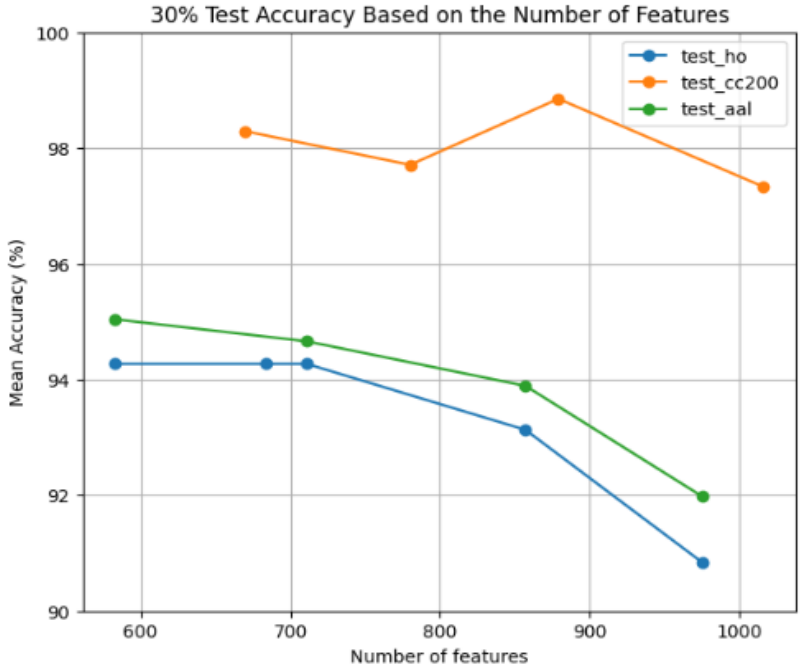

FIGURE 4.1.9 – 30% test accuracy Based on the Number of Features Selected by Lasso for AAL ,HO and CC200 atlases.

Atlas	alpha	n_features	10 KFold	10 StratifiedKFold	5 StratifiedKFold	30% test
HO	0.0005	711	95.53% (std=3.05%)	95.52% (std=2.64%)	93.68% (std=2.39%)	94.27%
CC200	0.001	670	99.43% (std=0.57%)	99.54% (std=0.56%)	99.20% (std=0.69%)	98.29%
AAL	0.0005	724	96.90% (std=1.99%)	96.10% (std=2.13%)	95.52% (std=1.47%)	94.66%

TABLE 4.3 – The best parameters obtained for the HO, AAL, and CC200 atlases using Lasso for feature selection and logistic regression for classification.

Feature selection based on RFE method :

The figures below provide a detailed summary of the model's performance, evaluated using cross-validation and Recursive Feature Elimination (RFE) for feature selection. This assessment was conducted for the three distinct brain atlases : HO , AAL, and CC200.
Each figure illustrates the results obtained for the model in terms of accuracy. Additionally, the figures also present the model parameters that led to the best results for each atlas. These parameters include the optimal values used during model training, providing valuable insights into the configurations that achieved the best performance.

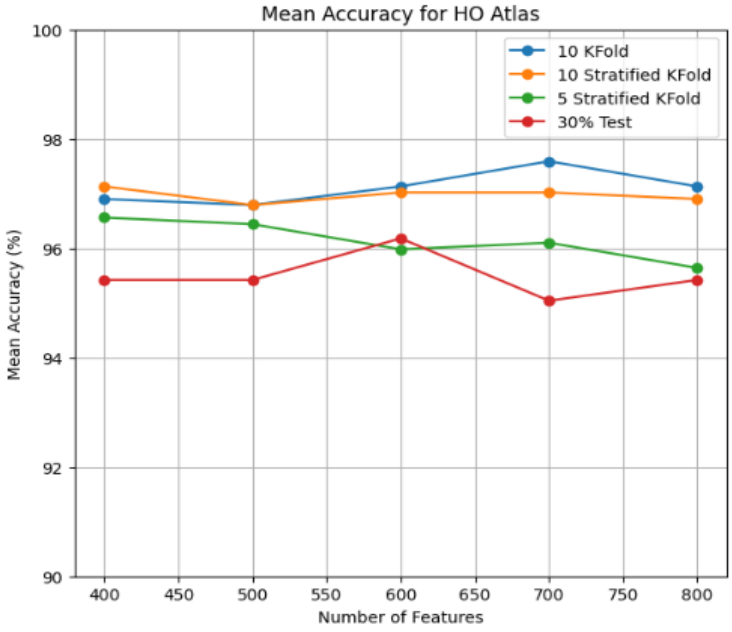

FIGURE 4.1.10 – The variation in average accuracy as a function of the number of features selected using the RFE method for the HO atlas.

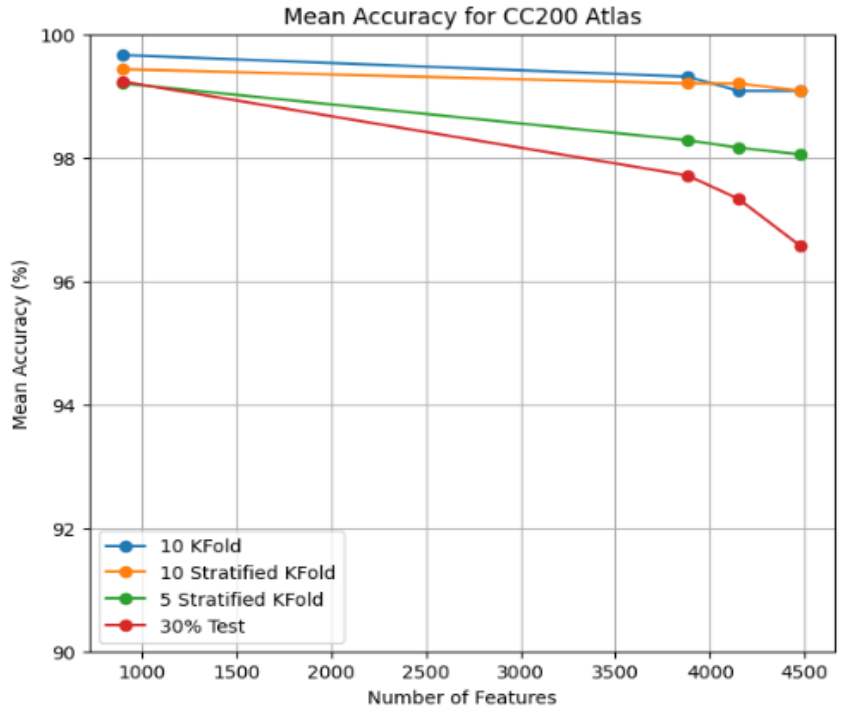

FIGURE 4.1.11 – The variation in average accuracy as a function of the number of features selected using the RFE method for the CC200 atlas.

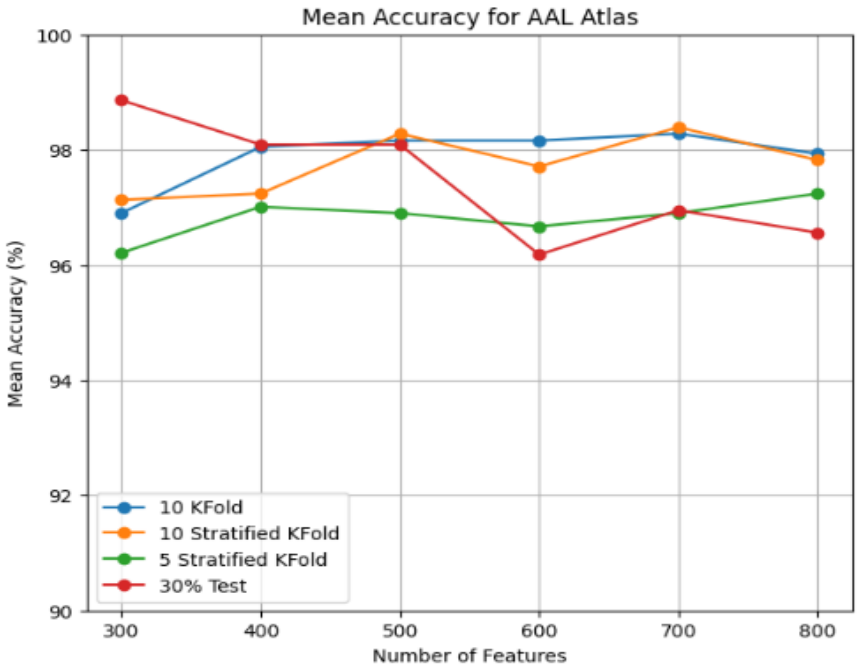

FIGURE 4.1.12 – The variation in average accuracy as a function of the number of features selected using the RFE method for the AAL atlas.

Atlas	n_features	10 KFold	10 StratifiedKFold	5 StratifiedKFold	30% test
HO	700	97.59% (std=2.20%)	97.02% (std=1.37%)	96.10% (std=1.23%)	95.04%
CC200	900	99.66% (std=0.52%)	99.43% (std=0.77%)	99.20% (std=0.69%)	99.23%
AAL	500	98.16% (std=1.95%)	98.28% (std=1.56%)	96.90% (std=1.29%)	98.09%

TABLE 4.4 – The best parameters obtained for the HO, AAL, and CC200 atlases using RFE for feature selection and logistic regression for classification.

After performing feature selection on the vector to identify the most relevant connections, we reinserted the selected indices into the original correlation matrix. We observed that all brain regions were present among the selected features, but not all possible correlations between these regions were included.
This means that the feature selection algorithm identified certain specific correlations between brain regions as the most important for distinguishing autistic subjects from non-autistic ones. However, it did not necessarily retain all possible correlations between all pairs of regions. In other words, certain specific interactions between pairs of brain regions are more informative for autism classification, while other interactions may be less relevant or redundant in this context.

4.1.1.5 Discussion

The results clearly demonstrate the importance of feature selection in enhancing the performance of machine learning models for classifying autistic and non-autistic subjects. By rigorously selecting the most relevant brain connections, we were able to identify specific correlations between brain regions that play a crucial role in distinguishing between these two groups. This approach reduces the complexity of the model while increasing its ability to capture the subtle nuances of brain interactions that are pertinent to autism diagnosis. In summary, feature selection proves to be an essential step in optimizing the efficiency and accuracy of classification models in this context.

4.2 Deep learning

Deep learning is a specialized area within machine learning that focuses on training artificial neural networks to simulate the structure and function of the human brain. This advanced form of machine learning has become increasingly popular due to its success in handling complex patterns and large datasets, especially in fields like computer vision, natural language processing, and speech recognition.

Artificial neural networks, the core of deep learning, are inspired by the human brain's neurons. These networks consist of multiple layers of interconnected nodes, or artificial neurons, that work together to analyze data and identify patterns. As data moves through each layer, the network refines its understanding, leading to a deeper and more abstract representation of the input.

Some essential concepts in deep learning include :

- Neural Networks : These models are structured with input, hidden, and output layers. The hidden layers, often called "deep" layers, are what give deep learning its name. The connections between neurons across layers are represented by weights, which are adjusted during the training process.
- Backpropagation : This is a key method used to train deep learning models. It involves calculating the error gradients for each weight in the network and using these gradients to update the weights, thereby reducing the model's error over time.
- Activation Functions : These functions introduce non-linearity into the network, allowing it to model complex relationships. Common examples include sigmoid, tanh, and ReLU.
- Convolutional Neural Networks (CNNs) : CNNs are specifically designed for grid-like data, such as images and videos. They use convolutional layers to automatically detect local patterns and spatial hierarchies in the data.
- Recurrent Neural Networks (RNNs) : RNNs are designed for sequential data like time series or natural language. They have connections that allow information to be passed across time steps, making them effective for tasks like language translation and speech recognition.
- Long Short-Term Memory (LSTM) : LSTM is a type of RNN that addresses issues like the vanishing gradient problem, enabling it to capture long-term dependencies in sequences.
- Generative Adversarial Networks (GANs) : GANs consist of two competing neural networks, a generator and a discriminator, which are trained together. They have been particularly successful in generating realistic data, such as images, and have many creative uses.

Deep learning has driven significant advancements across various fields, including image and object recognition, language translation, autonomous driving, drug discovery, and recommendation systems. Its ability to automatically identify complex patterns in data makes it a highly promising area of artificial intelligence. However, training deep learning models often requires large datasets and substantial computational resources, making them more demanding than traditional machine learning approaches.

4.2.1 Autism Dectecting Using Deep Learning

In our work, We attempted to reproduce the deep neural network (DNN) architecture used by the author in the paper [97]. This model, similar to the one

used in the previous section , takes a feature vector as input (without selection). The proposed architecture consists of two hidden layers, each with 32 neurons. A dropout layer with a probability of 80% was introduced between those layers to control overfitting. The hidden layers use ReLU activation function, while the final output layer uses the sigmoid activation function. Glorot and He weight initializers were used respectively with the sigmoid and ReLU activation functions. The Adam optimizer was employed with a relatively low learning rate of 0.0001. We used the same number of subjects as in the study, 866, and achieved an average accuracy of 60% ± 3.03% in a 5-fold stratified cross-validation by using Basc atlas.

On the other hand, we used the same architecture with the CC200 atlas, but after fine-tuning, we achieved an average precision of 70.14% with a standard deviation of ±2.84%, based on 5 stratified cross-validation folds. Additionally, with a 20% test split, we obtained a precision of 70.59%. The model accepts an input vector of size 19,900 and consists of three main layers : an input layer, two hidden layers, and an output layer.The two hidden layers each have also 32 neurons and use the ReLU activation function. To reduce overfitting, a dropout layer with a rate of 20% is added after the hidden layers. The output layer has a single neuron and uses the sigmoid activation function for binary classification. The model's weights are initialized with GlorotUniform. The model is optimized with the Adam optimizer, using a learning rate of 0.001. Training is performed with a batch size of 10 and 50 epochs. Additionally, an early stopping mechanism with a patience of 10 epochs is employed to halt training if the loss value does not improve. The figure below represents the confusion matrix using this model, where 0 represents neurotypical individuals and 1 represents individuals with autism.

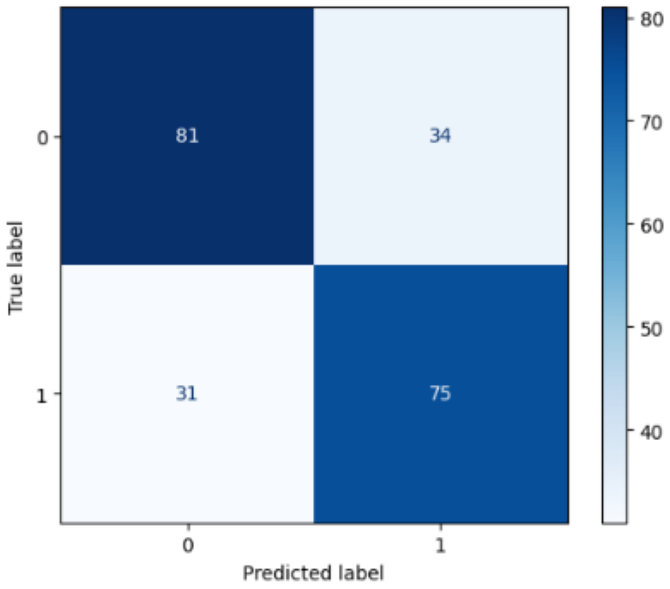

FIGURE 4.2.1 – Confusion matrix for 30% of the test set from the CC200 atlas using our fine-tuned DNN.

Feature selection based on the Lasso method :

We also attempted to select the 670 features using the Lasso method with the CC200 atlas, as done in the previous section, and used them as input for a deep learning model. The average accuracy achieved through 5-fold cross-validation was 98.28%, with a standard deviation of ± 0.37%, indicating the model's stability and reliability. Additionally, we achieved an accuracy of 98.47% on a separate 30% test set, further validating the model's performance.

For this experiment, the deep learning model's architecture was designed to be relatively simple but effective, considering the reduced input dimensionality. The model begins with an input layer that receives a vector of 670 features. This is immediately followed by a Dropout layer with a rate of 0.2, which helps prevent overfitting by randomly disabling 20% of the neurons during training.
The core of the model consists of a single hidden layer with 128 neurons, using the tanh activation function. The tanh function was chosen for its ability to map input values to a range between -1 and 1, providing a smooth gradient and avoiding the saturation issues that can occur with other activation functions.

Finally, the model's output layer contains a single neuron using the sigmoid activation function for binary classification. The model was compiled with the Adam optimizer, using a learning rate of 0.001, and was trained for 30 epochs with a batch size of 64, while reserving 20% of the training data for validation. This architecture effectively captured the complex relationships within the data while maintaining good generalization on the test sets.

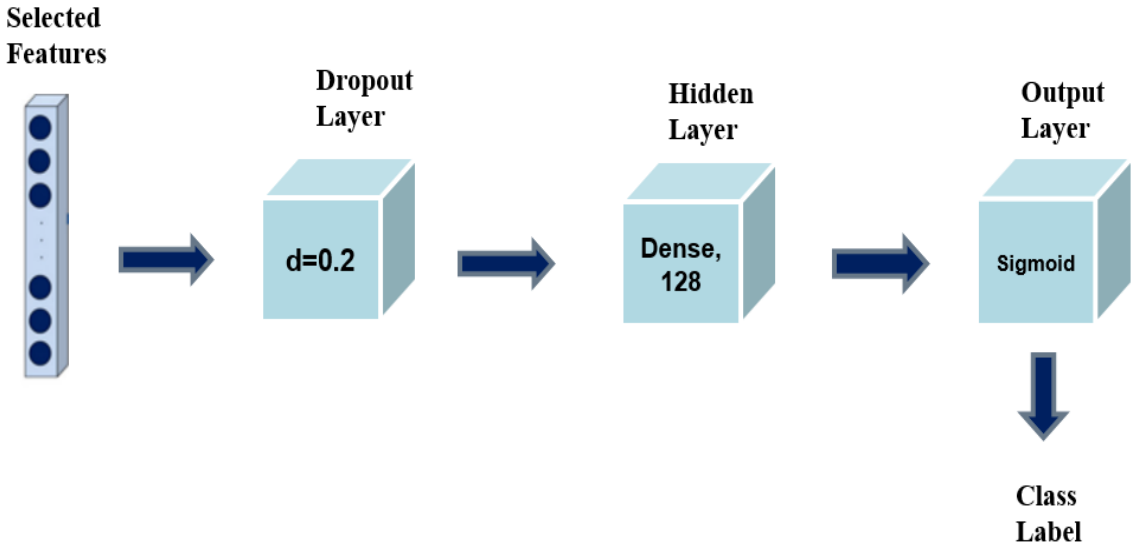

FIGURE 4.2.2 – Proposed deep neural network architecture for predicting ASD using features selected by Lasso and the CC200 atlas.

The figure below represents the confusion matrix using this model, where 0 represents neurotypical individuals and 1 represents individuals with autism.

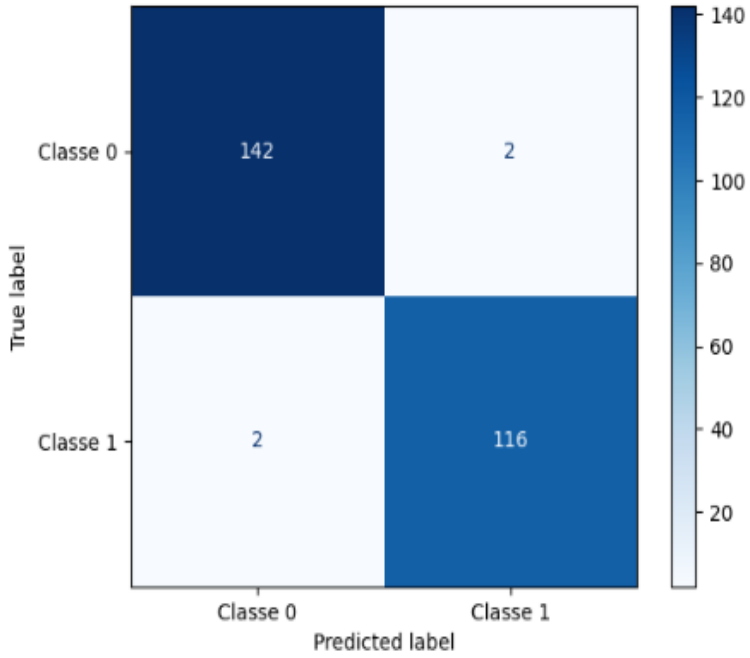

FIGURE 4.2.3 – Confusion matrix for 30% of the CC200 atlas test set using our proposed DNN with 670 features selected by Lasso.

4.2.1.1 Discussion

The results demonstrate that the performance of deep neural network configurations varied depending on the parameters and datasets used. The model using the CC200 atlas achieved an average accuracy of 70.14%, surpassing the 60% obtained with the Basc atlas. Additionally, another model using the CC200 atlas achieved an accuracy of 98.28% in cross-validation and 98.47% on a test set, after feature selection. These results highlight the importance of data quality and model simplicity in improving performance and clearly indicate the critical role of feature selection in optimizing results.

4.3 Conclusion

This chapter has thoroughly examined machine learning and deep learning models for identifying autism spectrum disorder (ASD). Various machine learning methods were assessed, with a focus on the importance of feature selection. When appropriate feature selection was applied, the performance of machine learning and deep learning models is much better then without using feature selection.

The results also indicate that, with adequate feature selection, the performance achieved with machine learning and deep learning models is nearly equivalent. This suggests that

feature selection plays a crucial role in enhancing model performance, regardless of the complexity of the approach used.

In summary, although deep learning models offer advanced capabilities for handling complex data, machine learning methods can achieve similar performance levels when effective feature selection is implemented.

Conclusion

This report provides a comprehensive examination of Autism Spectrum Disorder (ASD) through various lenses in neuroscience and brain imaging, focusing on advancements in diagnosis and data analysis. By establishing a solid foundation for understanding ASD, we explored its definition, associated risk factors, and the importance of early diagnosis. This exploration highlighted the inherent challenges of current diagnostic methods while emphasizing the complexity and variability of ASD symptoms.

Our analysis of brain development in relation to ASD demonstrated how brain imaging techniques, particularly functional magnetic resonance imaging (fMRI), are crucial for detecting and understanding connectivity abnormalities associated with autism. As a non-invasive method, fMRI is particularly suitable for use with children, offering valuable insights into altered brain networks in individuals with ASD.

In examining the data and methodologies essential for investigating ASD through neuroimaging, we discussed the ABIDE dataset and its preprocessed version. We provided a detailed description of this dataset, underscoring the importance of quality assessment to ensure the reliability of the data used in our analyses. The review of brain parcellation techniques, including the AAL, HO, CC200, CC400, and Dosenbach 160 atlases, revealed how each method offers unique advantages for segmenting the brain into regions of interest, which is fundamental for extracting meaningful time series data.

The time series derived from these parcellations were used to construct connectivity matrices, which are pivotal in understanding both functional and structural connectivity patterns in the brain. We discussed various types of connectivity matrices, such as correlation, partial correlation, covariance, and tangent space matrices for functional connectivity, as well as structural connectivity matrices. Each type of matrix provides a different perspective on brain connectivity, contributing to a comprehensive analysis of the neural networks involved in ASD.

In exploring the techniques and models used for autism diagnosis, we highlighted how MRI data can be represented and utilized within different classification models. We examined various representations of MRI data, including connectivity matrices, graph-based representations, 2D-slice images, 3D-volume reconstructions, time series analysis, and region-based representations. Each method offers a unique approach to capturing and interpreting intricate patterns in brain data, crucial for identifying neural signatures associated with ASD.

The assessment of machine learning and deep learning models for identifying ASD revealed the importance of feature selection. The results indicated that, with appropriate feature selection, machine learning models can achieve performance levels comparable to those of deep learning models. This suggests that feature selection plays a crucial role in enhancing model performance, regardless of the complexity of the approach used. In summary, while deep learning models offer advanced capabilities for handling complex data, machine learning methods can achieve similar performance levels when effective feature selection is implemented.

Overall, this report provides a thorough overview of brain imaging techniques, analytical methodologies, and machine learning and deep learning models used in studying ASD. The advancements in these areas pave the way for a better understanding and more accurate approaches for early diagnosis and intervention, offering promising perspectives for improving treatment and intervention strategies for individuals with ASD. In future work, we are thinking of integrating diverse data such as behavior analysis, eye tracking, and fMRI images, which could significantly enhance autism detection by capturing a broader range of behavioral and neurological markers. This multimodal approach may improve diagnostic accuracy and enable more personalized interventions for individuals on the autism spectrum.

Bibliographie

[1] Daniels, Amy M and David S Mandell, *Explaining differences in age at autism spectrum disorder diagnosis: A critical review*, Sage Journals Autism ,(2013).

[2] Billstedt, Eva, I. Carina Gillberg, and Christopher Gillberg, *Autism in adults: symptom patterns and early childhood predictors. Use of the DISCO in a community sample followed from childhood*, : Journal of Child Psychology and Psychiatry ,(2007).

[3] Taylor, Julie Lounds and Marsha Mailick Seltzer, *Changes in the Autism Behavioral Phenotype During the Transition to Adulthood*, Journal of Autism and Developmental Disorders ,(2010).

[4] *Autisme, la piste génétique*, Journal de la recherche de l'Institut Pasteur ,(2019).

[5] Dooley, Niamh and al. , *Is there an association between prenatal testosterone and autistic traits in adolescents?*, Psychoneuroendocrinology ,(2022).

[6] Bastaki, Kholoud N., Sura Alwan, and Farah R. Zahir, *Maternal Prenatal Exposures in Pregnancy and Autism Spectrum Disorder: An Insight into the Epigenetics of Drugs and Diet as Key Environmental Influences*, Advances in Neurobiology ,(2020).

[7] Angeline Mihailov, *Applying multidimensional approaches to disentangle autism spectrum disorder heterogeneity*, Université Paris-Saclay ,(2021).

[8] Dooley, Niamh and al. , *Is there an association between prenatal testosterone and autistic traits in adolescents?*, Psychoneuroendocrinology ,(2022).

[9] Bastaki, Kholoud N., Sura Alwan, and Farah R. Zahir, *Maternal Prenatal Exposures in Pregnancy and Autism Spectrum Disorder: An Insight into the Epigenetics of Drugs and Diet as Key Environmental Influences*, Advances in Neurobiology ,(2020).

[10] Havdahl, A. and al., *Genetic contributions to autism spectrum disorder*, Psychological Medicine ,(2021).

[11] Xie, Sherlly and al., *The Familial Risk of Autism Spectrum Disorder with and without Intellectual Disability*, Autism Research ,(2020).

[12] Pugsley, Kealan and al., *Environmental exposures associated with elevated risk for autism spectrum disorder may augment the burden of deleterious de novo mutations among probands*, Molecular Psychiatry ,(2021).

[13] Karimi, Padideh; Kamali, Elahe, *Environmental factors influencing the risk of autism*, Journal of Research in Medical Sciences ,(2017).

[14] Noens, Ilse and G. Nygren, *Comparisons Between the DISCO and the ADI-R and the ADOS*, International Meeting for Autism Research ,(2012).

[15] Slappendel, Geerte and al., *Utility of the 3Di Short Version for the Diagnostic Assessment of Autism Spectrum Disorder and Compatibility with DSM-5*, Journal of Autism and Developmental Disorders ,(2016).

[16] Dawkins, Tamara, Allison T. Meyer, and Mary E. Van Bourgondien, *The Relationship Between the Childhood Autism Rating Scale: Second Edition and Clinical Diagnosis Utilizing the DSM-IV-TR and the DSM-5*, Journal of Autism and Developmental Disorders ,(2016).

[17] Stiles, Joan and Terry L. Jernigan , *The Basics of Brain Development*, Neuropsychology Review ,(2010).

[18] Courchesne, Eric and al. , *Neuron Number and Size in Prefrontal Cortex of Children With Autism*, JAMA ,(2011).

[19] Hazlett, Heather Cody and al. , *Early brain development in infants at high risk for autism spectrum disorder*, Nature ,(2011).

[20] Tucker-Drob, Elliot M. and K. Paige Harden , *Gene-by-preschool interaction on the development of early externalizing problems*, Journal of Child Psychology and Psychiatry ,(2012).

[21] Hulzen, Kimm J.E. van and al. , *Genetic Overlap Between Attention-Deficit/Hyperactivity Disorder and Bipolar Disorder: Evidence From Genome-wide Association Study Meta-analysis*, Biological Psychiatry ,(2017).

[22] Hussain, Abul , *Growth and Development of Children*, SSRN Electronic Journal ,(2021).

[23] Nierhaus, Till et al. , *fMRI for the Assessment of Functional Connectivity*, Neuroimaging - Methods. InTech ,(2012).

[24] Frith, U. , *Autism: A Very Short Introduction*,(2008).

[25] Greicius, Michael , *Resting-state functional connectivity in neuropsychiatric disorders*, Current Opinion in Neurology,(2008).

[26] Edgar Canario, Donna Chen, Bharat Biswal , *A review of resting-state fMRI and its use to examine psychiatric disorders*, Psychoradiology,(2021).

[27] Pierre Desaunay, Fabian Guénolé, Francis Eustache, Jean-Marc Baleyte, Bérengère Guillery-Girard , *Autisme et connectivité cérébrale : contribution des études de neuroimagerie à la compréhension des signes cliniques*, Revue de neuropsychologie,(2014).

[28] Default mode network regions image

[29] *ABIDE dataset*

[30] Felix, Ebubeogu Amarachukwu and Sai Peck Lee , *Systematic literature review of preprocessing techniques for imbalanced data*, IET Software,(2019).

[31] Strother, S.C., *Evaluating fMRI preprocessing pipeliness*, IEEE Engineering in Medicine and Biology Magazine ,(2006).

[32] *ABIDE Preprocessed*

[33] Kamalaker Dadi, Mehdi Rahim, Alexandre Abraham and al. , *Benchmarking functional connectome-based predictive models for resting-state fMRI*, NeuroImage , (2019).

[34] *Quality Assessment*

[35] Tzourio-Mazoyer, N. et al., *Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain*, NeuroImage,(2002).

[36] Dosenbach, Nico U. F. and al. , *Prediction of Individual Brain Maturity Using fMRI*,Science (New York, N.Y.) ,(2010).

[37] Shumway, Robert H. and David S. Stoffer , *Time Series Analysis and Its Applications*, Springer International Publishing,(2017).

[38] Fulcher, Ben D., Max A. Little, and Nick S. Jones , *Highly comparative time-series analysis: the empirical structure of time series and their methods*, Journal of The Royal Society Interface ,(2013).

[39] B. Biswal, F. Zerrin 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 ,(1995).

[40] S. M. Smith, D. Vidaurre, C. F. Beckmann, M. F. Glasser, M. Jenkinson, K. L. Miller, T. E. Nichols, E. C. Robinson, G. Salimi-Khorshidi, M. W. Woolrich, et al., *Functional connectomics from resting-state fMRI*, Trends in Cognitive Sciences,(2013).

[41] J. Cabral, M. L. Kringelbach and G. Deco, *Functional connectivity dynamically evolves on multiple time-scales over a static structural connectome: Models and mechanisms*, NeuroImage,(2017).

[42] Nilearn github, *connectivity_ matrices*

[43] Erol, Aybüke; Hunyadi, Borbála, *Tensors for neuroimaging: A review on applications of tensors to unravel the mysteries of the brain*,Tensors for Data Processing ,(2022).

[44] F. Almuqhim, F. Saeed, *ASD-SAENet: A Sparse Autoencoder and Deep-Neural Network Model for Detecting Autism Spectrum Disorder (ASD) Using fMRI Data*, Front. Comput. Neurosci. 15, (2021).

[45] A. Sadiq, M.I. Al-Hiyali, N. Yahya, T.B. Tang, D.M. Khan, *Non-oscillatory connectivity approach for classification of autism spectrum disorder subtypes using resting-state fMRI*, IEEE Access 10 ,(2022).

[46] Z. Sherkatghanad, M. Akhondzadeh, S. Salari, M. Zomorodi-Moghadam, M. Abdar, U.R. Acharya, R. Khosrowabadi, V. Salari *Automated detection of autism spectrum disorder using a convolutional neural network*,Front. Neurosci. 13 ,(2020).

[47] J. Hu, L. Cao, T. Li, B. Liao, S. Dong, P. Li, *Interpretable learning approaches in resting-state functional connectivity analysis: the case of autism spectrum disorder*,Comput. Math. Methods Med. ,(2020).

[48] H. Jiang, P. Cao, M. Xu, J. Yang, O. Zaiane, *Hi-GCN, A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction*, Comput. Biol. Med. 127 ,(2020).

[49] J. Hu, L. Cao, T. Li, S. Dong, P. Li, *Li, GAT-LI: a graph attention network based learning and interpreting method for functional brain network classification*, BMC Bioinf. 22 ,(2021).

[50] Asrar G. Alharthi, Salha M. Alzahrani , *Do it the transformer way: A comprehensive review of brain and vision transformers for autism spectrum disorder diagnosis and classification* , Computers in Biology and Medicine ,(2023).

[51] M.S. Ahammed, S. Niu, M.R. Ahmed, J. Dong, X. Gao, Y. Chen, *DarkASDNet: classification of ASD on functional MRI using deep neural network*, Front. Neuroinf. 15 ,(2021).

[52] N. Dominic, Daniel, T.W. Cenggoro, A. Budiarto, B.Pardamean, *Transfer learning using inception-ResNet-v2 model to the augmented neuroimages data for autism spectrum disorder classification*, Communications in Mathematical Biology and Neuroscience ,(2021).

[53] Kristen Coyne,*MRI: A Guided Tour*

[54] H. Shahamat, M.S. Abadeh *Brain MRI analysis using a deep learning based evolutionary approach*, Neural Network. 126 ,(2020).

[55] X. Chen, Z. Wang, Y. Zhan, F.A. Cheikh, M. Ullah, *Interpretable Learning Approaches in Structural MRI: 3D-ResNet Fused Attention for Autism Spectrum Disorder Classification*, SPIE ,(2022).

[56] B. Hasan Atakan, S¸. Irmak, D. Onat, D. Salman Ul Hassan, Ç. Tolga, *BolT: Fused Window Transformers for fMRI Time Series Analysis*, arXiv e-prints ,(2022).

[57] X. Deng, J. Zhang, R. Liu, K. Liu, *Classifying ASD based on time-series fMRI using spatial–temporal transformer*,Comput. Biol. Med. 151 ,(2022).

[58] R. Liu, Z.A. Huang, Y. Hu, Z. Zhu, K.C. Wong, K.C. Tan, *Spatial–temporal Co-attention learning for diagnosis of mental disorders from resting-state fMRI data*, IEEE Transact. Neural Networks Learn. Syst. ,(2023).

[59] K. Gao, Z. Fan, J. Su, L.-L. Zeng, H. Shen, J. Zhu, D. Hu, *Deep transfer learning for cerebral cortex using area-preserving geometry mapping*, Cerebr. Cortex 32 ,(2021).

[60] A. Othmani, T. Bizet, T. Pellerin, B. Hamdi, M.-A. Bock, S. Dev, *Significant CC400 Functional Brain Parcellations Based LeNet5 Convolutional Neural Network for Autism Spectrum Disorder Detection*, Springer Nature Switzerland ,(2023).

[61] U.S. Shanthamallu, A. Spanias, *Machine and Deep Learning Algorithms and Applications*, Springer International Publishing,(2022).

[62] Zhao, X. Zhang, K.H. Thung, N. Mao, S.W. Lee, D. Shen, *Constructing multi-view high-order functional connectivity networks for diagnosis of autism spectrum disorder*,IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 69 ,(2022).

[63] F. Huang, E.-L. Tan, P. Yang, S. Huang, L. Ou-Yang, J. Cao, T. Wang, B. Lei, *Self-weighted adaptive structure learning for ASD diagnosis via multi-template multi-center representation*,Med. Image Anal. 63,(2020).

[64] C. Yang, P. Wang, J. Tan, Q. Liu, X. Li, *Autism spectrum disorder diagnosis using graph attention network based on spatial-constrained sparse functional brain networks*,Comput. Biol. Med. 139 ,(2021).

[65] N. Chaitra, P.A. Vijaya, G. Deshpande, *Diagnostic prediction of autism spectrum disorder using complex network measures in a machine learning framework*,Biomed. Signal Process Control 62 ,(2020).

[66] P. Lu, X. Li, L. Hu, L. Lu, *Integrating Genomic and Resting State fMRI for Efficient Autism Spectrum Disorder Classification*,Multimedia Tools and Applications,(2021).

[67] M. Cao, M. Yang, C. Qin, X. Zhu, Y. Chen, J. Wang, T. Liu, *Using DeepGCN to identify the autism spectrum disorder from multi-site resting-state data*,Biomed. Signal Process Control 70,(2021).

[68] L. Squarcina, G. Nosari, R. Marin, U. Castellani, M. Bellani, C. Bonivento, F. Fabbro, M. Molteni, P. Brambilla, *Automatic classification of autism spectrum disorder in children using cortical thickness and support vector machine*,Brain and Behavior 11 ,(2021).

[69] M. Raki´c, M. Cabezas, K. Kushibar, A. Oliver, X. Llado, *Improving the detection of autism spectrum disorder by combining structural and functional MRI information*,Neuroimage: Clinical 25,(2020).

[70] M.T. Ali, Y. ElNakieb, A. Elnakib, A. Shalaby, A. Mahmoud, M. Ghazal, J. Yousaf, H.A. Khalifeh, M. Casanova, G. Barnes, A. El-Baz, *The role of structure MRI in diagnosing autism*,Diagnostics 12,(2022).

[71] A. Brahim, N. Farrugia *Graph Fourier transform of fMRI temporal signals based on an averaged structural connectome for the classification of neuroimaging*,Artif. Intell. Med. 106 ,(2020).

[72] X. Ma, X.-H. Wang, L. Li, *Identifying individuals with autism spectrum disorder based on the principal components of whole-brain phase synchrony*,Neurosci. Lett. 742,(2021)

[73] J. Zhang, F. Feng, T. Han, X. Gong, F. Duan,*Detection of Autism Spectrum Disorder Using fMRI Functional Connectivity with Feature Selection and Deep Learning*,Cognitive Computation,(2022).

[74] F. Almuqhim, F. Saeed, *ASD-SAENet, A sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using fMRI data*,Front. Comput. Neurosci. 15 ,(2021).

[75] Y. Wang, J. Wang, F.-X. Wu, R. Hayrat, J. Liu, *AIMAFE: autism spectrum disorder identification with multi-atlas deep feature representation and ensemble learning*,J. Neurosci. Methods 343 ,(2020).

[76] M. Ingalhalikar, S. Shinde, A. Karmarkar, A. Rajan, D. Rangaprakash, G. Deshpande,*Functional connectivity-based prediction of autism on site harmonized ABIDE dataset*,Trans. Biomed. Eng. 68,(2021).

[77] Y. Liu, L. Xu, J. Yu, J. Li, X. Yu, *Identification of autism spectrum disorder using multi-regional resting-state data through an attention learning approach*,Biomed. Signal Process Control 69 ,(2021).

[78] H. Sewani, R. Kashef, *An autoencoder-based deep learning classifier for efficient diagnosis of autism*,Children 7 ,(2020).

[79] H. Sharif, R.A. Khan, *A Novel Machine Learning Based Framework for Detection of Autism Spectrum Disorder (ASD)*,Applied Artificial Intelligence,(2021).

[80] A. Sadiq, M.I. Al-Hiyali, N. Yahya, T.B. Tang, D.M. Khan, *Non-oscillatory connectivity approach for classification of autism spectrum disorder subtypes using resting-state fMRI*,IEEE Access 10 ,(2022).

[81] N. Wang, D. Yao, L. Ma, M. Liu, *Multi-site clustering and nested feature extraction for identifying autism spectrum disorder with resting-state fMRI*,Med. Image Anal. 75,(2022).

[82] M. Kunda, S. Zhou, G. Gong, H. Lu, *Improving multi-site autism classification via site-dependence minimization and second-order functional connectivity*,IEEE Trans. Med. Imag. ,(2022).

[83] J. Liu, Y. Sheng, W. Lan, R. Guo, Y. Wang, J. Wang, *Improved ASD classification using dynamic functional connectivity and multi-task feature selection*,Pattern Recogn. Lett. 138,(2020).

[84] C. Alvarez-Jimenez, N. Múnera-Garzon, ´ M.A. Zuluaga, N.F. Velasco, E. Romero, *Autism spectrum disorder characterization in children by capturing local-regional brain changes in MRI*, Med. Phys. 47,(2020).

[85] Y. Liu, L. Xu, J. Li, J. Yu, X. Yu, *Attentional connectivity-based prediction of autism using heterogeneous rs-fMRI data from CC200 atlas*,Experimental Neurobiology 29,(2020).

[86] L. Zhao, Y.-K. Sun, S.-W. Xue, H. Luo, X.-D. Lu, L.-H. Zhang, *Identifying boys with autism spectrum disorder based on whole-brain resting-state interregional functional connections using a boruta-based support vector machine approach*,Front. Neuroinf. 16,(2022).

[87] O. Dekhil, M. Ali, Y. El-Nakieb, A. Shalaby, A. Soliman, A. Switala, A. Mahmoud, M. Ghazal, H. Hajjdiab, M.F. Casanova, A. Elmaghraby, R. Keynton, A. El-Baz, G. Barnes, *A personalized autism diagnosis CAD system using a fusion of structural MRI and resting-state functional MRI data*,Front. Psychiatr. 10 ,(2021).

[88] M.A. Reiter, A. Jahedi, A.R.J. Fredo, I. Fishman, B. Bailey, R.-A. Müller, *Performance of machine learning classification models of autism using resting-state fMRI is contingent on sample heterogeneity*,Neural Comput. Appl. 33,(2021).

[89] P. Moridian, N. Ghassemi, M. Jafari, S. Salloum-Asfar, D. Sadeghi, M. Khodatars, A. Shoeibi, A. Khosravi, S.H. Ling, A. Subasi, R. Alizadehsani, J.M. Gorriz, S.A. Abdulla, U.R. Acharya, *Automatic autism spectrum disorder detection using artificial intelligence methods with MRI neuroimaging: a review*,Front. Mol. Neurosci. 15,(2022).

[90] S. Itani, D. Thanou, *Combining anatomical and functional networks for neuropathology identification: a case study on autism spectrum disorder*, Med. Image Anal. 69,(2021).

[91] J. Wang, L. Zhang, Q. Wang, L. Chen, J. Shi, X. Chen, Z. Li, D. Shen, *Multi-class ASD classification based on functional connectivity and functional correlation tensor via multi-source domain adaptation and multi-view sparse representation*,IEEE Trans. Med. Imag. 39 ,(2020).

[92] R. Kashef, *ECNN: enhanced convolutional neural network for efficient diagnosis of autism spectrum disorder*, Cognit. Syst. Res. 71,(2022).

[93] M. Emeç , M. Yurtsever , *DA novel ensemble machine learning method for accurate air quality prediction*, International Journal of Environmental Science and Technology,(2024).

[94] A. Kazeminejad, R.C. Sotero, *The importance of anti-correlations in graph theory based classification of autism spectrum disorder*, Front. Neurosci. 14,(2020).

[95] J. Hu, L. Cao, T. Li, B. Liao, S. Dong, P. Li, *Interpretable learning approaches in resting-state functional connectivity analysis: the case of autism spectrum disorder*, Comput. Math. Methods Med.,(2020).

[96] L. He, H. Li, J. Wang, M. Chen, E. Gozdas, J.R. Dillman, N.A. Parikh, *A multi-task, multi-stage deep transfer learning model for early prediction of neurodevelopment in very preterm infants*, Sci. Rep. 10,(2020).

[97] F.Z. Subah, K. Deb, P.K. Dhar, T. Koshiba, *A Deep Learning Approach to Predict Autism Spectrum Disorder Using Multisite Resting-State fMRI*,11, Applied Sciences,(2021).

[98] X. Yang, P.T. Schrader, N. Zhang, *A deep neural network study of the ABIDE repository on autism spectrum classification*, IJACSA International Journal of Advanced Computer Science and Applications ,(2020).

[99] M.S. Ahammed, S. Niu, M.R. Ahmed, J. Dong, X. Gao, Y. Chen, *DarkASDNet: classification of ASD on functional MRI using deep neural network*, Front. Neuroinf. 15 ,(2021).

[100] T.M. Epalle, Y. Song, Z. Liu, H. Lu, *Multi-atlas classification of autism spectrum disorder with hinge loss trained deep architectures: ABIDE I results*,Appl. Soft Comput. 107 ,(2021).

[101] Samvardhan Singh, *Types of neural networks: Convolutional Neural Networks*, (2023)

[102] Ayşe ALMACI, *NLP (Natural Language Processing) Techniques, (2023)*

[103] L. Shao, C. Fu, Y. You, D. Fu, *Classification of ASD based on fMRI data with deep learning*,Cognitive Neurodynamics 15,(2021).

[104] *ABIDE Preprocessed*, (accessed on 18 August 2024).