UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Investigating Functional Connectivity in Autism Spectrum Disorder with Graph Neural Networks

Henrique Vicente Lopes

Mestrado em Engenharia Biomédica e Biofísica

Dissertação orientada por:
Professor Doutor Alexandre Andrade

2024

"Research is formalized curiosity. It is poking and prying with a purpose."
— Zora Neale Hurston

Agradecimentos

A dissertação que aqui se apresenta é o culminar de 18 meses de dedicação e resiliência. Mas não é o resultado apenas deste período, decorre dos 5 anos passados na FCUL, e é em grande parte influenciada pelos Professores que tive o prazer de conhecer ao longo do meu percurso. Começarei por agradecer ao leitor por estar interessado nesta jornada.

Um agradecimento especial ao Professor Alexandre Andrade pela orientação, por se mostrar sempre disponível a ajudar-me, por estar tão ou mais entusiasmado do que eu pelo projeto, desde o início, mesmo quando as GNNs pareciam indecifráveis. A sua ajuda foi da maior importância. Agradeço à Professora Raquel pela disponibilidade e dedicação, ao Professor Nuno Matela pelo interesse que demonstra pelos seus alunos e por tornar qualquer assunto interessante, ao Professor Hugo Ferreira por nos incentivar a questionar, e à Professora Brígida Ferreira pela sua ajuda.

Uma palavra de agradecimento a todos os professores que contribuíram para o meu crescimento e que fizeram parte do meu percurso escolar, este trabalho também é um bocadinho vosso. Foram um exemplo desde o primeiro dia: professora Cristina, professora Edite, professora Graça, professora Lina, professor Pedro Gabriel, e professora Telma.

Quero agradecer a dois colegas e amigos que este projeto me trouxe e sem os quais este período teria sido muito mais desafiante e menos interessante: Duarte e João; ao Duarte pelo enorme interesse que sempre demonstrou pelo projeto, a tua motivação nos momentos em que não estava a correr tão bem foram preciosos; ao João pelas palavras de motivação que nunca faltaram; a ambos pelo apoio. Se não continuarmos colegas, que continuemos amigos. Agradeço à Leonor Pires, a minha conterrânea e companheira de tese, por toda a ajuda e por cada palavra amiga dada durante estes meses. Agradeço aos restantes amigos que a FCUL me trouxe e que fizeram parte destes maravilhosos 5 anos: Bea, Duarte (o CEO), Hayanna, Laura, Paiva e Oliveira, Maria, e Rafa.

Um obrigado aos meus amigos de Tavira, que seja desde a infância até às fraldas: Paulo, Dani, Afonso, Ian, e Pablo.

Agradeço à grande Bea e à grande Sara, que sem ninguém contar apareceram e tornaram a vida ainda mais agradável. Obrigado pelas jantaradas, pelas férias, e pelas aventuras, as já passadas e as que ainda estão por vir.

Obrigado à minha família, sem a qual nada disto seria possível. Obrigado mãe, por me conheceres tão bem desde pequeno, por me cativares para a Ciência, por seres constantemente aquela vozinha de apoio, por me ensinares que nada é impossível. Obrigado pai, por me ajudares a contrabalançar os estudos com o desporto, por me dares espaço para crescer, pelo orgulho que sempre vi nos teus olhos. Foram o início de tudo, obrigado por tudo. Um obrigado à Dina, por cuidar de mim como filho. Obrigado mana, é um orgulho ser teu irmão. Um obrigado aos meus avós, Idalina e António, por serem um exemplo de resiliência e por todo o apoio. À avó Alice que já não estando cá fisicamente estará sempre comigo.

Um obrigado à eterna Professora e à Tia Célia, por todo o apoio e motivação. À Cristiana que não sendo minha irmã, é como se fosse.

A ti, Mariana, por seres a minha companhia de todas as horas, por todo o amor e carinho. Obrigado por seres o meu pilar, por acreditares que consigo fazer tudo, principalmente obrigado por estares sempre lá, nem que seja para um sorriso rápido, obrigado pelo apoio incondicional. Palavras jamais serão suficientes.

Abstract

Autism Spectrum Disorder (ASD) is a complex array of neurodevelopmental disorders characterized by challenges in communication, social interaction, and restrictive and repetitive behaviors. Due to its heterogeneity and the lack of objective biomarkers, ASD diagnosis relies heavily on behavioral assessments, which may not accurately reflect its prevalence. This dissertation addresses this diagnostic challenge by exploring the potential of Graph Neural Networks (GNNs) to identify ASD.

GNNs can analyze relational data by representing individuals as nodes and their relationships (e.g., age, sex, and brain connectivity) as edges. This study leverages resting-state functional magnetic resonance imaging (fMRI) data from 871 subjects (403 with ASD, 468 typically developing) from the ABIDE database. The data were preprocessed to generate functional connectivity (FC) matrices, which were then structured into a graph format for GNN analysis.

A GNN model was developed and optimized using graph attention network (GAT) layers. The model was trained and tested in an inductive learning context to ensure generalizability. The model demonstrated stable performance compared to other GNN-based models in the literature.

To interpret the model's classifications, the GNNExplainer method was used to identify the most relevant brain region functional connections for ASD diagnosis. The study highlighted disruptions in visual processing, linguistic, and self-referential brain regions, consistent with existing literature. A significant finding was the under-connectivity between the left occipital temporal fusiform cortex and the left planum polare in ASD, related to face and object recognition.

This research underscores the potential of GNNs in improving ASD diagnosis by incorporating relational data. While the model did not surpass all existing deep learning approaches, it offered insights into the challenges of high-dimensional data and highlighted the importance of explainable AI methods in medical diagnostics.

Keywords: Graph Neural Network, Explainable Artificial Intelligence, Autism Spectrum Disorder, Functional Connectivity, Deep Learning

Resumo

A Perturbação do Espetro do Autismo (PEA) abrange um grupo clinicamente diverso de distúrbios do neurodesenvolvimento, caracterizados por traços comportamentais que afetam a comunicação e a interação social, bem como comportamentos e interesses restritivos e repetitivos. O termo "espetro" reflete a ampla gama de severidade nas dimensões cognitivas e comportamentais entre os indivíduos diagnosticados. As causas e mecanismos exatos da PEA permanecem pouco claros, sendo considerada um distúrbio complexo influenciado por fatores genéticos e ambientais e as suas interações. As estimativas de prevalência da PEA situam-se entre 1% e 1,5% da população em geral. Considerando a falta de biomarcadores clinicamente relevantes, o diagnóstico da PEA ainda se baseia na apresentação comportamental, o que faz com que a taxa de deteção atual seja inferior à sua prevalência real. Esta lacuna destaca a necessidade de investigar os mecanismos neuronais subjacentes à PEA e desenvolver métodos de diagnóstico mais objetivos. A PEA é altamente heterogénea em termos de etiologia, fenótipo e desenvolvimento, dificultando que modelos de aprendizagem profunda desenvolvidos obtenham bons resultados na tarefa de diagnóstico, pois estes recorrem apenas a informação individual sobre cada sujeito, sem considerar a informação relacional entre os sujeitos. Recentemente, as Redes Neurais de Grafos (RNGs) surgiram como uma abordagem promissora para enfrentar essa heterogeneidade. As RNGs podem operar em grafos onde cada nó representa um sujeito e cada aresta representa a relação entre dois sujeitos com base em informações complementares, como a idade ou o sexo. Esta abordagem tem potencial para contribuir para uma melhor compreensão e diagnóstico da PEA. Alguns estudos têm proposto RNGs para o diagnóstico da PEA com este tipo de abordagem. Contudo, a maioria utiliza um tipo de contexto de aprendizagem chamado de aprendizagem transdutiva, no qual a RNG é treinada e testada no mesmo grafo, não apresentando capacidade de generalização para novos dados. Por outro lado, num contexto médico é preponderante que as classificações realizadas pelo modelo sejam interpretáveis. Por conseguinte, métodos de inteligência artificial explicável (IAE) têm sido desenvolvidos especificamente para aumentar a transparência das decisões feitas por redes neurais de grafos. Deste modo, nesta dissertação, o objetivo centrou-se em explorar o potencial das RNGs para diagnosticar PEA, ao desenvolver uma RNG num contexto de aprendizagem indutiva em que o modelo é capaz de generalizar para dados independentes do grafo de treino, enquanto se investigam as alterações de conectividade funcional (CF) associadas à doença através de um método de IAE.

Para atingir estes objetivos, foram usados dados de ressonância magnética funcional (RMf) em estado de repouso de 871 sujeitos (403 com PEA e 468 com desenvolvimento típico). Estas imagens são provenientes da base de dados ABIDE e já tinham sido previamente processadas no âmbito do *Preprocessed Connectomes Project* (PCP). A abordagem de pré-processamento incluiu: correção temporal, correção de movimento, normalização da intensidade do voxel e remoção de ruído dos dados, através da regressão de potenciais efeitos perturbadores (como movimento da cabeça, respiração e pulsação cardíaca). Incluiu também filtragem com um filtro passa-banda (0,01 - 0,1 Hz) e os dados foram alinhados com o cérebro padrão *Montreal Neurological Institute 152* (MNI152). Cada cérebro foi segmentado em regiões

corticais e subcorticais utilizando o atlas anatómico Harvard-Oxford de 111 regiões. Foi gerada a matriz de CF para cada indivíduo recorrendo ao cálculo do coeficiente de correlação de Pearson entre as séries temporais extraídas de cada região. O conjunto das matrizes de CF foi dividido em conjuntos de desenvolvimento (80%) e teste (20%) para que, após o desenvolvimento do modelo de diagnóstico, a sua capacidade de generalização pudesse ser avaliada. Em seguida, os dados foram estruturados em formato de grafo. Os nós representaram individuos e foram representados pelas matrizes de CF vetorizadas. As arestas entre os sujeitos representavam as relações entre eles e foram baseadas na similariedade de idade, sexo, local de aquisição das imagens de RMf e na correlação entre as respetivas matrizes de CF. As arestas que ligavam sujeitos com pouca similaridade foram removidas.

Após a preparação dos dados, a RNG foi desenvolvida e otimizada. A rede neural de grafos construída é composta por quatro camadas de *graph attention network* (GAT) seguidas por um classificador com duas camadas totalmente conectadas. Nas camadas GAT foram incorporadas *skip connections* para evitar o problema frequente em RNGs de *oversmoothing*. Como hiperparâmetros a otimizar foram escolhidos: a função de ativação, o número de filtros de cada camada GAT, a taxa de aprendizagem, a taxa de *dropout*, o número de épocas de treino, e o parâmetro L2 de regularização. A otimização foi efetuada através da validação cruzada com 10 *folds* no conjunto de dados de desenvolvimento. Da otimização de hiperparâmetros, foi selecionado o modelo que apresentou os melhores valores de média de cada métrica. O modelo com melhor desempenho incluiu a função de ativação PReLU, 64 filtros em cada camada GAT, uma taxa de aprendizadem de 0,0001, uma taxa de *dropout* de 0,4 e o parâmetro L2 igual a $5,0 \times 10^{-2}$. Foi necessário desenvolver um critério de *early stopping* para regularizar mais o modelo e, particularmente, diminuir os efeitos de flutuações significativas no valor de *accuracy* de validação do modelo, que foram encontradas durante o treino.

Para verificar como o modelo desenvolvido se comparava a outros modelos atuais da literatura que usam RNGs para a classificação de PEA, o modelo foi comparado com dois modelos conhecidos, comparação essa que ocorreu recorrendo à aprendizagem transdutiva para uma comparação justa. Desta comparação, notou-se que as flutuações encontradas nos valores da *accuracy* de validação durante o treino também ocorriam em ambos os modelos de comparação. Em geral, o modelo desenvolvido apresentou resultados mais estáveis e melhor desempenho.

Em seguida, o modelo com o conjunto de hiperparâmetros que obteve a melhor *performance* foi treinado no conjunto de desenvolvimento e avaliado no conjunto de teste independente num contexto de aprendizagem indutiva. O modelo provou ser menos generalizável do que o ideal, notando-se principalmente uma dificuldade em classificar indivíduos com autismo. Este é, de facto, um problema comum a vários modelos de classificação de PEA na literatura que usam o ABIDE, e está associado à heterogeneidade inerente à doença e ao conjunto de dados do ABIDE, que inclui dados de vários locais com diferentes protocolos de aquisição, o que dificulta que o modelo capte toda a variablidade dos dados. Para além disso, considerando que a dimensionalidade das características dos nós é demasiado elevada relativamente ao número de dados, hipotetizou-se que o modelo está a sofrer do fenómeno de *curse of dimensionality*. Este fenómeno explica as flutuações presentes no valor da *accuracy* de validação e a menor capacidade de generalização.

Para tentar interpretar as classificações efetuadas pelo modelo no conjunto de dados independente, recorreu-se ao *GNNExplainer*, um método de IAE criado especificamente para atuar em RNGs, baseado na perturbação das características dos dados para identificar as características mais relevantes para a classificação feita pela RNG. Com o *GNNExplainer*, foram identificadas as conexões funcionais entre regiões cerebrais com maior relevância para o diagnóstico da PEA. De forma a verificar o impacto destas para os diagnósticos feitos pelo modelo, foram efetuados estudos de ablação, nos quais se removeram as

conexões funcionais mais importantes. Estes estudos, para além de mostrarem a relevância das conexões identificadas, permitiram verificar a existência de características redundantes, o que não é surpresa num caso de elevada dimensionalidade das características dos dados. As regiões cerebrais identificadas e as suas conexões funcionais enfatizam os papéis do processamento visual, particularmente do reconhecimento facial, e do processamento linguístico e autorreferencial na PEA. As disrupções encontradas são frequentemente relatadas na literatura e contribuem para os sintomas principais da PEA, estando especificamente associadas com as dificuldades na comunicação social. No entanto, apenas uma das dez conexões identificadas resistiu à análise estatística entre a amostra de indivíduos com PEA e os controlos e se mostrou significativa - a subconectividade entre o córtex fusiforme occipital temporal esquerdo e o *planum polare* esquerdo na PEA, que poderá estar associada ao reconhecimento e nomeação de objetos e faces de pessoas.

Os resultados desta dissertação revelam o potencial das RNGs para o diagnóstico da PEA. Apesar do modelo não ter superado a *performance* reportada por outros estudos que utilizam modelos de aprendizagem profunda na literatura, conseguiu superar dois dos modelos de classificação que usam RNGs para a classificação de PEA com o ABIDE, que foram treinados nas mesmas condições que o modelo desenvolvido. Inclusive, permitiu identificar outra causa possível para a dificuldade em criar modelos ótimos de diagnóstico para PEA usando este tipo de abordagem, para além da heterogeneidade da doença e da base de dados - a elevada dimensionalidade das características dos dados. Por outro lado, explorou um método de IAE na RNG desenvolvida, com resultados concordantes com a literatura relativa às disfunções de conexões funcionais encontradas associadas à PEA. Assim, os objetivos delineados para esta dissertação foram cumpridos, e este trabalho poderá contribuir para a superação dos problemas patentes no desenvolvimento de modelos de inteligência artificial de diagnóstico da PEA.

Palavras chave: Rede Neural de Grafos, Inteligência Artificial Explicável, Perturbação do Espetro do Autismo, Conectividade Funcional, Aprendizagem Profunda

Contents

List of Figures

List of Tables

Acronyms

AAL Automated Anatomical Labeling. 18, 41

AAP American Academy of Pediatrics. 10

ABIDE Autism Brain Imaging Data Exchange. v–vii, xiii, 2, 35, 37, 38, 40, 42, 43, 45–48, 63, 64, 73

AD Alzheimer's disease. 16, 40

ADAM adaptive moment estimation. 22, 50

ADHD attention-deficit/hyperactive disorder. 1, 8, 11, 64

ADI-R Autism Diagnostic Interview-Revised. 10

ADOS-2 Autism Diagnostic Observation Schedule. 10, 34

AE autoencoder. 38

AI artificial intelligence. 20, 34, 42, 73

APA American Psychiatric Association. 4, 9

ASD autism spectrum disorder. x–xiv, 1–13, 16–19, 24, 34–40, 42–50, 52, 54–58, 60–75, 96, 97

ATP adenosine triphosphate. 15

AUC area under the receiver operating characteristic curve. xiv, 40, 51, 52, 56–58, 62

BOLD blood-oxygen-level-dependent. x, 13, 15, 16, 19, 43, 44

CARS Childhood Autism Rating Scale. 10

CEN central executive network. 12, 17

CNN convolutional neural network. 26, 29, 38

CNVs copy number variations. 5

CPAC Configurable Pipeline for the Analysis of Connectomes. 38, 43

CSF cerebrospinal fluid. 11

DAN dorsal attention network. 17

DL deep learning. 2, 3, 13, 19, 20, 22, 29, 30, 34–39, 42, 64, 74

DMN default mode network. 12, 16, 17, 36, 37, 72

DNN deep neural network. 38

DSM the Diagnostic and Statistical Manual of Mental Disorders. xii, 1, 4, 5, 7, 9, 10, 96

DTI diffusion tensor imaging. 11, 34

EEG electroencephalography. 18, 34

EV-GCN Edge-Variational Graph Convolutional Network. xi, xiii, xiv, 40, 41, 50, 53, 58–61

FC functional connectivity. xi–xiv, 2, 16–19, 34–40, 42, 44–46, 48, 49, 54, 55, 63–71, 73, 74, 97

fMRI functional magnetic resonance imaging. 2, 3, 11, 13, 15, 16, 18, 19, 34, 35, 37, 38, 43, 64

fNIRS functional near-infrared spectroscopy. 34

GABA gamma-aminobutyric acid. 12

GAT graph attention network. xi, xiii, 27, 28, 31, 46, 47, 49–52, 56, 74

GCN graph convolutional network. xi, xiii, 26–28, 39–41, 48

GNN graph neural network. x, xii–xiv, 1–3, 20, 23–32, 34, 39–42, 45–50, 54–56, 59–63, 65, 66, 68, 69, 73, 74, 97

GRL graph representation learning. 23, 49

hi-GCN hierarchical Graph Convolutional Network. 40

HO Harvard-Oxford. xii, xiv, 18, 41, 43, 44, 67, 98–100

ICA independent component analysis. 19, 37

ID intellectual disabilities. 1, 8

IQ intelligence quotient. 38

JK jumping knowledge. x, 29, 30, 50

KNN k-nearest neighbors. 41

KS2 two-sample Kolmogorov-Smirnov. 47

LG-GNN local-to-global graph neural network. xi, xiv, 40, 41, 53, 58–62

LSTM long short-term memory. x, 29, 30, 39

MAMF-GCN multi-scale adaptive multichannel fusion deep graph convolutional network. 40, 41

MEG magnetoencephalography. 18, 34

ML machine learning. 13, 20, 22, 33–38, 42, 49, 63, 73

MLP multilayer perceptron. x, 21, 26, 38

MNI152 Montreal Neurological Institute 152. 18, 38, 43

MNS mirror neuron system. 7

MR magnetic resonance. 15

MRI magnetic resonance imaging. xiii, 11–15, 38, 43–46

NDD neurodevelopmental disorder. 1, 10

NMR nuclear magnetic resonance. 13

NN neural network. 2, 3, 20–23, 39

OCD obsessive-compulsive disorder. 8

PAE pairwise association encoder. 41

PCP Preprocessed Connectomes Project. 35, 38, 43, 44

PDD-NOS pervasive developmental disorder-not otherwise specified. 5

PET positron emission tomography. 11, 12, 34

RBF Radial Basis Function. 37

ReLU rectified linear unit. vi, xi, 21, 40, 50–52, 56

RF radio-frequency. 13, 14

RFE recursive feature elimination. 40, 48

ROC receiver operating characteristic curve. 51, 52

ROI region of interest. xii, xiv, 18, 19, 37, 38, 41, 43, 44, 48, 64–67, 71, 72, 97–100

rs-fMRI resting-state functional magnetic resonance imaging. 2, 12, 13, 16–19, 34–40, 42–45, 73

RSN resting-state network. x, 12, 16–19, 36, 37, 74

SCA seed-based correlation analysis. 19

SCQ Social Communication Questionnaire. 10

SLD specific learning disorder. 1

SMN somatomotor network. 12, 17, 36

sMRI structural magnetic resonance imaging. 11, 12, 34, 35, 38

SN salience network. 12, 17, 37

SPECT single-photon emission computerized tomography. 11, 12

SRS Social Responsiveness Scale. 10

SVM Support Vector Machine. 34, 38

TD typical developing. xi–xiv, 1, 2, 6, 11, 12, 19, 34–38, 42–44, 46, 49, 50, 52, 56–58, 60–63, 67, 69–73

TE echo time. xiii, 14, 44, 45

TE-HI-GCN Ensemble of Transfer HIerarchical Graph Convolutional Networks. 40

ToM Theory of Mind. 36, 72

TR repetition time. xiii, 14, 44, 45

XAI explainable artificial intelligence. xii, 42, 54, 64, 68, 70, 73, 74

Chapter 1

Introduction

This dissertation proposes a graph neural network (GNN) model to explore brain functional connectivity abnormalities to discriminate between autism spectrum disorder (ASD) and typical developing (TD) individuals. This chapter exposes the context and motivation of the main topics addressed in this dissertation project, as well as the objectives and structure overview.

1.1 Context and Motivation

In 2013, the Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5) [1] introduced a new diagnostic category called neurodevelopmental disorders (NDDs) of conditions with onset in early childhood that can persist through life [2]. These disorders are associated with the disruption of the events that lead to normal brain development [3]. In this way, NDDs are characterized by delays in the expected social, emotional, language, cognitive, and/or movement milestones, including intellectual disabilities (ID), communication disorders, autism spectrum disorder (ASD), attention-deficit/hyperactive disorder (ADHD), specific learning disorder (SLD), and motor disorders [1, 4], that affect 10-15% of the general population [5].

DSM-5 [1] recognizes that these disorders have related symptomatology and very frequently co-occur [2, 6], which complicates the diagnosis and, consequently, the estimation of the true prevalence of NDDs.

ASD is one of the most common NDDs and is characterized by impaired social communication and interaction, and by restricted, repetitive interests and behaviors [7], with epidemiological studies stating that its prevalence has risen substantially in the last two decades [5, 8, 9]. Although there is some controversy if the cause of this increase is the change in the diagnosis criteria over the years, the broadening of ASD definition, or the increasing awareness of ASD, these factors are considered a complement to the true unexplained increase in the prevalence of the disorder [8, 10], with recent estimates being between 1% and 1.5% of the general population [7]. Yet, according to Zwaigenbaum *et al.* [7], the current detection rate of ASD is lower than its actual prevalence. This is expected considering that ASD is highly heterogeneous in etiology, phenotype, and outcome. Multiple genes and environmental factors are involved in the ASD developmental course of symptom expression that in tandem with the co-occurrence of medical and mental comorbidities, are responsible for the clinical heterogeneity of ASD [5, 9]. Despite the risk factors already identified, there is no consensual etiology for ASD, as well as any biomarker clinically relevant, with ASD diagnosis being based on behavioral observation [11, 12].

Moreover, undiagnosed individuals are likely to underperform academically and experience social, emotional, and behavioral problems [4]. The estimated annual societal costs of ASD related to services

and lost productivity by patients and their parents are $236bn in the US and $47.5bn in the UK [7, 13].

Therefore, it becomes preponderant to study the neural mechanisms underlying the impairment observed in ASD subjects, and the development of more objective diagnostic methods for rapid identification and subsequent intervention in subjects at risk, enhancing long-term outcomes [4].

The advent of non-invasive neuroimaging techniques allowed researchers and clinicians to study the human brain, enabling advancements in the understanding of its structural and functional connections in healthy, developing, aging, and diseased brains [14]. In fact, with the development of functional magnetic resonance imaging (fMRI), neuroimaging biomarkers related to functional connectivity abnormalities, which are observed in individuals suffering from ASD, have been vastly investigated [15].

For many years, neuroimaging data analysis was based on statistical analysis to evaluate differences between groups of individuals [16]. Recently, deep learning (DL) has been gaining traction in the medical field, as DL algorithms provide promising solutions at the point where traditional analytical methods are inefficient, particularly for the analysis of neuroimaging data, as they can extract abstract and complex patterns that characterize neurologic and psychiatric disorders [16]. Furthermore, in recent years, graph theory insights and DL were integrated resulting in a new category of neural networks (NNs) called graph neural networks (GNNs). Graphs are a very intuitive method to model a large population of individuals (nodes) and their associations or similarities (edges) in addition to allowing the integration of different data sources efficiently. Node features in a graph can be derived from neuroimaging data of each individual, while the edge features can represent the similarities between individuals based on their phenotypic information, particularly in disease cases, which can help facilitate their diagnosis [17]. Thus, graphs seem a very promising approach to model the continuous spectrum of ASD as it considers the relationship information between ASD individuals, but also typical developing (TD) subjects, and thus GNNs may have the power to surpass the heterogeneity problem of ASD and ultimately improve its diagnosis.

1.2 Objectives

The main goal of this dissertation is to apply a GNN to the classification of FC data derived from resting-state functional magnetic resonance imaging (rs-fMRI) of ASD and TD subjects, investigating FC abnormalities in ASD and ultimately improving its diagnosis. To achieve this, intermediate goals need to be accomplished, namely:

1. Literature review of the current state-of-the-art of DL methods applied to ASD diagnosis, particularly GNNs, and functional connectivity abnormalities found to be associated with ASD.

2. Extraction of rs-fMRI data from Autism Brain Imaging Data Exchange I (ABIDE I) database and computation of the FC matrices.

3. Development and optimization of the proposed GNN.

4. Analysis of the incorporation of additional data to the model, such as phenotypic and clinical data, for the improvement of the GNN performance.

5. Application of the developed GNN to an independent test set.

1.3 Dissertation Overview

This dissertation is composed of 6 chapters. Chapter 1 contextualizes, and enumerates the main objectives of this study. Chapter 2 provides fundamental theoretic concepts about fMRI, ASD, and DL, namely general concepts regarding NNs and GNNs. The state-of-the-art to support this study is presented in Chapter 3. Chapter 4 discusses the materials and methods employed during this project. Chapter 5 includes the results and respective discussion. Finally, the overall conclusions as well as limitations and recommendations for future work are discussed in Chapter 6.

Chapter 2

Background Theory

2.1 Autism Spectrum Disorder

Autism spectrum disorder (ASD) is a broader term used to describe a clinically heterogeneous group of neurodevelopmental disorders that share core behavioral traits that impact social communication and interaction and include restrictive and repetitive stereotypical behavior patterns and interests [1, 18]. ASD affects diagnosed children very differently regarding the severity along many cognitive and behavioral dimensions, hence the term "spectrum" [19]. Until now, the exact etiology and pathogenesis of ASD are not fully understood, although it is believed that ASD is a complex multifactorial disorder with genetic and environmental factors, and their interaction influencing the development of this disorder [19, 11].

The present section aims to succinctly cover important time points in the history of ASD as a concept and a diagnostic construct, describe its clinical features, summarize some of the modern theories regarding the pathophysiology of this disorder, and describe the current diagnostic criteria, as well as associated challenges and prospects. Furthermore, neuroimaging techniques and developments in ASD will be discussed, including previous findings, limitations, and future directions.

2.1.1 Brief History of ASD

The first descriptions of Autism as a diagnostic construct occurred in the 1940s, with the independent works of Leo Kanner and Hans Asperger. In 1943, child psychiatrist Leo Kanner employed the term "autism" in his classic report "Autistic disturbances of affective contact" [20] to describe children with an extreme inability to relate to others, sensory sensitivity, impairment in communication, a tendency to interpret things literally, and repetitive behaviors [9]. Independently, in 1944, pediatrician Hans Asperger coined the term "autistic psychopathy" [21] to detail a condition affecting a different group of children that experienced impairments in nonverbal communication, empathizing, and related social skills, but had above-average linguistic skills [22, 23]. The core symptoms identified by Asperger were very similar to the ones identified by Kanner, but in higher-functioning individuals [9]. The "autistic psychopathy" of Asperger was later renamed "Asperger Syndrome" by child psychiatrist Lorna Wing in 1976 [24] that in her paper [25] recognized the similarities between the criteria proposed by Kanner and Asperger and introduced the idea that those two conditions were varieties of the same underlying abnormality [9].

Only in 1980, with the publication of the Diagnostic and Statistical Manual of Mental Disorders 3 (DSM-3) [26] published by the American Psychiatric Association (APA), autism was recognized as an official diagnostic category called "infantile autism" as was a very restricted definition in onset and development. In 1994, was published the Diagnostic and Statistical Manual of Mental Disorders 4 (DSM-4)

[27] with a separate set of criteria for Asperger's Syndrome. The criteria for Asperger's Syndrome differed from those for Autistic Disorder only in that they did not include impairments in language or cognition [22, 9]. In 2013 was published the 5^{th} and current edition of The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [1], which comprises the most used diagnostic criteria for ASD, worldwide, since its publication. In this edition, the conceptualization of autism heralded a substantial shift from a multi-categorical diagnostic system to a single diagnosis based on multiple dimensions with the collapse of the former diagnoses of autistic disorder, Asperger's Syndrome, and pervasive developmental disorder-not otherwise specified (PDD-NOS) into one unifying umbrella term as ASD [22, 9, 28].

In the last decades, the definition of ASD has evolved considerably, as well as the understanding regarding its pathophysiology and etiology. Today, ASD is a highly researched disorder, considered highly variable, in etiology [9], phenotype and development [29].

2.1.2 Pathophysiology

The etiology of autism spectrum disorder (ASD) is thought to be influenced by genetic, environmental, and a combination of both types of factors. It is a complex, multifactorial disorder, and the contribution of genetic and environmental factors to the etiology of the disorder is considered to vary from case to case [11, 30].

a. Genetic factors

ASD is deemed the most heritable complex neurodevelopmental disorder [31], with estimated heritability ranging from approximately 40% to 90% [18]. Multiple genetic etiologies have been postulated as contributing to the disorder, including rare, spontaneous single-gene mutations, interactions between common functional variants of multiple genes, or copy number variations (CNVs) [10]. More than 100 genes and genomic regions have been associated with ASD [18]. Some of the genes are involved in the development and functioning of the brain, including those that regulate the growth of neurons and the formation of synapses. Mutations or changes in these genes can lead to disruptions in the development and functioning of the brain, which can result in the symptoms of autism [32, 33, 34]. These genetic factors may occur separately or in combination to determine an individual's risk of developing the disorder, consequently leading to the diversity in both the spectrum and intensity of observed symptoms [10].

b. Environmental factors

Environmental factors have also been found to increase the risk of ASD. These factors include advanced parental age, preterm birth, birth trauma associated with hypoxia, maternal obesity, a short interval between pregnancies, gestational diabetes mellitus, valproate use during pregnancy [18], maternal nutritional status, prenatal stress, infection during pregnancy, exposure to toxins, heavy metals, or drugs [28]. However, whether these factors are causal or merely markers of risk is still unclear [18].

c. Gene-Environmental interaction factors

Environmental factors interact with genetic factors, contributing to existent heterogenous genetic profiles and fostering the clinical heterogeneity of autism spectrum disorder (ASD) [34]. The interplay between genetic and environmental factors, including epigenetic factors, influence the risk of ASD through several complex underlying mechanisms that lead to abnormalities in the devel-

opment and functioning of neural circuits in the brain *in utero* or after birth, ultimately inducing the disorder symptoms [30, 6, 35] (see Figure 2.1).

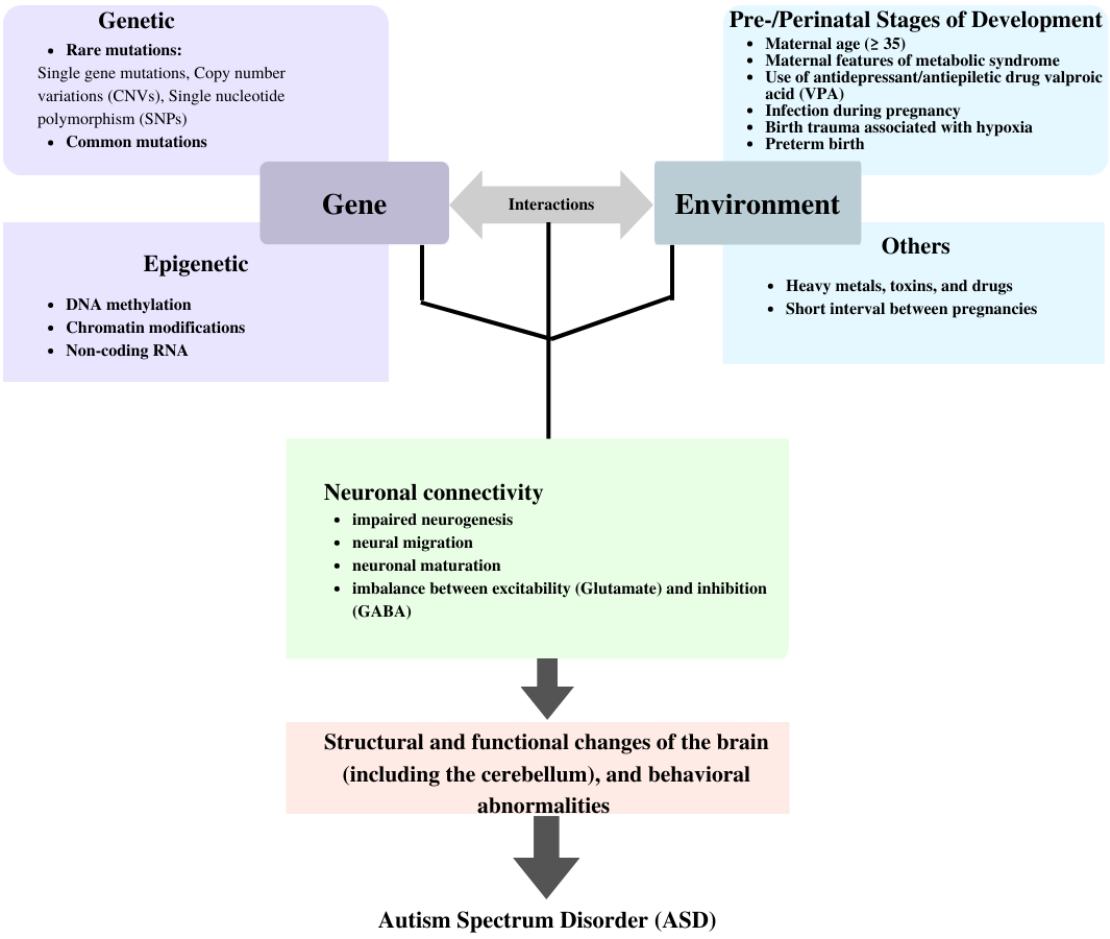


Figure 2.1: Pathophysiological mechanisms of ASD (adapted from [11]).

Considering the existence of multiple possible risk factors associated with ASD, research has focused on finding a degree of convergence regarding the developmental pathways of the disorder. One of the most cited explanatory models of atypical brain functioning in ASD is the neural connectivity model introduced by Belmonte *et al.* (2004) [36]. Since its publication it has been investigated and supported by several studies [37, 38, 39, 40, 41]. This framework proposes that ASD is primarily characterized by abnormalities in the development and functioning of neural circuits in the brain, which lead to deficits in social communication, sensory processing, and other cognitive functions. The theory is based on numerous post-mortem and *in vivo* neuroimaging findings (further discussed in section 2.1.5). One of the most often reported findings is the brain volume overgrowth in early stages of development in ASD individuals compared with TD controls [42, 37, 40], which is thought to compromise the shaping and fine-tuning of neural connections and consequently disrupt the processing of information [43]. Another common finding thought to be associated with these disruptions [35, 34] is the increased number of neurons found in autistic individuals [44]. Evidence of abnormalities in neural migration during the antenatal period has also been found, which contributes to brain malformation, as ectopic neurons have delayed maturation, and insufficient dendritic spine pruning [45, 39]. In this way, it supports the neuronal connectivity model

to explain the clinical manifestations of the disorder [45, 39]. In addition, aberrant functional connectivity present in ASD has been related to an imbalance of excitatory to inhibitory neurotransmission, particularly an increase of excitatory signaling and a decrease in inhibitory signaling [46, 38], in sensory, mnemonic, social and emotional systems [47].

Taken together, the disruptions in neuronal connectivity present in ASD can arise from impaired neural migration, synaptogenesis, dendritic morphogenesis, or an excitation-inhibition imbalance. While the neuronal connectivity framework is potentially the most validated theory to explain the pathogenesis of autism, offering a foundational understanding of how disrupted circuits contribute to ASD, other theories have been proposed that try to help explain the specifics of which circuits are disrupted and how these disruptions lead to distinct symptoms.

- **The "Broken-Mirror" Theory**

  Mirror neurons are brain cells part of the mirror neuron system (MNS) that fire when an individual performs an action and when the individual observes someone else performing the same action. This system is thought to play a role in understanding others' intentions, imitating actions, and developing empathy [48]. According to this hypothesis, in individuals on the autism spectrum, this system might not function properly, which is proposed to be the primary cause of their social disability, as affects action recognition, motion mimicry, theory of mind (capacity to understand subjective mental states), empathy, and language in autistic individuals [35, 49].

- **The Amygdala Theory**

  The amygdala, a brain region implicated in the processing of emotions and social information, is posited within this theory as a conceivable pivotal neural locus in the pathophysiology of autism. Structural and functional abnormalities in the amygdala could result in difficulties in recognizing and interpreting emotional cues, which are essential for understanding social situations, and observable in autistic individuals [50, 51]. Currently, it is suggested that the amygdala may not be singly responsible for the ASD deficits but abnormal connections between the amygdala and other brain regions may better explain them [52]. Indeed, the deficits in social behavior observed in ASD may be attributed to potential disturbances within the "social" brain network. This network encompasses functions such as facial recognition, attribution of mental states, emotional consciousness, self-reflection, empathy, and the discernment of social interactions and judgments, of which the amygdala constitutes a component [49].

Other theories could have been mentioned, but these were the ones considered more relevant to this project's scope. One can understand how they are not mutually exclusive and complement each other. Indeed, considering the complex nature of ASD, a single brain region is unlikely to account for all the traits of the disorder.

2.1.3 Clinical Presentation

Considering the complex and heterogeneous etiology described, it is anticipated that the clinical presentation is also characterized by substantial heterogeneity [9]. According to DSM-5, ASD is characterized by a core group of symptoms that consist of impaired social interaction and the presence of repetitive and inflexible behavior and circumscribed interests causing significant impairment in major life areas [1]. Moreover, its manifestations are extensive, depending on the severity, developmental level, and chronical age of the subjects [1].

The clinical features are often most marked in early childhood, with the common first symptoms involving delayed language development, frequently accompanied by a lack of interest in social interactions, unusual communication patterns, and absence of typical playing with odd and repetitive behaviors often present [1], with common onset before the age of three [31, 28]. There is growing evidence that autism has a heterogeneous developmental trajectory with possible subgroups identified, ranging from individuals with worsening symptoms to individuals with improving symptoms [18]. In some individuals, most symptoms of ASD persist until adulthood, especially regarding social functioning, although communication skills can improve over time as individuals reach adolescence and adulthood [53]. By contrast, in a smaller portion of individuals diagnosed as children, no clinically meaningful symptom is detectable - the so-called "Optimal Outcome" [18, 54]. Currently, there is no evidence if this outcome is associated with early intervention or is an etiologically distinct subtype of autism [18]. Additionally, there are individuals that although core symptoms of ASD are no longer an impairment in daily life, still struggle with novel situations and suffer from the anxiety needed to consciously try to understand what is socially intuitive for most individuals [18, 1].

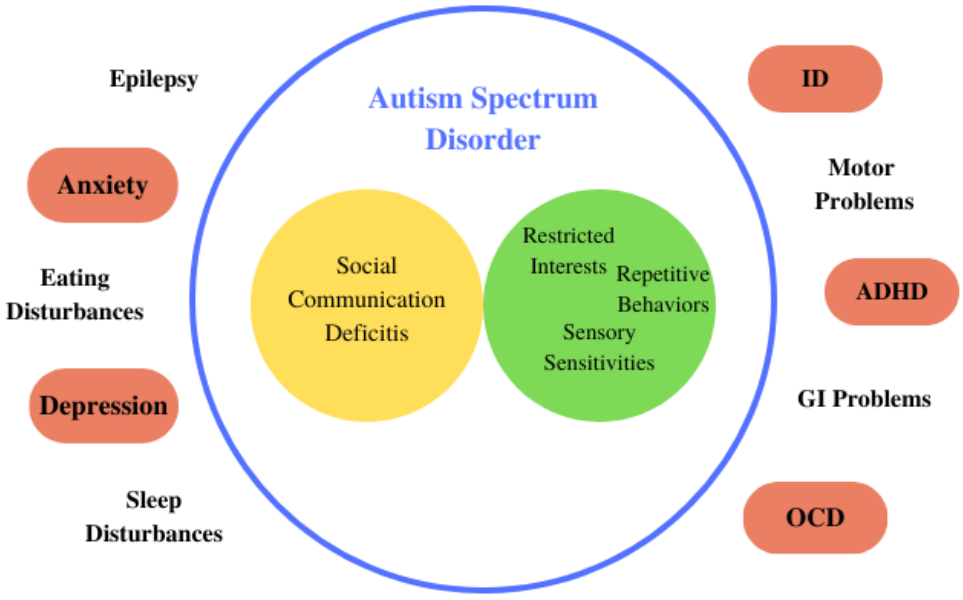


Figure 2.2: ASD core symptoms and co-occurring psychiatric and medical conditions (adapted from [22] and based on [1]).

Besides, secondary symptoms can include self-injury, hyperactivity, impulsivity, aggression, and co-occurring psychiatric disorders and medical conditions [30, 1] (see Figure 2.2). The psychiatric disorders include ADHD, obsessive-compulsive disorder (OCD), anxiety, major depression, and ID [18] and the medical conditions that can co-occur with ASD are epilepsy, sleep and eating disturbances, motor problems, constipation [1], and gastrointestinal problems [28]. In addition, reports show increases in cytokine levels and inflammation, with immune system dysfunction being also considered a comorbidity of ASD [55]. The presence of one of these comorbidities is likely to be linked to more severe autism-related core symptoms [28], as well as independence and well-being at each age [18]. The degree to which the mentioned symptoms affect each individual is widely variable, ranging from individuals profoundly affected — where an autistic person may be non-verbal, with severe ID and unable to function without significant support — to relatively high-functioning – where the individual is verbal with average or above-average intelligence that can live an independent life [56] (see Figure 2.3). Moreover, increasing evidence sug-

gests that individuals suffering from ASD have premature mortality [57, 58], increased risk of self-harm and ultimately suicide [18].

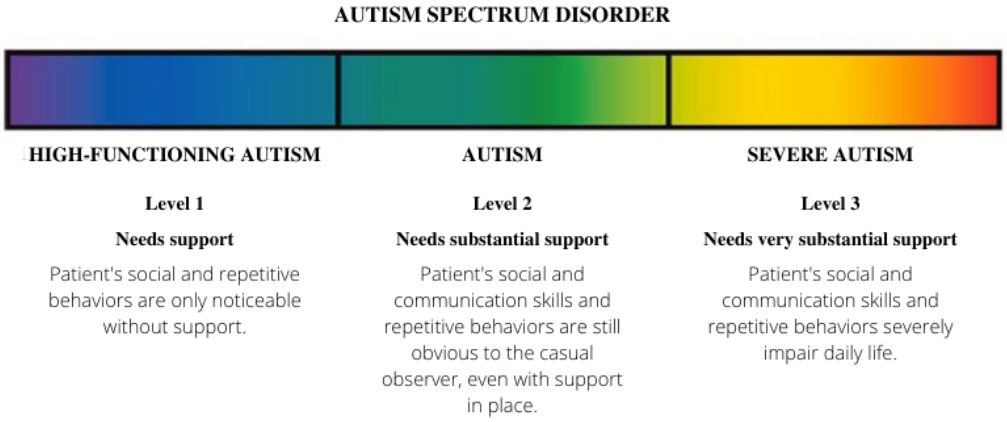


Figure 2.3: Visualization of the severity of ASD (adapted from [59]).

ASD is a very debilitating disorder with no cure [60]. Current treatment interventions aim at improving communication and social skills, play and daily living competencies, academic skills, inappropriate behavior, and ultimately the quality of life of individuals with ASD [60, 28]. The educational and behavioral interventions employed depend on the varied symptoms and functioning levels of each autistic individual [28]. In light of the limitations experienced by young children with autism in terms of communication and social interaction, early intervention is considered a critical priority [24, 18]. Such interventions hold promise for improving the individual's capacity to learn and enhancing the comprehension of parents who may otherwise struggle to understand their child's perplexing behavior. Notably, intervention efforts conducted during the preschool years, which coincide with heightened brain plasticity, may yield additional benefits, although the empirical evidence for this theory remains uncertain [18]. Some individuals can also be prescribed medication to treat the associated symptoms or co-occurring diagnoses, although currently is not used medication to treat the core symptoms of ASD [18].

Hence, ASD has substantial variability and heterogeneity regarding its clinical features and comorbidities, which vary in severity and can change during the lifespan. ASD is associated with individualized treatment, mainly focused on behavioral interventions, that should start to be implemented as soon as possible, to have the best chance of overcoming some of the disabilities. Thereby, it is fundamental to better understand the neurobiological mechanisms to provide biological-based diagnostic tools to allow an early diagnosis and subsequent intervention.

2.1.4 Diagnosis

Currently, the diagnosis of ASD is made based on behavioral presentation, by an experienced clinician [18], as outlined in APA's DSM-5 diagnostic criteria [61, 1], with an average age of diagnosis of 4.5 years [11]. This comprises five criteria (A-E) that should be met.

To be diagnosed with ASD, a person must show evidence of persistent deficits, past or present, in each of three social communication and interaction subdomains (Criterion A), and must have or have had impairment in two of the four different restricted, repetitive patterns of behavior, interests, or activities (Criterion B) [6, 1]. These symptoms have to be present from early childhood, but may not fully manifest until later or may be masked by learned strategies in later life (Criterion C). Besides, symptoms must

cause clinically significant impairment in social, occupational, or other important areas of current functioning (Criterion D). For last, the present disturbances are not better explained by intellectual disability or global developmental delay (Criterion E) [1].

In addition to establishing the diagnosis, the severity is assessed to determine the functioning level and identify co-morbid conditions. The severity should be separately rated for social communication and restricted, repetitive behaviors, with the recognition that severity may vary by context and change over time [1]. The DSM-5 suggests the use of specifiers to qualify ASD diagnoses. There are specifiers to describe the severity that spans through three levels (Figure 2.3), such as "requiring substantial support", and also relative to intellectual impairment, language deficits, and psychological and medical co-morbid conditions that may be present, for example "with or without accompanying language impairment" [22, 1]. Thus, the dimensional approach to diagnosing ASD effectively addresses both the shared features of core symptoms, while capturing the heterogeneity present in terms of the quantity and quality of these symptoms, thereby enhancing diagnostic specificity, facilitating more effective treatment planning, and enabling the identification of distinctive subgroups within the ASD population with differing developmental trajectories. [22].

The process of detection of ASD and other NDDs includes a screening phase followed by an exhaustive developmental examination phase that can confirm the diagnosis [62]. The American Academy of Pediatrics (AAP) recommends universal developmental surveillance and screening that comprehends ASD specific screening at 18 and 24 or 30 months [63]. Furthermore, high-risk children, such as with a sibling previously diagnosed with ASD, with birth complications or low birth weight, are advised to undergo additional developmental risk assessment [62].

Many assessment tools to evaluate if the diagnostic criteria are met have been developed to assist with the diagnosis of ASD in individuals at risk. These include parental questionnaires and interviews, clinical judgments, and direct interactions [61]. Nevertheless, clinicians often implement two tests considered the gold-standard for diagnosing ASD [61, 62]. The Autism Diagnostic Observation Schedule (ADOS-2) is a semi-structured, standardized measure of communication, social interaction, play/imagination, and restricted or repetitive behaviors, applied by a trained clinician to the individual being diagnosed [64]. During the evaluation, the clinician observes the subject performing a variety of imaginative activities and social tasks that normally evoke spontaneous behavior [30]. The Autism Diagnostic Interview-Revised (ADI-R) is a semi-structured interview with the parent or caretaker to review the developmental history and current behavior of the individual being evaluated, regarding language, reciprocal social interactions, and restrictive, repetitive, and stereotyped behaviors and interests [65]. Other screening tools can be used to assess ASD symptoms, such as the Social Responsiveness Scale (SRS), the Social Communication Questionnaire (SCQ), and the Childhood Autism Rating Scale (CARS) [24]. Thus, the diagnosis of ASD is based on combined clinician observation, caregiver reports, and assessment tools results.

While these screening tools are considered the gold standard, they still have a considered plethora of limitations [61, 62]. For instance, (i) interviews response rely on the memory and subjective understanding of the assessment questions by the caregiver [61, 62], contributing to evaluation and screening biases [62] in addition to be further complicated in adult populations [66]; (ii) the clinician responsible for the ASD detection need to be submitted to significant training, influencing the availability of clinicians but also infrastructures globally to assist ASD detection and management [62]; (iii) the current ASD questionnaires require a considered amount of time to implement, for example, the ADI-R can take between 90 and 150 minutes to complete [65]; and (iv) the diagnostic tools aforementioned are developed for western-population [62, 9], being susceptible to evaluation biases and subjective decision-making by

doctors from low and middle-income countries (LMICs), resulting in inaccurate results driven mostly by a lack of training and cultural differences [62]. Furthermore, there are other challenges related to the diagnostic process *per se*. Firstly, an early diagnosis is compromised for borderline and high-functioning individuals due to the high behavioral variance that characterizes ASD [67]. In addition, diagnosis is, also, particularly difficult in adult populations, as symptoms assessed are often masked in adult samples by coping strategies developed over the years, or might have been ameliorated by treatments and interventions [66]. Secondly, ASD is more than four times more diagnosed in boys than in girls [68], in addition to girls being diagnosed later than males, on average, with susceptibility to misdiagnosis attributed to stereotypical gender biases [69]. Lastly, similarities with other conditions, such as ADHD and speech delays contribute to misdiagnoses and delayed diagnoses [62].

To overcome such challenges, in the past few years, there has been an increasing focus on the identification of potential ASD biomarkers to help (i) stratify risk in order to determine which children should be screened, (ii) aid in early diagnosis, (iii) validate the behavioral findings of diagnostic testing, (iv) stratify patients into subgroups, and (v) predict treatment response [70, 61]. Furthermore, they can be clustered into different major types: behavioral, genetic, immune, medical history, metabolic, neuroimaging, neuropsychology, and nutritional [70].

Neuroimaging techniques have been used to study the atypical brain findings that are observed in individuals suffering from ASD [66], such as the altered neuronal connectivity. Subsequently, neuroimaging biomarkers have been substantially studied and show great prospects for the early diagnosis of ASD [71, 66, 70], and even for assisting behavioral diagnoses that have important promises in adult populations [66, 72].

2.1.5 Neuroimaging Techniques and Findings

Neuroimaging comprises powerful methods to visualize and study the neurological underpinnings of autism spectrum disorder (ASD), beyond behavioral observations, shedding light on how the brain of individuals in the autism spectrum may differ from the brain of typical developing (TD) individuals. A multitude of neuroimaging techniques are available and used in an ASD context. Nuclear imaging techniques, such as positron emission tomography (PET) and single-photon emission computerized tomography (SPECT), have been explored to examine metabolic and neurochemical changes in the autistic brain [73, 74]. Structural and functional magnetic resonance imaging have been extensively employed to investigate the neural anatomical and functional basis of ASD. While sMRI allows to map the brain's morphological alterations of ASD [75], fMRI is used to identify ASD-related aberrant brain activity while performing a task or at rest [76, 77]. Furthermore, fiber tracking techniques based on diffusion tensor imaging (DTI) that provide direct information about white matter connections and the integrity of communication pathways, as well as functional connectivity methods assessing inter-regional brain activity synchronization, offer novel insights into the operational dynamics of aberrant brain networks within individuals along the autism spectrum. These techniques hold the potential to identify neural markers and understand the underlying mechanisms at play [49].

Structural MRI studies, resorting to morphometric techniques and DTI, have reported numerous morphological brain alterations associated with ASD, mainly in cortical surface area and thickness, brain lateralization, gray matter volume, and white matter connectivity, particularly in the prefrontal cortex [78], temporal cortex [79], amygdala [80], cerebellum [81], corpus callosum [82], caudate nucleus [83], and cerebrospinal fluid (CSF) [84].

While certain findings have demonstrated consistency across various studies, the landscape of neu-

roimaging investigations pertaining to ASD is also marked by heterogeneity of findings [49]. Regarding sMRI, for instance, there has been a divergence in reported results concerning white matter volumes among individuals with ASD [85]. These disparities can be attributed to numerous factors encompassing demographic variations, the heterogeneity evident in clinical presentations, distinctions in scanner parameters, fluctuations in in-scanner head motion, and the divergent methodologies applied to data analysis [49].

For functional MRI studies, task-based protocols focus on paradigms that target specific behavioral deficits observed in individuals diagnosed with ASD, such as responsiveness to social stimuli, recognition of facial emotions, and reward-related behavior. Conversely, there has been a growing interest within ASD research regarding the expanding popularity of rs-fMRI [86], which have consistently demonstrated altered patterns of connectivity between brain regions that underlie the clinical presentation of individuals with ASD. Aberrant intrinsic connectivity has been identified to impact resting-state networks (RSNs) associated with self-reference, social cognition, decision-making and cognitive control, and sensory integration (default mode network (DMN), salience network (SN), central executive network (CEN), somatomotor network (SMN), respectively), with both under- and over-connectivity reported when compared to TD controls [86, 49, 87]. A growing trend of studies has been reporting long-range under-connectivity and short-range over-connectivity of the DMN [88], which is speculated to be linked to the excess number of neurons found in ASD individuals' brains, hampering the proper functioning of extensive, long-range interactions among different regions of the brain [86]. Note that this finding integrates with the pathophysiology of ASD discussed, as the excess number of neurons may be due to impaired neural migration and insufficient dendritic pruning. Furthermore, PET and SPECT studies support the increase in excitatory signaling and a decrease in inhibitory signaling theory discussed in the pathophysiology subsection (2.1.2) as they have found atypical ratios of excitatory and inhibitory neurotransmitters (namely, serotonin, dopamine, and gamma-aminobutyric acid (GABA)) levels in individuals with ASD, contributing to altered neural signaling patterns [89, 90, 38].

ASD exhibits an unconventional trajectory of brain maturation, that potentially influences autistic symptoms throughout an individual's lifespan [66]. Findings from longitudinal and cross-sectional neuroimaging studies have revealed age-specific alterations in brain anatomy and distinctive neurodevelopmental trajectories in ASD [86, 49]. As an illustration, autism manifests with a notable brain overgrowth during infancy and early childhood, succeeded by a heightened pace of reduction in dimensions, potentially progressing towards degenerative processes from adolescence to adulthood when declines in structural volume are observed [86]. Besides, age-related abnormalities in functional connectivity in the DMN and the "social" brain have also been reported in rs-fMRI studies [49]. In this way, ASD is potentially associated with atypical structural and functional developmental trajectories across the lifespan. While these investigations have indeed delineated distinctions between autistic individuals and those classified as typical developing, their scope has been constrained by relatively small sample sizes. Additionally, a further constraint arises from the predominant utilization of conventional neuroimaging methodologies, which were originally designed to detect brain abnormalities by averaging data across two or more participant cohorts. This approach has the potential to obscure inherent heterogeneity and, consequently, may provide only restricted insights into the presence of neuropathological conditions at the individual subject level [18].

Taking all of these factors into consideration, neuroimaging research in ASD holds the promise of identifying brain markers of the disorder that could aid in its early detection or serve as a complementary diagnostic tool across all age groups. Furthermore, it has also played a pivotal role in uncovering the underlying mechanisms of ASD and, through longitudinal studies, has contributed to a comprehensive

understanding of how the disorder progresses over an individual's lifespan. By pinpointing potential targets, neuroimaging research could significantly contribute to a more precise treatment approach for ASD [49]. Nonetheless, it is crucial to acknowledge and address the mentioned challenges. Advanced techniques, including machine learning (ML) and deep learning (DL) methods that, by extracting complex patterns from neuroimaging data, can help identify specific characteristics, coupled with larger sample sizes could potentially unveil novel insights into atypical brain connectivity [86].

2.2 Functional Magnetic Resonance Imaging

The emergence of functional magnetic resonance imaging (fMRI) dates back to the early nineties. Over the past three decades, this category of imaging techniques has transformed the study of brain function by enabling the investigation of dynamic regional brain activity patterns [91, 92]. Its popularity stems from several factors: widespread accessibility (feasible on a clinical 1.5T scanner), non-invasiveness (eliminating the need for radioisotope injection or other pharmacological agents), relatively low cost, and provision of good spatial resolution, signal reliability, robustness, and reproducibility [92, 93]. fMRI scans rely on the same foundational atomic physics principles as conventional magnetic resonance imaging (MRI). Beyond that, fMRI leverages the blood-oxygen-level-dependent (BOLD) contrast to detect changes in brain activity related to increased oxygen consumption in activated brain regions [91].

Recent advancements in functional neuroimaging have introduced novel methodologies beyond simple activation detection. These methods explore functional interactions between brain regions, facilitating the investigation of hypothesized disconnectivity patterns across various neurological disorders [94]. This section delves into the fundamental physics principles governing fMRI (such as the MRI principles), and particularly the BOLD effect. Subsequently, the resting-state functional magnetic resonance imaging (rs-fMRI) framework is discussed. Finally, it is introduced the notion of brain connectivity, along with the associated computational and analytical tools employed for its exploration.

2.2.1 MR principles

The primary components of an MRI scanner are the main magnet, the radio-frequency (RF) coils, and three gradient coils. The MRI images are generated through the manipulation of the nuclear magnetic resonance (NMR) phenomenon, primarily targeting the hydrogen nuclei present in human tissues. Pioneered by Bloch and Purcell in 1946 [95, 96], NMR arises when atomic nuclei with nonzero spin, typically those containing an odd number of protons, are placed in a strong external magnetic field. These nuclei can then be excited by a specific RF pulse with a characteristic frequency, causing them to resonate and subsequently emit an electromagnetic signal. This emitted signal can then be detected to form an MRI image [97]. Hydrogen atoms are single-proton nuclei in high concentration in the human tissues, being normally targeted. Due to their nonzero spin, they have both angular and magnetic momentum, allowing them to interact with magnetic fields akin to small magnets due to the NMR effect [98].

The main magnet produces a strong and constant external magnetic field (B_0). When subjected to B_0, hydrogen protons' magnetic moments tend to align in one of two orientations relative to B_0: parallel (low-energy state) or anti-parallel (high-energy state). The predominance of protons in the low-energy state results in a net macroscopic magnetization, referred to as longitudinal magnetization M_z in the direction of the main magnetic field (B_0) (i.e. z axis) [98]. As hydrogen protons possess angular momentum, their axes precess around the B_0 axis instead of being perfectly aligned parallel or anti-parallel to it [97]. The frequency of precession (ω_0) is directly proportional to the magnitude of the magnetic field B_0 by a

factor γ (the gyromagnetic ratio, which is an empirical constant unique to each nucleus). This relation is described by the Larmor equation [97, 99]:

$$\omega_0 = \gamma B_0. \tag{2.1}$$

The signal detected in an MRI exam originates from the transverse magnetization (M_{xy}). This magnetization is created by applying a RF pulse - a short pulse lasting microseconds, also known as the B_1 magnetic field - emitted by the RF coils at the Larmor frequency, transversely to B_0. During its application, hydrogen protons are excited, transitioning from the parallel to the anti-parallel orientation, reducing M_z. Eventually, M_z reaches zero, corresponding to a scenario where the number of protons in both orientations is equal, resulting in the cancellation of all protons magnetic moments. Concurrently, the application of the RF pulse causes these protons to precess in phase about B_1 direction, contributing to M_{xy} [98]. The transverse magnetization is a variable magnetic field (precessing in the transversal plane at the Larmor frequency). When a conductive receiver coil, such as the RF coils, is in proximity, an alternating voltage can be induced in it [98, 99, 97].

Upon applying the RF pulse, the hydrogen protons begin to lose phase coherence while returning to a lower energy state, a process known as proton relaxation. Proton relaxation encompasses two simultaneous phenomena: transverse relaxation, where magnetization tends to decrease in its transverse direction, and longitudinal relaxation, where magnetization regrows along its longitudinal direction [98, 97]. Longitudinal relaxation, also termed spin-lattice relaxation, arises from interactions between spins and their surroundings, leading to proton energy loss and a subsequent increase of M_z. This process is characterized by a time constant T1, representing the time required for approximately 63% of M_z to regrow. Conversely, transverse relaxation, or spin-spin relaxation, results from the spin dephasing caused by interactions between neighboring protons' magnetic fields and is characterized by T2. T2 corresponds to the time taken for M_{xy} to decrease to approximately 37% of its initial value [98]. In practice, phase coherence loss is accelerated by inhomogeneities within B_0. Instead of directly observing T2, the composite decay factor T2* is observed, accounting for both internal inhomogeneities of spins and the external inhomogeneities of the B_0 magnetic field [98, 99]. Crucially, T1 and T2 relaxation times vary across tissue types. Therefore, they can be leveraged to generate contrast in MRI images, by employing specific pulse sequences that emphasize either T1 or T2 weighting. The most fundamental contrasts are T1-weighted, T2-weighted, and proton density [100].

Following the 90° RF pulse, an additional pulse can be applied to the signal at a 180° angle relative to the existing magnetization, thus generating a signal echo that is subsequently measured. The interval between the RF excitation (initiated by the 90° pulse) and the ensuing echo is termed the echo time (TE) in a sequence known as spin-echo. To acquire an MRI image, hydrogen protons are excited multiple times via successive 90° RF pulses. The time span between consecutive RF pulses is denoted as the repetition time (TR) [98]. TE and TR are adjusted by the user to achieve the different contrasts. T1-weighted images are produced when TR is on the order of the tissue's T1, with TE chosen to be short (compared to the tissue's T2). T2-weighted images are obtained when TE is comparable to the magnitude of T2, and TR is long (compared to the tissue's T1). Proton-density images are produced with long TR and short TE [99].

For the final three-dimensional image to be reconstructed it is crucial to localize spatially the resulting signal. This localization is achieved through the application of three gradient fields. Each gradient is an additional magnetic field that varies in B_0 according to its spatial location and is induced by one of the gradient coils. These magnetic fields are applied in the three orthogonal directions. As a result, the

measurements of precession frequency can differentiate signals obtained from various spatial positions. The Fourier transform is then used during image reconstruction to translate the spatially encoded signal into a three-dimensional image. [101, 97].

To investigate the functional dynamics of the brain by measuring metabolic activity in an fMRI examination, the MRI scanner can be used with a focus on specific pulse sequences. Currently, gradient recalled echo (GRE) sequences with T2* contrast, particularly fast gradient-echo echo-planar imaging sequences, are commonly favored for fMRI acquisitions [102].

2.2.2 BOLD effect

The blood-oxygen-level-dependent (BOLD) effect is dependent on the hemoglobin magnetic properties, which is used as a natural contrast agent. Specifically, while oxygenated hemoglobin (or oxyhemoglobin) is diamagnetic and is magnetically indistinguishable from brain tissue, intensifying the MR signal, deoxygenated hemoglobin (or deoxyhemoglobin) is highly paramagnetic, which results in local magnetic field distortions dependent on the deoxyhemoglobin concentration that decreases the net MR signal [92, 91, 103].

The neural signaling processes within the brain cells require energy in the form of adenosine triphosphate (ATP), primarily generated through glycolytic oxygenation of glucose within the mitochondria. When changes in neuronal activity in a brain region occur following a change in brain state induced by a stimulus or a task, there is a heightened local demand for energy. This increase in energy requirement escalates the need for oxygen by brain cells, prompting the extraction of oxygen from the blood in the surrounding capillaries and adjacent tissues. Subsequently, as neurons fire and oxygen is consumed, hemoglobin molecules get deoxygenated, with an increase of deoxyhemoglobin levels, and a decrease of oxyhemoglobin levels. In response, the capillaries vasodilate, facilitating increased blood flow to restore local oxygen levels. However, for reasons not fully understood, more oxygen is delivered than necessary to meet the elevated oxygen requirement. Consequently, a situation develops where there are elevated levels of oxyhemoglobin and reduced levels of deoxyhemoglobin [92].

The processes described where increases in brain activity lead to increased blood flow and oxygenation in the corresponding brain regions, resulting in measurable changes in MRI signal characterize the BOLD response. In this way, the scanners are not measuring neural activity directly, but a downstream consequence of neuron firing - the MR signal generated by the triggered hemodynamic response (i.e. local changes in neuronal activity lead to changes in blood flow and oxygenation to meet the metabolic demands of active neurons) and the magnetic properties of deoxyhemoglobin [91, 104]. In Figure 2.4 it is illustrated the BOLD response for an instantaneous stimulus. In the first stage, the MR signal decreases due to factors such as the large concentration of paramagnetic material owning to the conversion of oxyhemoglobin in deoxyhemoglobin and the transient increase in blood volume without a change in the deoxyhemoglobin concentration following capillary vasodilation. The second stage includes a rise in the MR signal as a result of the overcompensation in blood oxygen supply facilitated by heightened cerebral oxygenated blood flow. This stage raises the levels of diamagnetic material and dilutes the deoxyhemoglobin concentration, thereby increasing the MR signal, which constitutes the main signal measured during fMRI acquisition [105, 106]. The third stage represents a signal decrease before returning to baseline. The precise cause for the undershoot observed remains subject to debate. Proposed explanations include prolonged neuronal activity following the initial stimulus, which leads to a sustained demand for oxygen. Additionally, hypotheses suggest a prolonged increase in cerebral blood volume or a decrease in cerebral blood flow as contributing factors [106].

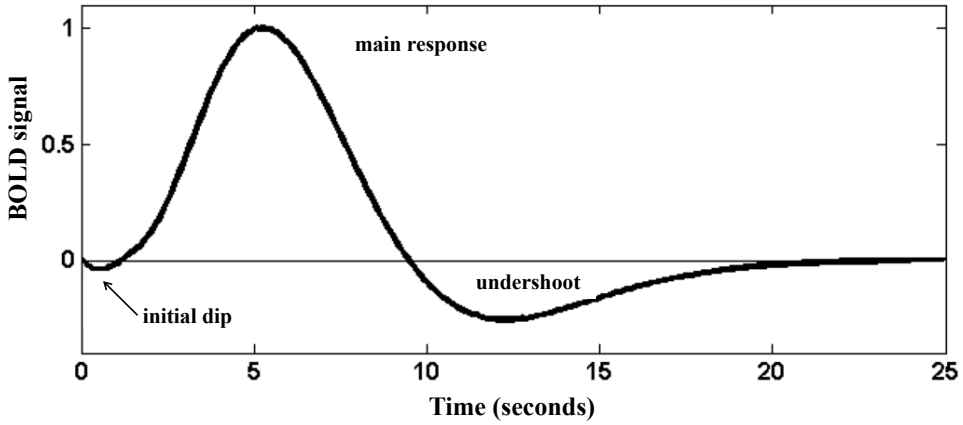


Figure 2.4: Representation of a typical blood-oxygen-level-dependent (BOLD) response to a brief stimulus at time zero. After a possible initial dip, associated with the initial uptake of oxygen before substantial hemodynamic alterations, the primary BOLD response peaks due to the predominant influence of blood flow. This is followed by a post-stimulus undershoot, frequently of significant duration, before returning to baseline levels (from [107]).

2.2.3 Resting-State fMRI

fMRI studies can be categorized into two main types based on their objectives: task-based fMRI and resting-state fMRI. Task-based fMRI investigates the spontaneous brain activity modulations in the BOLD signal evoked by performing a specific cognitive task (e.g. finger-tapping, naming, memorization), with the aim of associating brain regions with specific tasks. Rs-fMRI gained prominence following the seminal work by Biswal and colleagues, who found the presence of coherent low-frequency spontaneous fluctuations in fMRI signal at resting-state [94, 103]. Therefore, rs-fMRI focus on measuring these spontaneous and low-frequency neuronal oscillations [94], being conducted without the performance of specific cognitive tasks. During these scans, participants are instructed to remain awake and may be asked to keep their eyes closed, keep their eyes open, or focus on a crosshair, while refraining from engaging in any cognitive activity [103]. The low-frequency spontaneous fluctuations measured reflect the activity of intrinsic connectivity networks also known as resting-state networks (RSNs)[94]. A RSN is composed by several anatomically separated brain regions that demonstrate synchronous BOLD signals fluctuations during rest, being described as functionally connected [94, 108]. Various RSNs have been consistently delineated and named in the literature, including the most prominent network, the default mode network (DMN), and others, such as the dorsal attention network and the ventral attention network, which are related to attention [109, 103, 108]. Their significance lies in the correspondence between their topographical organization and activation maps observed in task-based fMRI paradigms. These networks encompass critical brain regions often termed as "eloquent" areas within the somatosensory, language, and visual networks. This knowledge holds significant clinical value, particularly for neurosurgeons performing pre-surgical planning to identify and preserve these crucial areas [108]. Furthermore, investigating the dynamic behavior of RSNs can unveil potential variations between healthy and diseased states, potentially serving as biomarkers for neurological and psychiatric disorders, including Alzheimer's disease (AD), epilepsy, and ASD [103].

It is worth noting that depending on the granularity of how a network is defined, there is no single number of RSNs but at the highest level, the brain can be thought to consist of seven networks, that have been consistently identified in the literature, resorting to various brain functional connectivity analysis methods. Particularly, the seven RSNs are: the default mode, visual, somatomotor (or sensorimotor),

limbic, dorsal attention, ventral attention (or salience), and frontoparietal control (or central executive) networks [109, 110] (see Figure 2.5).

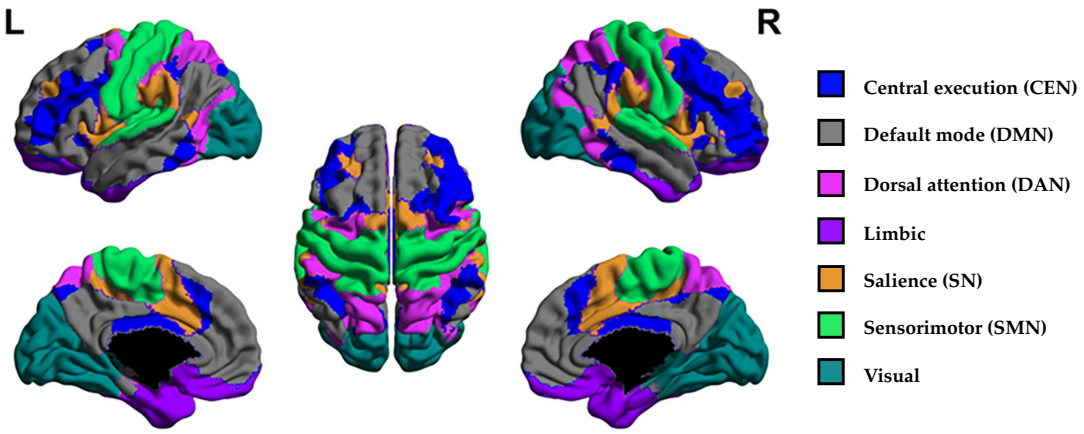


Figure 2.5: Large-scale RSNs (adapted from [110]).

These RSNs can be associated with specific anatomical and functional characteristics. For instance, the default mode network (DMN) includes the posterior cingulate cortex, medial prefrontal cortex, and lateral parietal cortex [103]. The DMN is typically active during rest periods when individuals are not engaged in specific cognitive tasks but deactivates when tasks are performed [103, 111]. Therefore, it is linked to mind wandering and social cognition, including introspection, emotional processing, and thinking about others' mental states [103]. The visual network comprises the calcarine sulcus, lingual gyrus, inferior area of the precuneus, and lateral geniculate nucleus of the thalamus [103]. It responds primarily to visual stimuli, such as motion analysis and pattern recognition [111]. The somatomotor network (SMN) involves the primary and secondary somatosensory cortices, premotor cortex, and supplementary motor area [111], contributing to motor tasks [112]. The limbic network includes regions like the orbitofrontal cortex, hippocampus, hypothalamus, amygdala, and thalamus, playing roles in memory, learning, decision-making, and emotion regulation [111, 113]. The dorsal attention network (DAN) encompasses the intraparietal sulcus and frontal eye fields, contributing to spatial attention, saccade planning, and visual working memory [110, 114]. The central executive network (CEN) is composed by the dorsolateral prefrontal cortex and lateral posterior parietal cortex. It supports active tasks, working memory, and rule-based problem-solving. Its activity is relatively stronger during higher-order cognitive efforts and is negatively correlated with DMN activity [115]. The salience network (SN) includes regions such as the dorsolateral cingulate cortex, bilateral insula, presupplementary motor area, amygdala, and other subcortical areas. The SN moderates the switching between internal and external processing, mediated by the DMN, and task-related directed attention to external stimuli, maintained by the CEN. Additionally, the SN influences the processing of reward, motivation, emotion, and pain [115, 110].

Rs-fMRI research has yielded intriguing insights into alterations in RSN between healthy and diseased brains. For instance, in ASD, changes in neuronal activation within regions of the DMN and CEN have been found to correlate with the severity of ASD symptoms [115]. The subsequent subsection delves into the methodologies employed to estimate and delineate brain functional connectivity, aiming to explore discrepancies in activations within these RSNs.

2.2.4 Brain Connectivity

The brain relies on both its anatomical and functional connections to process and integrate the information necessary for performing tasks, such as motor control or visual perception [116]. While anatomical and functional connections are inherent components of the broader concept of "Brain Connectivity", it is imperative to distinguish functional connectivity (FC) from anatomical (or structural) connectivity, as well as effective connectivity. FC describes the temporal interdependence between the neuronal activations of anatomically distant brain regions, while effective connectivity elucidates the directionality of these relationships. On the other hand, anatomical connectivity describes brain regions physically connected by synaptic contacts or white matter fiber tracts [117, 116, 32].

Connectivity analyses investigate variations in connectivity type and strength between brain regions, conducted either at the level of specific neural systems or across the entire brain. In whole-brain connectivity analyses, each type of brain connectivity is typically represented in a connectivity matrix, where each row and column correspond to different brain regions, and each matrix element denotes the strength of connectivity between every pair of considered brain regions [116, 118, 119]. In brain connectivity analyses, connectivity matrices are often undirected, indicating the presence and strength of a connection but not its direction, although models employing effective connectivity can ascertain directionality [119]. Brain activity measurements for connectivity analysis can be obtained from multiple sensors or spatial locations, using techniques such as electroencephalography (EEG), magnetoencephalography (MEG), calcium imaging, or fMRI data. A significant portion of research has been focused on FC analysis, derived from fMRI data [117, 120, 121]. FC is particularly vulnerable to neuropathologies. Subsequently, FC analyses aim to identify relevant disruptions in FC that could serve as potential biomarkers for neuropsychiatric disorders such as ASD. These clinical biomarkers hold promise for aiding disease diagnosis, staging, risk prognosis assessment, and prediction and monitoring of clinical response to interventions [118].

To derive FC from rs-fMRI data, preprocessing steps involving conventional and specific methods for FC analysis must be applied to the acquired rs-fMRI data. The recommended preprocessing steps for acquired rs-fMRI data, as suggested by the scientific community, encompass multiple procedures. These include slice time correction, motion and distortion correction, temporal filtering, spatial smoothing, physiological noise correction, functional-structure co-registration, and spatial normalization to a common template, such as the Montreal Neurological Institute 152 (MNI152) [122, 72]. Additionally, quality assurance should be performed to detect and rectify any issues that may corrupt the acquired images. These issues could arise during data acquisition or preprocessing, from scanner problems such as extreme scanner noise or signal drift, or from excessive subjects' motion. Quality assurance should be conducted continuously, starting immediately after scanning, and should be integrated into the preprocessing pipeline using visual inspection and simple tests to ensure data quality before the next analysis step [122].

Following the preprocessing procedures, various postprocessing methods can be employed to perform FC analysis. In rs-fMRI studies, functional connectivity analysis begins with node definition. As FC can be evaluated with both intra- and inter-network FC metrics, nodes can delineate brain regions or RSNs [111]. Node definition can be accomplished through atlas-based or data-driven based approaches. Atlas-based approaches resort to established brain templates from the literature. These include region of interest (ROI)-based atlases, where each node represents a specific brain region, and network-based brain atlases, where each node corresponds to an RSN. The Harvard-Oxford (HO) atlas [123] and the Automated Anatomical Labeling (AAL) atlas [124] are two examples of ROI-based atlases. An example

of a network-based atlas is the functional atlas defined by Yeo et al. [109]. Using established atlases ensures that the defined nodes align with existing literature, facilitating direct comparisons of results across different works [111]. Conversely, node definition can be performed using data-driven methods, such as group-independent component analysis (ICA) [125].

Independent component analysis is a multivariate statistical method aimed at decomposing a time-series signal into a set of mutually independent and temporally associated time courses. ICA identifies groups of voxels or regions with co-activated signals, effectively grouping them into distinct functional networks. Each network, or component, extracted by ICA represents a set of brain regions that exhibit FC with each other [111, 126]. Given the well-established notion of the brain's organization into functionally discrete networks, known as RSNs, ICA can be employed to spatially identify distinct RSNs [127]. Additionally, alongside the identification of components from the data, their number is also estimated, minimizing researcher bias in defining RSNs [126]. Therefore, offers the advantage of providing the best representation of structured components present in the dataset.

After identifying the nodes in analysis, the BOLD signal time-series is extracted for each defined node. The FC analysis can focus on analyzing the functional activity of one specific brain ROI in relation to other regions throughout the brain, in a type of analysis called seed-based connectivity analysis. **Seed-based connectivity** analysis involves delineating a ROI, referred to as the seed region. The FC of each voxel in the brain is computed in relation to the seed region. Once the seed region is defined, the BOLD time-series are extracted for the seed region and each voxel across the brain. Seed-based correlation analysis (SCA) stands out as one of the most prevalent FC analysis method [128, 111]. In this, the correlation between the time-series of the seed region and those of each voxel is computed using the Fisher-transformed bivariate correlation coefficient. The resulting correlation values are then mapped onto a connectivity map, with each voxel representing the FC measures [103]. Despite its widespread use, SCA poses challenges for examining whole-brain FC due to its reliance on the selected seed region. Moreover, the necessity of choosing a seed region presupposes prior knowledge of the system [103].

Alternatively, after parcellating the brain into multiple brain ROIs or RSNs, brain functional activity can be studied as a whole by computing a FC matrix [125, 111]. For each subject, a FC matrix is computed, which describes all possible pairwise functional connections between the defined nodes [129]. Various connectivity metrics, including Pearson's correlation coefficient, partial correlation, mutual information, and coherence, can be used to define connections between nodes [125]. Pearson's correlation coefficient, commonly chosen for its sensitivity to both direct and indirect FC in the brain, is frequently employed to estimate pairwise similarity between BOLD time-series, leading to the generation of a FC matrix [129].

Graph theory analysis can be applied to extract high-level information regarding network functionality and topology. A graph can be computed from the FC matrix, typically involving thresholding and binarization of the FC matrix. Usually, graph metrics are determined. These include node degree, which helps identify crucial brain regions that play an important role in network communication (hubs), path length, and modularity [129, 111]. Alternatively, weighted analyses are also attractive, considering that brain network dynamics define an intrinsically weighted system, where functional interaction varies between different pairs of brain regions [129].

Following preprocessing of rs-fMRI data from all subjects, statistical analysis is typically employed to investigate inter-individual or inter-group differences. There has been a growing trend in employing DL techniques to extract pertinent features from intricate fMRI datasets. When combined with brain FC, these methods hold promise in identifying specific functional dysconnections in individuals with ASD, that could potentially serve as biomarkers to differentiate between ASD and typical developing (TD)

individuals [72].

2.3 Deep Learning on Graphs

Artificial intelligence (AI) systems capable of automated learning from data that could automatically improve their knowledge and performance with experience emerged as significant advancements in a division of AI called machine learning (ML) [130]. However, building a ML system required careful feature engineering and considerable domain expertise for extracting suitable internal representation for the ML system [131]. In the last decade, deep learning (DL), a sub-field of ML, gained immense popularity. By addressing efficiently the learning challenges, DL techniques are capable of performing representation learning, not only discerning the mapping from representation to output but also uncovering the inherent representation themselves. Learned representations frequently outperform the manually crafted ones in a majority of tasks, facilitating rapid adaptation to novel tasks with minimal human intervention within the deep learning framework [132, 133]. The breakthroughs of DL are spread across many fields such as Computer Vision, Natural Language Processing, and Biomedical Imaging [133], in which the data have an underlying Euclidean structure (e.g. images, text, and videos). In numerous scientific fields, many real-world objects and problems yield data that naturally conform or are best expressed along non-Euclidean structures, such as graphs. These graph structures are suitable for representing the dependencies and inter-relationships between various entities [134, 135]. This motivated a growing interest in generalizing DL advances to graph-based tasks, giving rise to the rapidly expanding field of graph neural networks (GNNs) [136]. Nevertheless, graph data can be irregular, with variable counts of unordered nodes, and complex topological structures derived from the possible different number of neighbors [137]. This complexity inherent to graph data imposes significant challenges on standard DL algorithms that served as motivations for the development of the GNN architecture.

This section starts with an introduction to fundamental concepts in DL and provides foundational definitions of key graph theory concepts. This serves as a groundwork for facilitating a comprehensive understanding of the forthcoming examination of the GNN framework. In the final part of this section, current methods on the explainability of GNNs will be examined.

2.3.1 Deep Learning Basics

Deep learning is based on neural networks (NNs). Commonly a NN consists of several neural units organized in many layers working together to learn representations of data with multiple levels of abstraction [138]. A NN is composed of an input and output layer, with, or without, layers in between - called hidden layers. The neural units, or neurons, within each layer, are interconnected with neurons in adjacent layers [139]. Each connection has an associated weight ($w_i, i = 1, 2, ..., m$). Each neural unit computes the weighted sum of the inputs ($x_i, i = 1, 2, ..., m$) by first multiplying each input by its respective weight, then summing those values and adding a bias term (b) [140]. In a simple NN, with only an input layer and an output layer composed of one neuron, the resultant value (z) is the argument for the activation function ϕ that generates the output signal of the model:

$$y = \phi(z) = \phi((\sum_{k=1}^{m} w_k x_k) + b) \tag{2.2}$$

The activation function performs a non-linear transformation of z and can be as simple as a step function, or a more complex one, such as a sigmoid. Some of the most used include sigmoid, hyper-

bolic tangent, or rectified linear unit (ReLU)[139]. These functions imply that the neuron is activated in response to specific weighted sum values. The inclusion of this non-linearity is crucial for the network to learn and represent complex, non-linear relationships within the data (expressive power). Otherwise, merely combining linear operations would constrain the model's expressive capacity to linear processes, since a succession of linear operations ultimately collapses into a singular linear operation. The fundamental concept underlying this is that by composing these simple yet non-linear modules it's possible to learn highly complex functions [131]. A multilayer perceptron (MLP) is a feedforward NN with at least one hidden layer. In this, each neuron operates similarly to the neuron model exposed [131]. Figure 2.6 illustrates a multilayer perceptron (MLP) with one hidden layer and output determined by the following expression:

$$y = \phi_2 \left(\sum_{j=1}^{4} w_j \phi_1 \left(\sum_{k=1}^{3} w_{jk} x_k + b_j \right) + b \right) \tag{2.3}$$

where x_k represents the value of the k-th input neuron, with $k = 1,2,3$, w_{jk} represents the weights associated with the connection from the k-th input neuron to the j-th neuron in the hidden layer (with $j = 1,2,3,4$), b_j is the bias term for the j-th hidden node, ϕ_1 is the activation function applied to the hidden layer, w_j is the weight associated with the connection from the j-th neuron in the hidden layer to the output unit, b is the bias term for the output unit, and ϕ_2 is the activation function applied to the output layer.

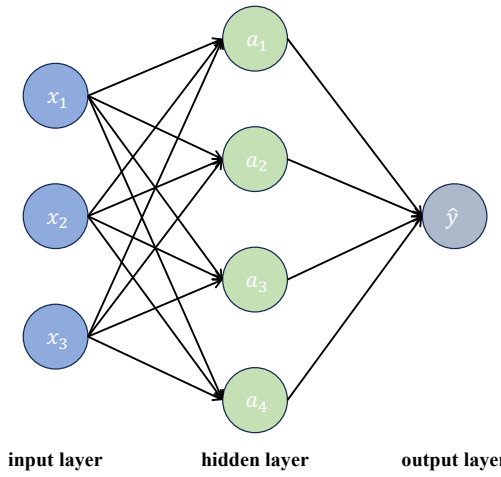


Figure 2.6: Example of an MLP network with one hidden layer. $a_i, i = 1,2,3,4$ is the activation output of each hidden neuron, and $\hat{y}$ is the output of the model that could also be written as $a^{[2]}$, with the activations of the hidden units being written as $a_i^{[1]}$. The superscripts indicate the respective layer.

Overall, a NN establishes a connection between the input values it receives and the desired behavior at the output layer [132]. This linkage is established through a continuous process referred to as training, wherein the network is repeatedly exposed to data examples [131]. The training of a NN includes forward and backward propagation while minimizing a loss function in the training data. In the forward propagation step, the initial information (inputs) traverses through the network, passing through the successive hidden layers via the weighted connections, culminating in the generation of the output, represented by $\hat{y}$. The prediction of the model is compared to the ground truth (y) using the loss function to measure the performance of the model. The backpropagation step allows the information from the cost to flow backward and compute the gradient of the loss with respect to the weights of the connections. By using gradient descent algorithms the weights (and biases) are updated based on these gradients in order to minimize the loss function [132]. For binary classification tasks, binary cross-entropy is the

common choice for the loss function, categorial cross-entropy for multi-class classification tasks, and mean squared error for regression tasks.

The selection of hyperparameters significantly influences the performance of the model. Key hyperparameters that are universally relevant across various NN architectures include the number of epochs and the learning rate. An epoch denotes a complete iteration wherein the entire training set is forwarded and backward-propagated through the NN once. Typically set to a large value, the number of epochs allows the NN to converge towards a minimum loss value, balancing the achievement of optimal convergence with a reasonable training time. The learning rate is a crucial hyperparameter to tune given its significant impact on model performance. The learning rate corresponds to the step size for updates to the weights and biases in the gradient direction in each iteration. Among the algorithms designed to compute an optimal learning rate, the adaptive moment estimation (ADAM) optimizer [141] emerges as a potent method, known for its relative ease of configuration and computational efficiency. It computes individual adaptive learning rates for distinct parameters, derived from estimates of the first and second moments of the gradients.

A well-trained NN should exhibit the ability to generalize its learned knowledge to new, unseen data. During the training process, the model is presented with a training set, with the primary goal of minimizing the training error. This error is often quantified by the loss function. A generalization error also called a test error, is measured on a separate test set drawn from the same underlying probability distribution. In addition to achieving a low training error, the desirable outcome is a minimal difference between the training error and the test error [132]. The model might struggle to fit the data if it is overly simplistic, leading to an increase in assumptions about the data (high bias). Consequently, the model fails to attain a satisfactorily low training error value and is in a situation known as underfitting. Overfitting stands also as a key challenge in ML. In this situation, the difference between the training and test errors may be too large (high variance). Essentially, the model might excessively fit the training data while struggling to generalize well to the test set data [132]. Conventionally, to address underfitting involves constructing a more complex DL model (introducing more layers and/or hidden units), to capture discriminative features from the training data. On the other hand, mitigating overfitting often involves acquiring more data, which subsequently provides the model with a more comprehensive and diverse representation of the underlying distribution, promoting generalization. In addition, various regularization techniques have been developed to address overfitting. These techniques are designed to diminish test error without significantly impacting training error.

Traditional regularization strategies involve the addition of a penalty term to the loss function. Among these methods, one extensively employed approach is L2 regularization, which adds the $\Omega = \frac{1}{2}||\omega||^2$ term to the loss function. This penalty, commonly referred to as weight decay, identifies features with low covariance with the output target and assigns them smaller weights compared to features with higher covariance. By effectively reducing the influence of less relevant features, L2 regularization helps prevent the model from fitting noise in the training data too closely, thus mitigating overfitting [132]. Another prominent regularization method is the dropout. This involves randomly dropping neurons from the NN during training, effectively preventing the network from becoming overly reliant on specific hidden units. This mechanism encourages the NN to learn more robust feature representations, hindering its tendency to overfit the training data. Lastly, early stopping stands as a straightforward yet widely adopted regularization technique in DL. This technique requires a validation set. The primary objective is to minimize generalization error by reducing the validation set error. During the training of deep NNs, the model often exhibits the capacity to overfit the training data, resulting in an observed reduction in the training error but with an increase of the generalization error (after decreasing for several iterations).

Early stopping serves to avoid overfitting by reverting to the parameter setting at the point where the validation set error is at its minimum (ideally reflecting the lowest generalization error). To achieve this, the validation set error is computed at each iteration and the training stops as soon as the error on the validation set increases. While this represents a naive approach, more complex conditions may be employed to address concerns, such as preventing premature cessation of NN training before the model adequately fits the training data [132].

2.3.2 Graph

Before delving into an exploration of deep learning on graphs, it is essential to provide a formal description of the concept of "graph data". A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined by a set $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ of $N = |\mathcal{V}|$ nodes and a set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ between nodes in $\mathcal{V}$, where an edge going from node $v_i \in \mathcal{V}$ to node $v_j \in \mathcal{V}$ is denoted by $e_{ij} = (v_i, v_j) \in \mathcal{E}$ and indicates the existence of a relationship between the two nodes. The neighborhood of a node $v_i \in \mathcal{V}$ is defined as $\mathcal{N}(v_i) = \{v_j \in \mathcal{V} | (v_i, v_j) \in \mathcal{E}\}$. A convenient way to represent graph connectivity is by its adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ that represents the existence of a relation between any two nodes. In the case of a binary or unweighted graph, if any two nodes v_i and v_j are connected by an edge e_{ij} then the respective entry of the matrix is $\mathbf{A}_{ij} = 1$, and 0 otherwise. Some graphs have weighted edges, where each entry of the matrix $\mathbf{A}$ corresponds to the weight of the connection between any two nodes, i.e. the entries in $\mathbf{A}$ are arbitrary real-values rather than $\{0, 1\}$. A graph can also be categorized as undirected if the edges are undirected, meaning that any two nodes have a link with no directional information, and directed if the edges between nodes have a direction. For undirected graphs, the adjacency matrix is symmetric, and for directed ones may not be [142]. In many cases, each node $v \in \mathcal{V}$ have attached a feature vector, $\mathbf{x}_v \in \mathbb{R}^m$. The set of all node features is most often represented using a real-valued matrix $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times m}$ obtained by stacking all the node features as $\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\right]^\mathsf{T}$ [143].

In some real-world scenarios, the data do not directly conform to a graph structure. In this way, there are two possible scenarios: structural and non-structural scenarios. In the former, the graph structure is explicitly defined, as observed in applications involving chemical molecules. These instances typically feature small graphs wherein nodes represent atoms, and edges symbolize chemical bonds. Conversely, in non-structural scenarios, like real-world applications in Computer Vision, the graph must be constructed from the available data, either through manual or automatic means [144].

2.3.3 Graph Representation Learning

Upon obtaining the graph expression of the input data, the subsequent stage involves the application of graph representation learning techniques. Graph representation learning (GRL) seeks to derive dense and continuous low-dimensional vector representations for nodes. This process serves to mitigate noise or redundant information while preserving the intrinsic structural information within the graph. Within the learned representation space, the relationships between nodes, initially represented by edges or other high-order topological measures in graphs, are captured by the distances between nodes in the vector space, effectively encoding the structural characteristics of a node into its representation vector [145]. This process must fulfill two essential criteria. Firstly, it should enable the reconstruction of the original graph from the learned representation space. Secondly, the graph representation derived should robustly support graph inference tasks, including the inference of node labels [145]. Overall, there are different GRL techniques that can produce a latent representation from the original graph. One of them and more currently studied are graph neural network (GNN)s (for a detailed summary on traditional and modern

graph embedding methods refer to [142]). The learned representation can then be fed to downstream tasks, such as node classification, clustering, and link prediction [146]. In this work, the goal is to predict if a subject has ASD or not, in a node classification task.

2.3.4 Node classification

Node classification is a variation of supervised learning that deviates from the independently and identically distributed (i.i.d.) assumption prevalent in supervised learning scenarios. In supervised learning each data point is statistically independent from all others, and typically, it is assumed that the data points come from identical distributions to ensure generalizability of the developed model to new data points. However, in node classification, a set of interconnected nodes is modeled, challenging the common i.i.d. assumption [142]. Given a single graph where some nodes are labeled while others remain unlabeled, the objective is for node classification models to learn a predictive model that effectively identifies the class labels for the unlabeled nodes. Node classification is frequently referred to as both supervised and semi-supervised learning. The factor that mostly affects this nomenclature is how different nodes are used during model training [142]. Two settings that differ in the use of the test nodes in embedding operations during training are possible (see Figure 2.7):

- **Transductive learning** computes the node embeddings in the latent space for the entire graph, including both labeled and unlabeled nodes. However, the training is only performed using the labeled nodes, i.e., the unlabeled nodes are not used to compute the loss and update the node classification model parameters. At test time, the model is evaluated in the unlabeled nodes. In summary, the features of every node are known during training, but the labels of the test nodes are not. In this way, this setting is also referred to as semi-supervised learning. If there are new unseen nodes to classify, the model in this setting has to be re-trained.

- **Inductive learning** computes the node embeddings in the latent space and the loss function only considering the training nodes, meaning that the test nodes - and all their edges - are completely unobserved during model training. A successful inductive model should generalize to new unseen nodes.

Transductive learning tries to leverage information from the labeled nodes, and the graph structure, particularly the similarity between connected node pairs, to learn a specific function for the problem at hand. This is in contrast to inductive learning, which strives to develop a general function to classify future instances, from specific examples [147].

2.3.5 Graph Neural Networks

Graph neural networks (GNNs) are an extension of deep learning-based methods to graph-structured data. There is a general GNN framework to build GNNs models in which a form of neural message passing is performed. In this mechanism, vector messages are exchanged between nodes and updated using neural networks [142].

The neural message passing framework in GNNs iteratively updates each node representation by aggregating the information of itself and its neighbors, generating node representations that leverage the node feature information, and the graph structure [148]. At each message passing iteration/layer k, the hidden representation $\mathbf{h}_u^{(k-1)}$ of each node $u \in \mathcal{V}$ is updated. In the process, each node u and those within its neighborhood undergo a transformation of their hidden representations, achieved by the MESSAGE

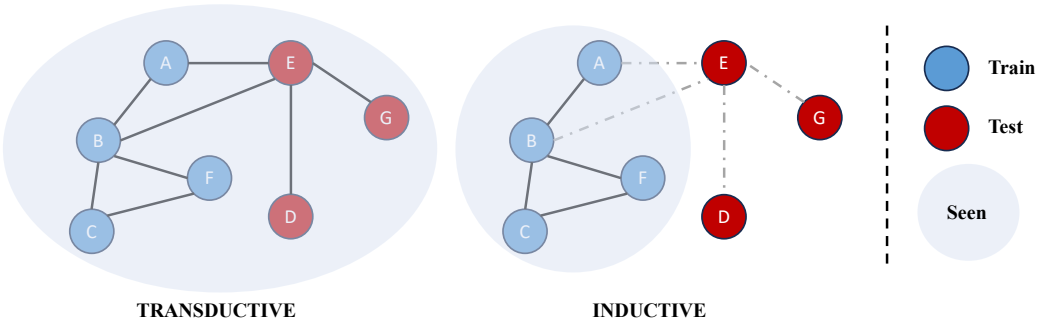


Figure 2.7: Graph split for transductive and inductive learning. The figure illustrates how for transductive learning both train and test nodes are used to compute node embeddings during training (but not to compute the loss function), conversely to what happens in inductive learning, where the test nodes are completely unseen during training. The train nodes are labeled and the test nodes are unlabeled. Note that the graph could be split into train, validation, and test sets, with the validation set behaving identically to the test set in each situation.

function that generates a message $\mathbf{m}_v^{(k)}$ for each node. Then, the set of messages from the neighborhood nodes of a given node u is aggregated via the AGGREGATE function that outputs $\mathbf{a}_u^{(k)}$. Lastly, the UPDATE function yields the updated embedding $\mathbf{h}_v^{(k)}$ by combining $\mathbf{a}_u^{(k)}$ with the generated message $\mathbf{m}_u^{(k)}$ (see Figure 2.8) [148]. This process is expressed as follows:

$$\mathbf{m}_v^{(k)} = \text{MESSAGE}^{(k)}\left(\mathbf{h}_v^{(k-1)}, \forall v \in \{\mathcal{N}(u) \cup u\}\right) \tag{2.4}$$

$$\mathbf{a}_u^{(k)} = \text{AGGREGATE}^{(k)}\left(\mathbf{m}_v^{(k)}, \forall v \in \mathcal{N}(u)\right) \tag{2.5}$$

$$\mathbf{h}_u^{(k)} = \text{UPDATE}^{(k)}\left(\mathbf{m}_u^{(k)}, \mathbf{a}_u^{(k)}\right), \forall u \in \mathcal{V}, \tag{2.6}$$

where MESSAGE, AGGREGATE, and UPDATE are arbitrary differentiable functions (i.e. neural networks). The initial embeddings for all nodes are set to the input features, i.e. $\mathbf{h}_u^{(0)} = \mathbf{x}_u, u \in \mathcal{V}$.

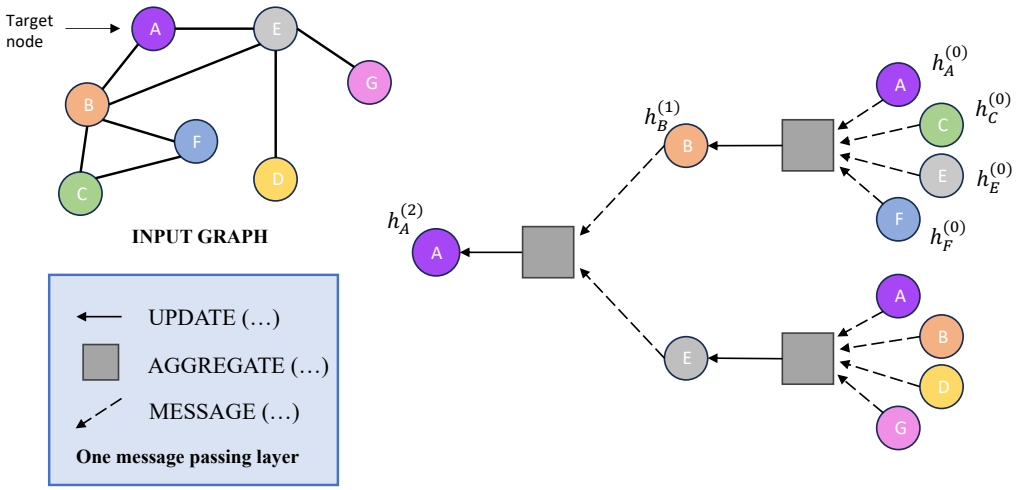


Figure 2.8: Computational flow of the message passing mechanism in a GNN. The figure illustrates how a single node aggregates messages from its local neighborhood, leading to a recursive propagation process, and forming a tree-like structure. This depiction demonstrates the cascading aggregation process of a two-layer version of a message-passing model.

Thus, with each successive iteration, individual nodes gather information from their immediate local surroundings, and as these iterations continue, the embeddings of each node progressively incorporate a richer and more comprehensive set of information from distant regions within the graph.

To translate the abstract GNN framework given into an implementation, concrete instantiations to

the MESSAGE, AGGREGATE, and UPDATE functions must be provided. The basic GNN framework resembles that of a standard MLP. This similarity is rooted in the utilization of linear operations, succeeded by a single element-wise non-linearity. Specifically, this process entails the summation of incoming messages from neighbors, followed by the integration of neighborhood information with the node's prior embedding through a linear combination, culminating in the application of an element-wise non-linearity. The message-passing function in the basic GNN can be formally defined as follows:

$$\mathbf{h}_u^{(k)} = \sigma\left(\mathbf{W}_u^{(k)}\mathbf{h}_u^{(k-1)} + \mathbf{W}_{\mathcal{N}(u)}^{(k)} \sum_{v\in\mathcal{N}(u)} \mathbf{h}_v^{(k-1)} + \mathbf{b}^{(k)}\right) \tag{2.7}$$

where $\mathbf{W}_u^{(k)}$, $\mathbf{W}_{\mathcal{N}(u)}^{(k)}$ are learnable weight matrices for node u and its neighboring nodes $\mathcal{N}(u)$, respectively, σ is an element-wise non-linearity (i.e. activation function), and $\mathbf{b}^{(k)}$ is the associated bias term, frequently omitted for notation simplicity.

The basic GNN given by equation 2.7 can achieve strong performance. However, can be generalized and improved upon in several ways. GNN variants mostly differ in how each node aggregates and updates the representations of its neighbors with its own [149, 150].

2.3.5.1 Graph Convolutional Networks

Convolutional neural networks (CNNs) have demonstrated considerable success in addressing a multitude of problems, notably in the domain of image classification, where the data is inherently structured in a grid-like format. These networks efficiently leverage their localized spatial features, by applying them to all the input positions for parameter learning [151]. With the desire to leverage the advantages of convolutions in the graph domain, numerous works have tried to generalize convolutions to graph-structured data. One of the most popular graph neural network architectures - the graph convolutional network (GCN) - was proposed by Kipf and Welling (2017) [152]. This GCN architecture defines the message-passing function as

$$\mathbf{h}_u^{(k)} = \sigma\left(\mathbf{W}^{(k)} \sum_{v\in\mathcal{N}(u)\cup\{u\}} \frac{\mathbf{h}_v^{(k-1)}}{\sqrt{|\mathcal{N}(u)||\mathcal{N}(v)|}}\right). \tag{2.8}$$

The information is aggregated from 1-hop neighborhoods for every node by the sum function. However, in contrast to the basic GNN, the GCN performs a symmetric normalization before aggregating each message. By performing this normalization the GCN can avoid unstable and weak performance, particularly in tasks where node features hold more significance than structural information. This distinction arises from the inherent nature of graphs, wherein nodes typically do not possess a fixed number of neighbors, unlike the pixels in images. For instance, a node with a large number of neighbors (and therefore connections, which in graph theory is referred to as node degree) would generate an embedding much larger than a node with a low degree. To facilitate effective comparison between nodes with varying numbers of neighbors, the embeddings must be in the same range, which can be achieved by symmetric normalization. Furthermore, note that the update function in this network is implicitly determined by the aggregation function [142].

Relation with Spectral Graph Convolutions. Convolution GNNs can be categorized into two categories: spatial-based and spectral-based. The GCN outlined by equation 2.8 is derived as a simplification from the spectral-based graph convolution network proposed in Defferrard *et al.* (2016) [153]. The orig-

inal spectral approach was initially detailed in Bruna *et al.* (2014) [154] that formulated the convolution operation within the Fourier domain by performing the eigendecomposition of the graph Laplacian (i.e. factorization of the Laplacian matrix into a canonical form). However, this method entailed computationally intensive operations and resulted in non-spatially localized filters. Defferrard *et al.* (2016) [153] proposed an approximation technique employing Chebyshev expansion of the graph Laplacian in a layer currently referred to as ChebConv. This method eliminates the need to compute Laplacian eigenvectors and produces spatially localized filters. Kipf and Welling (2017) [152] considered the assumption that closely related nodes share labels and constrained the filters to operate solely within a 1-hop neighborhood around each node, proposing the GCN. This network became one of the most popular GNN architectures. By stacking very simple graph convolutional layers (outlined by equation 2.8) it is possible to build very powerful models, that conversely to previous works, operate directly on graph's adjacency matrix [142].

Despite their success in node classification tasks, spectral approaches have a fundamental limitation: the filters learned depend on the Laplacian of the graph and are inherently tied to the specific graph structure. This dependence on the graph topology restricts their generalization to unseen data with varying graph structures, confining its applicability to transductive tasks.

Nevertheless, spatial-based approaches directly formulate convolutions on the graph, operating in sets of spatially close neighbors, being able to generalize to new data [151]. The graph attention network (GAT) introduced by Veličković *et al.* (2018) [151] is an example of a spatial-based convolutional graph network, which is directly applicable to node classification tasks in inductive learning.

2.3.5.2 Graph Attention Network

The GAT integrates attention mechanisms to learn the significance of each neighbor to a node. Subsequently, this learned importance is used to weigh each neighbor's influence during the aggregation step [142]. The hidden representation of each node is given by:

$$\mathbf{h}_u^{(k)} = \sigma\left(\sum_{v \in \mathcal{N}(u)} \alpha_{uv} \mathbf{W}^{(k)} \mathbf{h}_v^{(k-1)}\right) \tag{2.9}$$

$$\alpha_{uv} = \frac{\exp\left(LeakyReLU\left(\mathbf{a}^{\mathrm{T}}[\mathbf{W}\mathbf{h}_u || \mathbf{W}\mathbf{h}_v]\right)\right)}{\sum_{l \in \mathcal{N}(u)} \exp(\left(LeakyReLU\left(\mathbf{a}^{\mathrm{T}}[\mathbf{W}\mathbf{h}_u || \mathbf{W}\mathbf{h}_l]\right)\right)} \tag{2.10}$$

where α_{uv} denotes the attention weight of node $v \in \mathcal{N}(u)$ to node u, $\mathbf{a}$ is the trainable attention vector, and $\mathbf{W}$ is the (trainable) weight matrix associated with the linear transformation applied to each node, in layer k (note that the superscripts were omitted in equation 2.10 for notation simplicity). Delving into the process, the initial step involves a shared learnable linear transformation parameterized by $\mathbf{W}$ applied to every node. Secondly, a shared attentional mechanism (a) computes the attention coefficients that give the importance of a node v´s features to node u, and is defined by:

$$e_{uv} = a(\mathbf{W}\mathbf{h}_u, \mathbf{W}\mathbf{h}_u). \tag{2.11}$$

The attention coefficients are computed only to the first-order neighbors of u. Thirdly, a normalization with the $softmax$ function of the attention coefficients is performed to make them comparable across

different nodes, as defined by:

$$\alpha_{uv} = softmax(e_{uv})_v = \frac{\exp(e_{uv})}{\sum_{l \in \mathcal{N}(u)} \exp(e_{ul})} \tag{2.12}$$

Finally, the definitive output features for each node are determined through a linear combination of the node features, weighted by the (normalized) attention weights. This computation is followed by the potential application of a non-linearity, denoted as σ [151].

To stabilize the learning process, K-independent attention mechanisms can perform the transformation defined in equation 2.9 in parallel. Each of these mechanisms produces a set of node features, which can subsequently be concatenated or averaged to form a final output representation. Each independent attention mechanism is referred to as an attention head, and this approach is denoted as multi-head attention (for further details please refer to [151]).

In contrast to the GCN, where the importance of a neighbor v for a target node u is determined by the weight of the connecting edge, the GAT enables the automatic learning of the importance of node v to node u by considering their respective features, and (softly) select its most relevant neighbors. This capability proves advantageous in scenarios where the data does not inherently conform to a graph structure, and subsequently the graph generated might be noisy [155]. However, in 2022, Brody *et al.* [150] proposed an improved variant of the GAT. The authors refer that the successive application of linear layers ($\mathbf{a}^{\mathrm{T}}$ and $\mathbf{W}$ in equation 2.10) in the GAT leads to a shared ranking of attention coefficients across all nodes, a limitation that renders the model incapable of handling scenarios where different nodes hold varying relevance to distinct target nodes. Recognizing this issue, Brody *et al.* [150] referred to it as static attention, and proposed GATv2. GATv2 distinguishes itself from the standard GAT by altering the operation sequence, which enables the network to perform dynamic attention. Dynamic attention empowers GATv2 to assign unique relevance values to different nodes for varying target nodes. Moreover, GATv2 demonstrates heightened robustness to edge noise, effectively mitigating the impact of noisy edges, while GAT's performance degrades considerably with increasing noise.

2.3.6 Challenges and enhancements in Graph Neural Networks training

2.3.6.1 Over-smoothing

During the training of GNNs a prevalent issue known as over-smoothing can compromise the performance of GNN models [156]. Over-smoothing occurs when too many GNN layers are stacked together to form the model architecture. As previously explained, in each iteration, corresponding to each GNN layer, node features from the local neighborhood undergo aggregation, becoming part of the new vector representation of the target node. In this way, the fundamental idea of over-smoothing is that after several iterations (by stacking several GNN layers), all the nodes within the graph will tend to exhibit highly similar representations [142]. This phenomenon affects the performance of downstream tasks, such as node classification, as it becomes difficult to distinguish between different classes from very similar node embeddings [155].

This issue suggests that augmenting the number of layers in GNN models does not yield performance benefits. Addressing over-smoothing becomes a pertinent question. In scenarios of over-smoothing, the updated node representation becomes excessively reliant on the aggregated incoming message from neighbors at the cost of preserving the node representations from previous layers. Skip connections have been employed to alleviate this phenomenon and will be introduced in the subsequent section [142].

2.3.6.2 Skip Connections

Skip connections equip the model with an architectural advantage that preserves information from previous layers of message passing during the update step and enhances convergence, stability, and overall model performance. Skip connections are inspired by the concept of residual learning introduced in CNNs by He *et al.* [157] in the context of image classification [158]. In general, these connections allow the network to directly pass information from one layer to its subsequent layers, bypassing any intermediate layers. By establishing skip connections, GNNs can avoid compounding effects of aggregation mechanisms by enabling the interplay between local and global information while the network deepens. With it they prevent the network from completely relying on the aggregation of local neighborhood information, helping the model to overcome the over-smoothing problem [142].

In addition to addressing the challenge of over-smoothing, the incorporation of skip connections in GNNs serves a dual purpose by mitigating the vanishing gradient problem [159]. This issue hinders the effective propagation of gradients through deep networks, leading to their diminishment ("vanishing") as they are backpropagated from the output layers to the earlier layers [159]. Similar to its impact on CNNs, the vanishing gradient problem is also a noteworthy concern in GNNs. By introducing supplementary pathways for gradient flow, the integration of skip connections prevents the attenuation of gradients. Consequently, this facilitates the scalability of the designed GNN models to deeper architectures, enhancing their overall performance.[142].

Several skip connections methods have been developed which can be used in conjunction with the GNNs described [160, 142]. Residual connections pass the output from the previous layer to the layer ahead by matrix addition. They are computationally simple as they don't increase the number of parameters [160]. Dense connections are inspired by DenseNets proposed by Huang *et al.* [161]. Unlike residual connections, dense connections concatenate the output embeddings of a layer with those of the next layer[160]. In dense connections, every layer is connected to all the preceding layers. The idea behind the concatenation is to use features that are learned from earlier layers in deeper layers as well [142]. Jumping knowledge (JK) connections [162] leverage the representations at each layer of message-passing by aggregating the node embeddings of every layer as the final output of the GNN. This aggregation can be achieved through concatenation, max-pooling, or long short-term memory (LSTM) attention layers. JK connections that use concatenation can be seen as a simplified case of dense connections by omitting the complex skip connections at the intermediate layers. JK connections often lead to consistent improvements in varied tasks, such as in social, bioinformatics, and citation networks, being generally a useful component to employ in GNN models [142, 163, 158].

2.3.6.3 Node Sampling: Subsampling and Mini-Batching

Mini-batch training is commonly performed in DL to accelerate model convergence while limiting the memory footprint when compared with full-batch training. However, when dealing with graphs, which are relational data structures with interconnected nodes, special considerations are necessary to perform mini-batch training in GNN without compromising the relationships within the graph.

Recall the computational flow of a two-layer GNN in Figure 2.8. The best way to guarantee that the information is passed across the graph without losing information is to form mini-batches of target nodes considering their computational flows (from now on called computation graphs). To achieve this, embeddings for a set of M different nodes in a mini-batch are computed using M distinct computation graphs. However, training on different computation graphs independently introduces challenges. Firstly, redundant computations may occur when multiple nodes share neighbors. Secondly, the computation graph

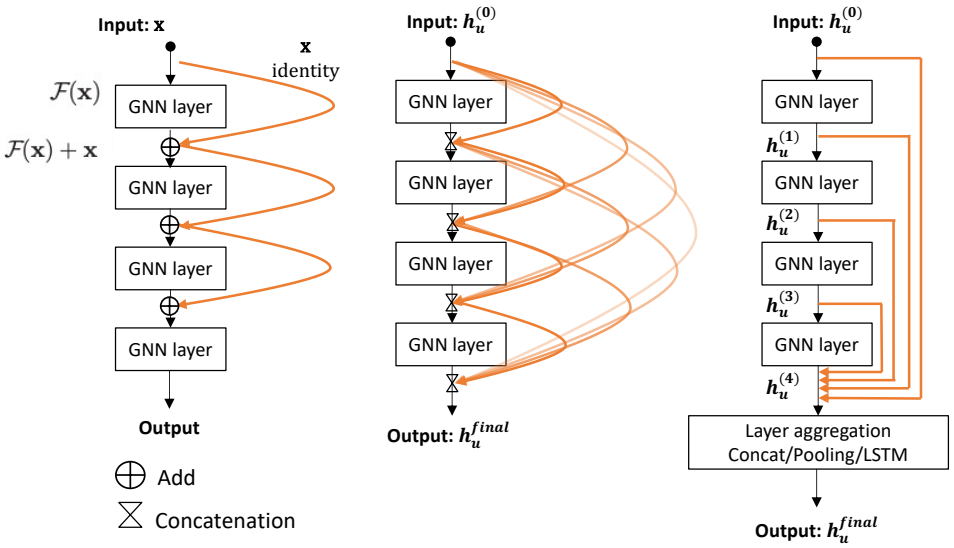


Figure 2.9: Illustrations of 4-layer GNNs featuring distinct types of skip connections. On the left, residual connections exhibit a configuration where each layer is bypassed, and its output is summed with the node embeddings of the subsequent layer. In the middle, dense connections showcase a structure where the node features of each layer are concatenated with those of the previous layers. On the right, JK connections produce an output through the aggregation, employing concatenation, max-pooling, or LSTM attention layers, of the node embeddings from every layer.

size grows exponentially with the model depth (i.e., the number of layers), leading to a problem known as "neighbor explosion" [163]. To tackle this issue, Hamilton *et al.* [146] proposed a strategy commonly known as neighborhood sampling. The key concept is to randomly subsample neighbors for each node, employing a small fixed sample size [142], typically 2 to 50 neighbors are selected by one node in the next layer, instead of using the entire neighborhood [163, 164]. Subsequently, the pruned computational graphs are used to efficiently compute the node embeddings, reducing computational costs. Although this technique holds good results, a few caveats need to be taken into consideration. For instance, as the number of layers increases, the pruned computation graph still experiences exponential growth, and it does not address the problem of computation redundancies when multiple nodes share neighbors. The development of sampling and mini-batching techniques for training GNNs is an extensively researched area, with various proposed methods [146, 165, 166, 167]. For instance, Wei-Lin *et al.* (2019) [166] proposed the ClusterGCN method, which samples subgraphs by combining graph clusters obtained through graph clustering algorithms. Besides, ClusterGCN updates the node embeddings layer-wise, enabling the re-use of embeddings from the previous layer and reducing the computational cost, although it may result in systematically biased gradient estimates.

2.3.7 Explainable AI

DL models are widely used across various domains, prompting interest in applying them to critical real-world scenarios [168]. However, their perceived black-box nature due to opaque decision-making processes raises concerns about trustworthiness, particularly in fields like healthcare [169]. To ensure safe and reliable deployment, DL models must not only provide accurate predictions but also offer human-interpretable explanations for their reasoning [170, 168]. GNNs excel at capturing complex relational and node feature information within graphs, but lack interpretability tailored to their unique data structure (i.e. inherent irregularity and discreetness). Consequently, there is a growing need for methods specifically designed to explain GNN predictions [168].

Explainability methods can be categorized at four different levels based on their characteristics: stage, target, applicability, and methodology type. At the stage level, explanations can be ante-hoc or post-hoc [171]. Ante-hoc (or intrinsic) explanations are provided by inherently interpretable models, while post-hoc explanations elucidate already constructed models (i.e. pre-trained models), requiring an external method for explanation. In terms of target, methods can yield instance-level or model-level explanations. Instance-level (or local) explanations are specific to individual examples, whereas model-level (or global) explanations offer overarching insights and a general understanding of model behaviour [168]. Regarding applicability, post-hoc explanations are model-aware or model-agnostic. Model-aware (or model-specific) explanations are derived from methods that directly analyze the model parameters to explain the relationship between input and output. Conversely, model-agnostic methods are independent of the model architecture focusing solely on inputs and outputs without examining internal model mechanisms [170]. Finally, explainability methods can be classified based on their methodology for computing importance scores, which include gradients/features-based methods, perturbation-based methods, decomposition methods, counterfactual methods, surrogate methods, and generation methods [170].

The attention mechanisms integrated into networks such as the GAT enhance interpretability by analyzing the attention scores computed by them during training, yielding intrinsic explanations. These mechanisms help extract relevant information from the input graph that aids in model predictions. However, given the interconnected nature of graph structure and node features, it is critical to discern the type of information that can be retrieved from the explanations provided by each proposed method. This distinction is particularly important for the attention scores, which solely elucidate predictions by focusing on graph structure and lack node feature information. Furthermore, these explanations are constrained to specific GNN architectures [172].

Ying *et al.* (2019) introduced the first method aimed at explaining the predictions produced by any GNN model based on the message-passing framework on graphs, operating as a post-hoc model-agnostic approach. This method, termed *GNNExplainer*, was developed recognizing the intricate interplay between nodes, their corresponding features, and the connections (edges) within the input graph that collectively influence the model's predictions [172]. Considering a specific instance, this perturbation-based method, generates both a node feature mask to mask out unimportant node features, and a graph mask that delineates the most important subgraph for the prediction of a specific instance's label, within the computation graph of the instance.

For a specific prediction $\hat{y}_i$, GNNExplainer generates an explanation (G_S, X_S^F) (see Figure 2.10). G_S represents the most influential subgraph of the computation graph (G_c) of node v_i (i.e. $G_S \subseteq G_c$), illustrated by the green arrows in Figure 2.10. X_S is the associated feature matrix of G_S, and X_S^F represents the masked version of this matrix using the mask F, with $F \in \{0,1\}^m$ (m is the dimension of each node feature vector). X_S^F encompasses the most influential node features for explaining $\hat{y}_i$ when considering the subgraph G_S, as depict in Figure 2.10 B. The goal is to identify G_S (and X_S^F) that retains as much information as possible compared to the use of the computation graph G_c (and X_S) for predicting $\hat{y}_i$. The authors proposed to maximize the mutual information metric to determine (G_S, X_S^F) through the optimization of the following objective function:

$$\max_{G_S,F} MI(Y,(G_S,F)) = H(Y) - H(Y|G = G_S, X = X_S^F), \tag{2.13}$$

where, Y indicates the probability of nodes belonging to each of C classes, and $H()$ represents the entropy. MI quantifies the change in the probability of prediction $\hat{y}_i$ when the computation graph of node v_i is confined to subgraph G_S and its node features restricted to X_S^F. In essence, for a node v_i, if removing

a node v_j or its connecting edge from the computation graph of v_i significantly impacts the prediction probability $\hat{y}_i$, decreasing it, then the presence of this node or edge, respectively, exerts a significant influence on the prediction at node v_i.

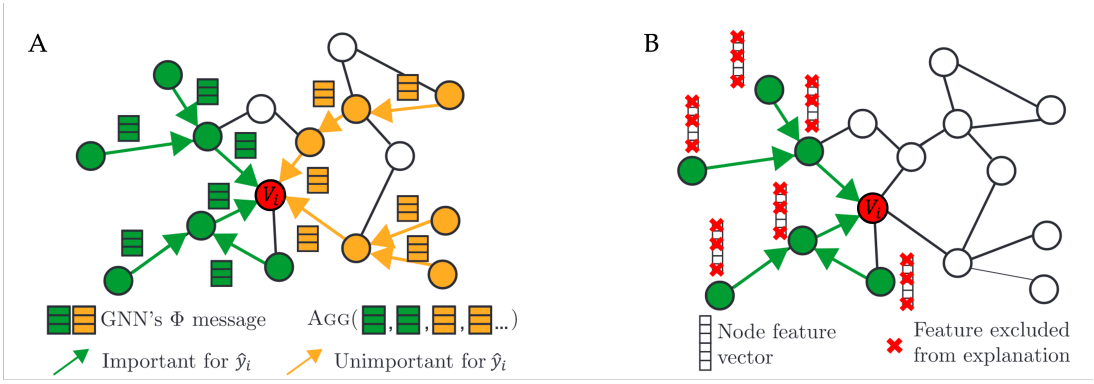


Figure 2.10: Illustration of the GNNExplainer framework. A. Computation graph G_c of node v_i (highlighted in green and orange). The edges represented in green are deemed critical by the GNNExplainer in forming message-passing pathways for the propagation of useful node feature information across G_c and aggregated at v_i to arrive at prediction $\hat{y}_i$. B. Additionally, the GNNExplainer identifies the most important node feature dimensions among the nodes forming the subgraph delineated by the green edges, and which are identified by the method as the pivotal for predicting $\hat{y}_i$ (adapted from [172]).

In equation 2.13, the entropy of Y, $H(Y)$, remains constant for a trained GNN, simplifying the optimization task to minimizing the conditional entropy:

$$\min_{G_S,F} H(Y|G = G_S, X = X_S^F). \tag{2.14}$$

However, due to the exponentially large number of possible subgraphs G_S in the computation graph G_c, direct optimization of equation 2.14 is infeasible. To address this, the authors propose treating the explanation as a distribution of plausible explanations instead of a single graph, turning the discrete optimization into a continuous process optimized by gradient descent. This involves learning a mask $M \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ of the computation graph with adjacency matrix A_c, with the process being reduced to:

$$\min_{G_S,F} H(Y|G = A_c \odot \sigma(M), X = X_S^F) \tag{2.15}$$

where $\odot$ represents the element-wise multiplication, and σ is the sigmoid function mapping the mask to $[0,1]^{|\mathcal{V}|\times|\mathcal{V}|}$. A similar approach is applied to learn mask F that masks the unimportant features in the feature matrix of the subgraph G_S. To account for features important for prediction but with values close to zero, the method marginalizes over all feature subsets and uses a Monte Carlo estimate to sample from the empirical marginal distribution for nodes in X_S during training, with X being reparameterized to: $X = Z + (X_S - Z) \odot F$, where Z is a m-dimensional random variable sampled from the empirical distribution.

It is noteworthy that they incorporate multiple regularization terms to ensure that the computed explanations possess desired properties, such as compactness. This is achieved through the inclusion of two parameters: one that limits the number of nodes in the subgraph G_S and another that restricts the maximum number of features retained in X_S^F.

Subsequent to this work, various other explainability methods for GNNs have been proposed, each with its distinct set of characteristics and applications. Notably, GNNExplainer has been extensively applied and validated in numerous studies [170]. Another notable method is the PGExplainer [173], which focuses on identifying the subgraph structures crucial for the predictions made by the GNN model,

without identifying the relevant node features. It is important to recognize that explainability in graph-based models is not as mature as in other areas of ML, such as computer vision, with only a limited number of methods available for each application.

Chapter 3

State of The Art

Considering the goal of this dissertation project, this chapter will conduct a literature review regarding the state-of-the-art of ASD diagnosis resorting to ML approaches with functional connectivity data. Firstly, section 3.1 presents an overview of the current ML methods applied in ASD and their challenges. Secondly, a detailed review of the functional connectivity patterns that characterize ASD and of the relevant ML and DL methods that use FC data to diagnose autistic individuals is performed in section 3.2. Finally, in section 3.3 different approaches using GNNs and FC data in ASD are analyzed.

3.1 Artificial Intelligence in the Diagnosis of ASD

The promise that AI-based technologies can surpass human performance in recognition of symptoms, and early diagnosis while at the same time reducing human error by analyzing large and complex amounts of data has evoked a particular interest in using AI for the diagnosis of many pathologies [174]. As outlined in Chapter 2, the diagnosis of ASD is only based on behavioral features. Thereby, if successful, AI-based technologies will diminish subjectivity and contribute to the diagnosis process's reproducibility. ML and DL techniques can be used in the analysis of biomedical and neuroimaging data as well as novel observational data, such as eye movements, facial expressions, and postural data, or even assisting with current assessment techniques (questionnaires, such as ADOS-2) [175]. Over the years, several brain imaging methods including fMRI, sMRI, diffusion tensor imaging (DTI), electroencephalography (EEG), magnetoencephalography (MEG), functional near-infrared spectroscopy (fNIRS), and positron emission tomography (PET) have been used to extract ASD biomarkers [72, 176]. Given this dissertation scope, the following state-of-the-art on the application of AI in ASD diagnosis will target the use of ML and DL techniques to analyze brain imaging data.

In a 2021 review regarding brain imaging-based ASD classification using ML (including DL) methods, Xu *et al.* [177] provides an analysis of 119 studies. As outlined by the authors the number of papers on this topic has been increasing year after year. The use of functional imaging has also been growing, with more studies resorting to this type of data instead of structural imaging data. In line with their analysis, models based on fMRI data can achieve better performance in the classification task.

In a recent review and meta-analysis, Song *et al.* [176] analyzed 44 studies, published between 2010 and 2021, to evaluate "the use of ML with neuroimaging data to distinguish ASD from TD individuals". Most ML approaches used a Support Vector Machine (SVM) model (18 out of 44 studies) for the classification problem (to refer some [178, 179, 180, 181, 182, 183, 184, 185]). The majority of the studies resorted to resting-state functional magnetic resonance imaging (rs-fMRI) data (22 out of 44 studies), followed by sMRI (8 out of 44 studies) as the second most used type of data. The pooled sensitivity

and specificity achieved by these models were 86.25 95% CI (81.41, 90.08) and 83.31 95% CI (78.12, 87.48), respectively.

Despite the very promising results that have been achieved with ML models in ASD diagnosis, several factors have to be considered. For example in the paper by Song and colleagues [176], the corpus embodied small-sized studies. Studies with small sample sizes are prone to overfitting. In fact, as indicated by the authors [176], the pooled sensitivity and specificity decreased, respectively, to 83.23 95% CI (76.79, 88.16) and 78.90 95% CI (70.85, 85.19) when considering only studies with large samples (more than 100 participants). Beyond that, Song *et al.* pointed out that due to the disproportional male:female ratio in the prevalence of ASD 10 studies did not include female participants. Nevertheless, studies that considered both female and male participants obtained a better sensitivity in discriminating ASD from TD individuals. Focusing solely on males has the potential to introduce a bias in the outcomes and constrain the ability to generalize the findings. Thus, to address the core neuroimaging features of ASD both female and male participants should be included in the study sample [176]. Besides that, the results of the meta-regression performed by them suggest a similar performance in the classification task regardless of age [176]. Finally, they highlighted that conventional ML algorithms are incapable of conducting feature extraction, as they have been developed for classification relying on predefined optimal features, while DL algorithms can extract automatically optimal features from the data. Notwithstanding, to achieve a high classification performance, DL networks demand large datasets to train, which may not always be clinically feasible [176].

In 2012, the ABIDE database described in Di Martino *et al.* [186] was made available for researchers widely. ABIDE comprises a large sample of rs-fMRI, the corresponding sMRI data as well as the respective phenotypic information collected from laboratories all around the world, and currently, encompasses two subsets: ABIDE-I and ABIDE-II (for more details regarding ABIDE see Chapter 4 or [186, 187]). Furthermore, preprocessed images resorting to different pipelines and brain atlases of ABIDE-I are publicly available, courtesy of the Preprocessed Connectomes Project (PCP). Normally, from the 1112 participants of ABIDE-I are only used 871 subjects (403 ASD and 468 TD) that met the imaging quality and phenotypic information criteria [178, 17, 188, 189]. Not surprisingly, the ABIDE dataset became very common in ASD studies, which corresponds to one of the key reasons for the increased attention in detecting and finding fMRI biomarkers, in recent years [176]. The use of large datasets is important for the reliability of the results, as a large quantity of data brings increased statistical power [190].

Notwithstanding, inherent problems with ABIDE should be considered. Not only it is a multi-site set considering 17 different acquisition sites in ABIDE-I, with high inter-site variability in terms of neuroimaging data, but there is also significant variation in terms of the number of individuals and their demographic and diagnosis distribution across different acquisition sites, with some sites having only male participants and a higher prevalence of neurodiverse individuals, while other sites have a starker diagnosis imbalance for females compared to males. Furthermore, despite ASD being a spectrum disorder, the prediction of diagnosis when resorting to ML models is performed as a binary classification problem. In fact, this is a general problem regarding studies of ASD prediction, due to the unavailability of datasets with different ASD subtypes [72].

ASD is characterized by atypical FC (as discussed in the subsection 2.1.2). Thereby, numerous studies resorting to ABIDE aim to optimize ML or DL models for the diagnosis of ASD [188, 191, 192, 193], with encouraging results in performance and further identification of atypical functional connectivities that play important roles in the classification task and are possibly relevant FC biomarkers of the disorder.

In the next section will be first identified the most relevant FC findings in ASD individuals encountered in the literature and subsequently will be presented and discussed several significant studies that

using ML and DL studies and FC data aimed at finding ASD biomarkers.

3.2 Functional Connectivity and AI in ASD

Several methods, metrics, and statistical analyses have been employed to extensively investigate functional brain connectivity in individuals with ASD. In general, after deriving the FC data from rs-fMRI images, most studies identify atypical FC between or within RSNs in ASD groups and, further, try to establish relations between these alterations and the ASD core symptoms. In addition, some studies endeavor to determine associations between the identified FC findings and the severity of symptoms in individuals with ASD.

One of the first rs-fMRI studies focused on the comparison of intrinsic DMN connectivity in ASD and TD individuals [194]. Kennedy and Courchesne used seed-based analysis to examine the resting-state FC data of the DMN. DMN is suspected to be important in socio-emotional behavior, and both are impaired in ASD, thus, they hypothesized a significantly altered connectivity in this RSN. The authors found a significant under-connectivity in the DMN in the autistic patients when compared to the TD individuals.

Since DMN is considered to be important in socio-emotional behavior, it is contemplated to be implicated in the Theory of Mind (ToM) hypothesis [195]. ToM refers to a person's capacity to comprehend subjective mental states, including thoughts and desires, whether a scenario is real or hypothetical. This ability enables individuals to understand and predict the behavior of others based on their mental states. ToM is considered to be implicated in autistic behavior[30].

Thus, numerous works have focused their analysis in the DMN [196, 197, 198, 199]. In 2016, Lee *et al.* [200] performed a voxel-wise whole-brain analysis of the rs-fMRI data. They found under-connectivity in the medial prefrontal cortex, posterior cingulate cortex, inferior parietal lobule, and sensorimotor regions and proposed that under-connectivity in the medial prefrontal cortex and posterior cingulate cortex (DMN regions associated with ToM) may be related to social functioning impairments in ASD. Wang *et al.* confirmed this finding [199]. As highlighted by the authors, decreased FC between the medial prefrontal cortex and the posterior cingulate cortex is a robust finding in ASD [199]. Wang and colleagues performed a meta-analysis of resting-state FC studies of DMN to identify common abnormalities in ASD patients, in comparison with TD individuals. Furthermore, they provided evidence to confirm a suspicion in the literature regarding the link between disruption in the precuneus (another essential node of the DMN associated with ToM) and the struggle in inferring mental states of others. Moreover, the authors reported abnormal intrinsic FC in the cerebellum, particularly over-connectivity. Cerebellar regions are intimately linked to somatomotor and brainstem circuits and are involved in fundamental sensorimotor actions. Thus, increased FC of the cerebellum may be related to the subjacent sensorimotor processing dysfunction in ASD individuals [199].

The SMN is also associated with the ASD symptomatology [201]. Deficits in this RSN include difficulties with social communication and engagement, abnormal sensory responsivity, and repetitive and constrained behaviors [202]. In fact, a 2015 study proposed that sensorimotor deficits are primary features of ASD, possibly occurring before social and communication deficits [87]. In 2020, Liu *et al.* investigated which connectivities among whole-brain resting-state FC contribute more to ASD severity. They found that the majority of SMN-related intrinsic FC was negatively correlated with severity [201]. Furthermore, Liu and colleagues confirmed the results of [197], that the DMN-related functional connectivity magnitudes were negatively correlated with the severity of ASD. These areas are marked by an under-connectivity, as pointed out by Assaf *et al.* [197]. Additionally, both under- and over-connectivity

were found between numerous DMN areas and visual, motor, somatosensory, subcortical, ventral attention, salience, and reward networks using seed-based analysis, while resorting to a network analysis approach (in this case, graph theory metric) ASD individuals showed increased intrinsic FC of the DMN with other networks compared to TD individuals [203].

The role of the insula cortex (region of the SN) in ASD has also been examined. This region is considered critical in emotional and social processing by supporting the neural representation of the own physiological state [204]. Particularly, it was reported decreased connectivity in this region [205]. The insular cortex together with the amygdala (also part of the SN) and regions including the medial prefrontal cortex, angular gyrus, and posterior cingulate cortex are referred to as 'social' brain regions. Accordingly as mentioned above for the medial prefrontal cortex and posterior cingulate cortex, the insula, amygdala, and angular gyrus are also characterized by a decreased resting state FC in ASD, as found in the study by Hagen *et al.* [205] using both ICA and seed-based analysis approaches. The regions mentioned are part of the SN (insula cortex, and amygdala) and the DMN (medial prefrontal cortex, angular gyrus, and post cingulate cortex). In 2021, Chen *et al.* [206] focused on characterizing the inter-network resting state FC between SN and DMN. The authors used a large sample (325 ASD/356 TD) from the ABIDE and considered the medial prefrontal cortex as the seed for the seed-based analysis. They reported a significant increase in FC between the medial prefrontal cortex and the right anterior insula in the ASD group in comparison with the TD group and extrapolated that considering the medial prefrontal cortex the core region of the DMN and the anterior insula the core region of the SN, these two networks are strongly connected in ASD individuals [206].

In line with this finding, there is an emerging theory of weaker functional network segregation in ASD [88]. Specifically, with a combination of intra-network under-connectivity and inter-network over-connectivity in children and adults [88].

Until now were summarized several findings regarding resting-state FC in ASD patients relative to TD individuals with the goal to present a review of the relevant RSNs and specific nodes of these networks with functional dysfunction in autistic individuals and its relation with the core symptoms of ASD. While the discoveries presented are very promising, there are several inconsistencies in the literature. In 2017, Hull *et al.* performed a thorough review in an attempt to draw overall conclusions from systematically examining the resting state fMRI autism literature [207]. They reported that the overall discrepancies could come from the diversity of the sample, in sex and age, different designs of the resting-state scan, and the preprocessing and methodology of analysis (for example, seed-based analysis, ICA, and graph theory), with particular emphasis on the latter [207]. Furthermore, Hull *et al.* emphasized that the diversity of findings can be also influenced by the heterogeneous nature of the condition, which is characterized by individual variations in FC organization [207].

Despite the lack of unanimity regarding connectivity, researchers have continued to examine the use of altered connectivity as a biomarker for the diagnosis of ASD [207], particularly using ML and DL classifiers.

This type of study typically follows a general pipeline to link functional connectomes to the target phenotype that comprises crucial generic steps: 1. definition of brain regions (ROIs) from rs-fMRI images or using already defined reference atlases, 2. quantifying functional interactions from time series signals extracted from these ROIs and 3. comparisons of functional interactions across subjects using supervised learning [208]. Yang *et al.* (2022) [189] followed a similar pipeline. The study included 871 subjects, specifically 403 ASD and 468 TD individuals from the ABIDE I repository. They used the BASC444 functional atlas to define the ROIs, subsequently extracted the time-series and constructed the FC matrices, with Pearson's correlation coefficient, that were fed to a Radial Basis Function (RBF)

kernel SVM to achieve an accuracy of 69.43%, with associated sensitivity and specificity of 64.57% and 73.61%, respectively. These results illustrate the potential of using a ML classifier with intrinsic FC connectivity data to identify ASD subjects. Furthermore, Yang and colleagues also compared different models, particularly, testing a DNN with 8-hidden layers. Despite the DNN underperforming relatively to the kernel SVM, achieving an accuracy of 68.45% (sensitivity of 62.70% and specificity of 73.61%), the authors highlighted the potential of DL models trained on large datasets to extract important features to the classification and further improve the overall performance of the models to distinguish between ASD and TD individuals [189].

In 2020, Sherkatghanad *et al.* [188] developed a CNN model to detect ASD from rs-fMRI data. They resorted to the preprocessed ABIDE I dataset of the PCP and chose the Configurable Pipeline for the Analysis of Connectomes (CPAC), which includes slice timing correction, correction for motion, skull-striping, normalization of voxel intensity, and nuisance regression to delete the signal fluctuations caused by head motion, respiration, cardiac pulsation, and scanner drift. Also includes band-pass filtering ($0.01 - 0.1$Hz) and spatial registration to the MNI152 template space. They used the CC400 functional parcellation atlas to define the ROIs of 871 subjects (403 ASD and 468 TD). The brain connectivity matrices were constructed by calculating Pearson's correlation coefficient between each region. The developed CNN architecture achieved an average accuracy of 70.20%, with corresponding sensitivity of 77.00% and specificity of 61.00% [188]. They, further, used the salience technique to visualize the ROIs that play important roles in the classification task, having identified the right supramarginal gyrus, fusiform gyrus, and cerebellar vermis as playing a significant role in ASD diagnosis [188]. These results support the hypothesis of the disruption of the anterior-posterior FC in ASD [188] and reveal evidence of the potential in using intrinsic FC and DL approaches to classify ASD individuals and possibly extract discriminant features of the disorder.

Unsupervised learning neural networks have also been used. Almuqhim *et al.* (2021) [192] developed the ASD-SAENet resorting to a sparse AE architecture for feature reduction dimensionality. The features encoded by the AE were then utilized as input to a DNN with two hidden layers for the classification of ASD and TD subjects. The dataset was, also, the ABIDE I preprocessed with CPAC pipeline, although the brain was parcellated in 200 regions using the CC200 functional atlas. Interestingly, besides being evaluated in the whole ABIDE I dataset, the model was also evaluated on each site separately, to demonstrate how the ASD-SAENet performs on small datasets, and to validate the generalizability of the model across different data acquisition sites and MRI scanners [192]. In the whole dataset, the model achieved an accuracy of 70.8% (sensitivity of 62.2% and specificity of 79.1%). The average accuracy across sites was 64.6%, revealing superior generalizability [192] when compared to the anterior work by them [209] (that was the state of the art).

Considering the heterogeneity of ASD and the atypical structural and functional connections that characterize the disorder, the possibility of integrating clinical information, such as age, sex, or acquisition site [17], and/or structural brain imaging, such as sMRI [210], with fMRI data have been researched promisingly to improve the performance of DL models in ASD diagnosis, as outlined in the review article by Khodatars *et al.* [72]. A multi-modal imaging approach, considering structural (sMRI) and functional (rs-fMRI) data, was proposed by Rakić *et al.* (2020) [210] using the ABIDE I dataset for training a combination of stacked AEs and a MLP, which achieved a mean accuracy of 85.06%. Nonetheless, this study did not consider phenotypic data, which could improve the performance results. Diagnostic habits of medical doctors include the use of both phenotypic and imaging data. Thus, in a different study by Dvornek *et al.* [211] they employed a multi-modal approach considering phenotypic data, including age, sex, handedness, full IQ, and eyes status during the fMRI scan and rs-fMRI data from the ABIDE I, in a

single LSTM model, that achieved better accuracy than a model using only the rs-fMRI data (70.1% and 67.9%, respectively).

Combining phenotypic information into a NN built for imaging data is not a straightforward task [211]. One researched solution is a graph representation of a population of individuals. A graph is a powerful and intuitive way of integrating imaging and non-imaging data, encoded in the nodes and edges. Recently, DL in graphs has been gaining attention with the emergence of GNNs [212]. These graph-based models can process unstructured relational data, leverage multi-modal data, and benefit from exploiting the associations between subjects in a population graph. Numerous studies have employed GNNs in disease prediction, particularly GCNs in the diagnosis of ASD with FC combined with phenotypic data, as will be discussed in the next subsection.

3.3 Graph Neural Networks to study FC in ASD

GCNs are a relatively recent concept [154] that appeared from the desire to perform convolutional operations in irregular domains, such as graphs. The GCN models applied in disease prediction for ASD using FC data are divided into two groups according to the type of graph considered: individual and population graph. In the former, each graph represents one subject, nodes are brain regions and the edges are functional correlations between time-series observations from those regions. In the latter, each node represents a subject, and edges represent the pairwise similarity between them, which can be determined by considering phenotypic information (Figure 3.1) [135]. Thus, this approach is particularly promising due to its capability to deal with large populations and exploit different modalities of data in a single model encoding complex pairwise similarities and in parallel considering individual subject features and characteristics. Several works are available in the literature that have applied GCN models for population-based ASD prediction [17, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222]. Table 3.1 presents a summary of some of the most interesting studies in the literature that will be subsequently discussed.

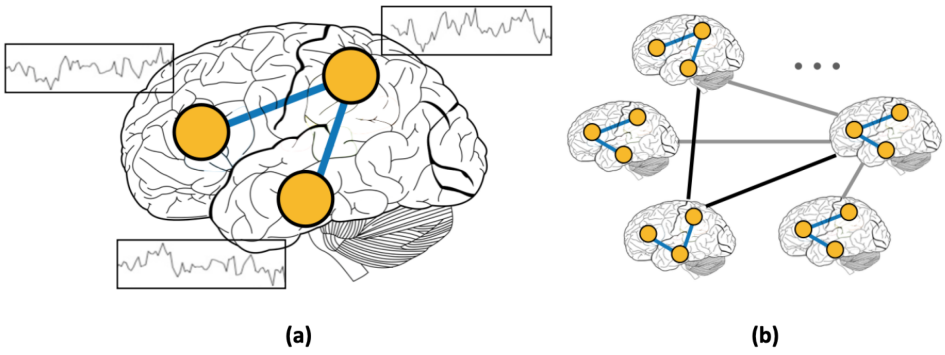


Figure 3.1: Types of graphs proposed in GCN models for ASD prediction: a) individual graphs, and b) population graphs (from [215]).

Parisot *et al.* (2017) was the first work in this domain [223]. In 2018, they published their extended work [17]. In this, Parisot and colleagues employed a GCN for the diagnosis of ASD individuals, which receives as input a population graph with a multi-modal approach and is denominated in the literature as PopGCN. Each node of the graph is represented by a reduced version of the vectorized FC matrix of each individual and the phenotypic information is used to determine the similarity measure between subjects and is encoded in the edges [17]. In this case, the diagnosis task consisted of a binary node classification

Table 3.1: Overview of the GCN approaches used in the diagnosis of ASD with FC and phenotypic data from ABIDE-I. ChebConv: Chebyshev Convolution. acc.:accuracy. spe. specificity. sen.:sensitivity. prec.:precision. NR: Not Reported in the paper.

Reference	Phenotypic data	GNN name	GNN Operator	#GNN layers	Type of Learning	Performance
[17]	Gender and Site	PopGCN	ChebConv	1	Transductive	70.4% acc. and 0.75 AUC
[216]	Gender and Site	hi-GCN	ChebConv	12	Inductive	67.2% acc., 0.75AUC, 68% spe. and 66% sen.
[219]	Gender and Site	TE-HI-GCN	ChebConv	NR	Inductive	76.5% acc., 0.76 AUC, 80% sen., 78% prec. and 78% F1
[218]	Age, Gender and Site	EV-GCN	ChebConv	4	Transductive	82.2% acc., 0.85 AUC and 83% F1
[220]	Gender and Site	MAMF-GCN	Snowball	9	Transductive	88.6% acc. and 0.94 AUC
[221]	Gender and Site	LG-GNN	ChebConv	5 local GNN 4 global GNN	Transductive	81.8% acc., 81% spe., 83 % sen., 0.85 AUC and 83% F1

problem, with a transductive setting, in which only a subset of the graph nodes are labeled during training and used for the optimization process, while all node features and respective similarities are fed to the GCN [17] with the unlabeled nodes being also aggregated and transformed during the graph convolution operation in training. The model had a shallow architecture consisting of one graph convolutional layer followed by a ReLU activation function and then a fully connected output layer activated by a softmax function. For the convolutions, the authors considered spectral graph convolutions and following the work of Defferrard *et al.* [224] restricted the class of considered filters to polynomial ones, that can be approximated by a truncated expansion in terms of Chebyshev polynomials [17], instead of performing the expensive Laplacian eigendecomposition. Feature selection was performed using the ridge classifier as the recursive feature elimination (RFE) method to reduce the feature vector to 2000 features and avoid overfitting problems, as the feature vector (6105) has high dimensionality compared to the graph size (871 nodes). They found out that gender and acquisition site are the phenotypic features considered that best explain the similarities between subjects and using them reported the model's best performance, with an accuracy of 70.4% and AUC 0.75 when establishing the (Chebyshev) polynomial order to $K = 3$ [17].

The associations between nodes (i.e., subjects) lack a standard definition and it is up to each researcher to decide and analyze the best way to define these relationships. As reported by Parisot and colleagues, the integration of redundant or wrong information can negatively impair the model's performance [17]. Therefore, conceiving an effective strategy for building the population graph is critical and far from obvious.

Jiang *et al.* (2020) [216] proposed a two-level hierarchical GCN (hi-GCN) framework capable of jointly learning the graph embedding from both the brain functional network and the population network in rs-fMRI images, which is generalizable to unseen data and therefore capable to be applied in an inductive setting. In 2021, Li *et al.* [219] built on top of the work developed by Jiang and colleagues [216] and proposed an Ensemble of Transfer HIerarchical Graph Convolutional Networks (TE-HI-GCN), i.e. applied a hi-GCN framework where they developed a novel strategy to apply transfer learning in GCNs. One of the main problems in network embedding learning with GCN is the need for a large collection of training data. Hence, Li and colleagues implemented a transfer learning scheme to learn generic graph structural features by leveraging the commonality in two related domains (ASD and AD), improving the performance results in relation to the previous work.

Both works [216, 219] compared with [17] employed an end-to-end hierarchical scheme in which the nodes features' embedding is learned automatically rather than extracted and the structure of the brain

functional network influence the similarity between nodes when constructing the population network. In 2022, Huang and Chung [218] developed a dynamic graph-based model called Edge-Variational Graph Convolutional Network (EV-GCN) capable of learning edge weights. They introduced the pairwise association encoder (PAE), a trainable module that determines the pairwise association between subjects based on their non-imaging information, such as phenotypic data, that aims at optimizing the inter-subject connectivity in the population graph. The GCN proposed is a well-regularized spectral graph convolutional network with four layers. To avoid overfitting the proposed network possesses an edge dropout layer [225]. Moreover, GCNs can suffer from an over-smoothing problem. When the number of layers of the network increases the nodes' features become increasingly similar as information propagates through the network, leading to difficulty discriminating between different nodes (as discussed in Section 2.3.6.1). Thus, to alleviate this problem, the authors adopted jumping connections (discussed in Section 2.3.6.2 [162]). Beyond that, the edge dropout step also helps with this, as it increases the graph sparsity [218]. In the end, their adaptive graph-based model achieved better results than the previous ones [17, 216, 219], and revealed the promise of using a trainable model to optimize the graph connectivity from the non-imaging data.

Later, Pan *et al.* (2022) [220] incorporated the PAE module [218] in their multi-scale adaptive multi-channel fusion deep graph convolutional network (MAMF-GCN). This model employs a multiple-graph approach with two different graph structures considered - phenotypic and functional - to exploit each modality specificity and two different atlases - AAL and HO - to construct different functional connectivity matrices to leverage multi-scale spatial and topological information. The phenotypic and functional graphs differ in the way the edge weights are determined. The former uses the PAE module to determine the similarity between subjects using phenotypic data. The latter is a construct resorting to the KNN highest (cosine) similarity. Each of them is fed to a specific channel. Besides that, the model encompasses also a common channel to obtain the correlation between different channels as although each modality and atlas has specific information, some information is certainly shared. In order to preserve the rich information from each modality and at multiple scale-spaces, Pan and colleagues proposed a multi-channel-based attention mechanism to flexibly incorporate the features of each channel. Furthermore, in order to extract the relevant information hidden in the complex multi-modal and multi-scale data, the model has to be extended to a deeper level. However, in this case, it could suffer from over-smoothing. Thereby, the authors employed a "snowball GCN module" proposed by Luan *et al.* in [226], that enables multi-scale feature information to be connected within the hidden layer while extending to deeper architectures. Their results revealed that harnessing multi-modal and multi-scale spatial data can improve substantially the prediction power of GCN models, obtaining better performance than the prior works.

A very interesting end-to-end approach was proposed by Zhang *et al.* (2023) [221]. Beyond leveraging the advantages of using a population graph-based model to learn the similarity between multiple subjects, Zhang and colleagues exploited the strengths of individual graphs in identifying and analyzing disease-related local brain regions and biomarkers in a model called local-to-global graph neural network (LG-GNN). The model is composed of two GCNs: local ROI-GNN and global Subject-GNN. They introduced a novel pooling strategy that employs an attention mechanism to selectively retain the most discriminative feature embeddings produced by the local ROI-GNN. This method reduces the size of the output embedding of the local ROI-GNN, which is given as input to the global Subject-GNN. Thus, significantly decreasing the total number of trainable parameters and avoiding overfitting, which ultimately yields substantial improvements for the global Subject-GNN and results in high performance. Besides its use as a pooling strategy, the attention mechanism aids the model's interpretability, identi-

fying the most discriminative brain regions: Inferior Frontal Gyrus, Precentral Gyrus, Frontal Orbital Cortex, Parahippocampal Gyrus, Frontal Operculum Cortex, Central Opercular Cortex, Planum Polare, and Accumbens. These brain regions are involved in language, execution, and memory tasks that are impaired in ASD [227, 228].

3.4 Final Considerations

This review aimed to first present a general view of the state-of-the-art of AI methods applied to ASD using neuroimaging data and respective challenges. Subsequently, perform a detailed literature review on the atypical functional connectivity found in ASD individuals and depict the state-of-the-art ML methods applied to ASD diagnosis using FC data. Finally, recent GNN approaches implemented in ASD were analyzed.

As reviewed, the availability of a large dataset such as ABIDE contributes to the elevated number of works developed resorting to rs-fMRI data and to improve the reliability of the results. Moreover, the use of additional phenotypic information mimics assessments by medical doctors and contributes to performance improvement in models resorting to this multi-modal approach. Population graphs have an important advantage in integrating both data types, considering the association between individuals to help in overcoming the inherent ASD heterogeneity. GNNs compared to other DL methods reveal a substantial enhancement in performance. Particularly, end-to-end approaches that learn the optimal node and edge features present the best results compared to methods resorting to hand-crafted ones. Furthermore, the use of attention mechanisms to select the most discriminative features in distinguishing ASD from TD individuals is particularly promising to overcome the lack of model interpretability and possibly identify biomarkers of the disorder. However, many of the developed GNN models are limited to a transductive learning setting, restricting their applicability to new data. Moreover, no studies were found that explored explainable artificial intelligence (XAI) methods on graphs for disease prediction.

Considering the above, this dissertation project focuses on the development of an inductive GNN model to differentiate autistic individuals from TD subjects by incorporating both rs-fMRI and phenotypic data. The model's generalizability will be assessed by testing it on an independent test set. Finally, the predictions made by the developed model on the independent test set will be interpreted using a XAI method for GNNs to analyze FC abnormalities characteristic of ASD.

Chapter 4

Materials and Methods

4.1 Data

The rs-fMRI images used during this project were obtained from Autism Brain Imaging Data Exchange (ABIDE) database [186]. ABIDE represents a collaborative initiative aimed at consolidating and disseminating previously collected rs-fMRI datasets from individuals with ASD and age-matched TD subjects to advance the understanding of the neurobiology of ASD. The ABIDE initiative is currently composed of two large-scale collections: ABIDE-I and ABIDE-II. In selecting the dataset for this project, ABIDE-I was preferred for two primary reasons. Firstly, it is well-established in the field. Secondly, it is readily available in a preprocessed state through common pipelines provided by the PCP initiative [229], facilitating the replication and extension of the work developed in this manuscript. The ABIDE-I is composed of data collected from 1112 subjects (539 ASD subjects and 571 TD individuals). It includes rs-fMRI images, with corresponding structural MRI and phenotypic information acquired in 17 international sites. The phenotypic data includes information about the subjects' age at the scan, the subject sex, cognitive assessment, medication status, and the communication and behavioral assessment values obtained with the diagnostic tools currently used, which were discussed in Section 2.1.4.

The ABIDE-I was preprocessed by four different preprocessing pipelines: the Connectome Computation System, the Configurable Pipeline for the Analysis of Connectomes (CPAC), the Data Processing Assistant for Resting-State fMRI, and the Neuroimaging Analysis Kit. The choice of the CPAC pipeline was deliberate, aligning with the preprocessing method commonly employed in existing literature for fair comparisons. The CPAC pipeline comprises slice time correction, motion correction, skull-stripping, global mean intensity normalization, nuisance signal regression, band-pass filtering ($0.01 - 0.1$Hz), and registration of fMRI images to a standard anatomical space - Montreal Neurological Institute 152 (MNI152). Subsequently, the mean BOLD signal time-series for 111 brain ROIs were extracted in accordance with the Harvard-Oxford brain parcellation atlas, including both cortical and subcortical regions (ROIs representing left/right white-matter, left/right grey-matter, left/right cerebrospinal fluid and brainstem were not considered) [123]. Moreover, the PCP initiative includes information about a quality assurance assessment conducted manually by three human experts who evaluated parameters such as the ghost-to-signal ratio, incomplete brain coverage, and high movement peaks. This process resulted in a subset of 871 subjects out of the initial 1112.

In this way, the dataset used during this study includes 403 ASD subjects (mean age 17.1 years, with age range 7.0 - 58.0 years) and 468 TD individuals (mean age 16.8 years, with age range 6.5 - 56.2 years), acquired in 17 different sites. There was no prior coordination between sites, resulting in variations in scan acquisition protocols. Moreover, the subsets pertaining to each site exhibit discrepancies in terms

of the distribution of diagnostic groups, sex proportions, and age distributions, which are illustrated in Figures 4.1, 4.2, and 4.3, respectively.

Figure 4.1: Distribution of each diagnostic group per acquisition site. In blue is illustrated the distribution of individuals that suffer from ASD, and in red are represented the TD subjects.

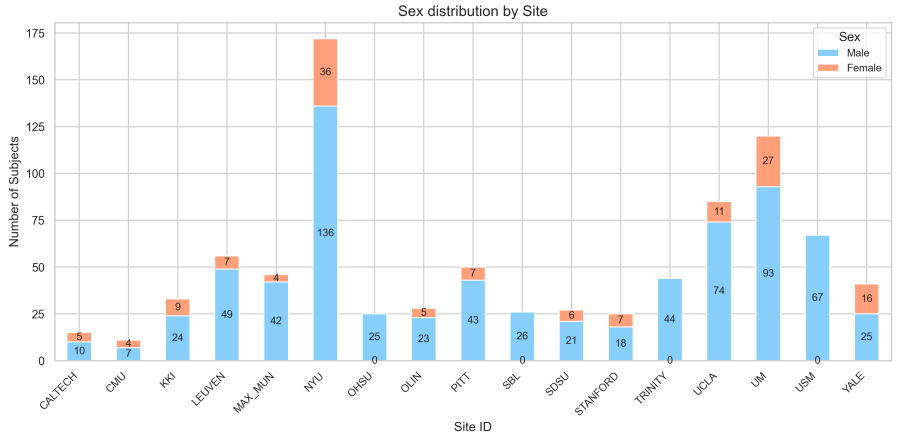


Figure 4.2: Distribution of the number of individuals of the female sex (in orange) and male sex (in blue) per acquisition site.

The MRI scanners used varied across the 17 independent sites, as well as the acquisition parameters, including the repetition time (TR), echo time (TE), and flip angle. Table 4.1 displays the diversity of MRI scanners and their associated parameters employed at each site. The field strength remained consistent across all sites, set at 3.0T.

4.2 Functional connectivity matrices

After downloading the preprocessed rs-fMRI data from the PCP initiative, FC matrices were computed. The mean BOLD signal time-series were extracted for each of the 111 brain ROIs included in the HO parcellation atlas, which provides information about intra-network correlations. The 111×111 FC matrices were computed using Pearson's correlation coefficients, allowing for the calculation of the linear correlation between each pair of brain ROIs. Each FC matrix was then Fisher transformed to improve normality (refer to Figure 4.4 for an example of a FC matrix). Since Pearson's correlation coefficient does not indicate the direction of FC, the resulting matrix is symmetric.

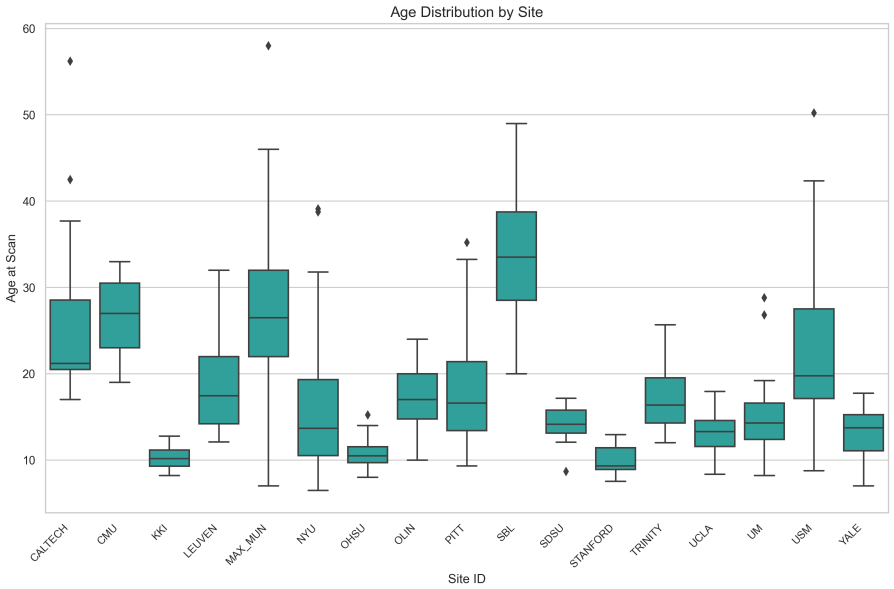


Figure 4.3: Boxplots illustrating the age distribution per acquisition site.

Table 4.1: Details regarding the scanners and scanning parameters employed at each site in the pre-processed ABIDE-I database. Variations in MRI scanner manufacturers and models, as well as differences in repetition time (TR), echo time (TE), and flip angle, are noticeable.

Site	Sample size	Manufacturer	MRI scanner	Field Strength (T)	TR (ms)	TE (ms)	Flip angle (degrees)
CALTECH	15	SIEMENS	TrioTim	3.0	2.000	30	75
CMU	11	SIEMENS	Verio	3.0	2.000	30	73
KKI	33	PHILIPS	Achieva	3.0	2.500	30	75
LEUVEN	56	PHILIPS	Intera	3.0	1.667	33	90
MAX_MUN	46	SIEMENS	Verio	3.0	3.000	30	80
NYU	172	SIEMENS	Allegra	3.0	2.000	15	90
OHSU	25	SIEMENS	TrioTim	3.0	2.500	30	90
OLIN	28	SIEMENS	Allegra	3.0	1.500	27	60
PITT	50	SIEMENS	Allegra	3.0	1.500	25	70
SBL	26	PHILIPS	Intera	3.0	2.200	30	80
SDSU	27	GE	MR750	3.0	2.000	30	90
STANFORD	25	GE	Signa	3.0	2.0	30	80
TRINITY	44	PHILIPS	Achieva	3.0	2.0	28	90
UCLA	85	SIEMENS	TrioTim	3.0	3.0	28	90
UM	120	GE	Signa	3.0	2.0	30	90
USM	67	SIEMENS	TrioTim	3.0	2.0	28	90
YALE	41	SIEMENS	TrioTim	3.0	2.0	25	60

4.3 Graph neural network model

This dissertation aims to develop a GNN for ASD classification utilizing FC data derived from rs-fMRI images, in a population-based approach to leverage relational information between each pair of subjects. Additionally, the study seeks to explore FC abnormalities and phenotypic characteristics among individuals with ASD. To accomplish these goals, two critical steps must be undertaken after computing the FC matrices. Firstly, the model needs to be optimized. Secondly, the optimized model's performance should be evaluated on an independent test set to assess its generalization capability to new data points.

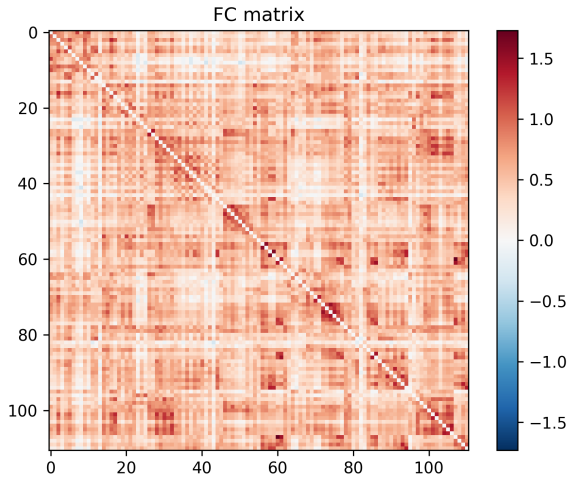


Figure 4.4: Example of a symmetric FC matrix obtained after completing all preprocessing steps.

For the evaluation of the independent test set to accurately gauge the model's generalization ability, an inductive setting is imperative. This ensures that the model does not have access to features of test subjects during training (for a detailed distinction between transductive and inductive settings, refer to Section 2.3.4). As discussed in the state-of-art in Chapter 3, current GNN models in ASD prediction through FC data have predominantly utilized GNN layers that are inherently transductive. To ensure a fair comparison with those models, transductive learning was chosen for model optimization. This comparison provides valuable insights into the model performance relative to state-of-the-art approaches. To facilitate this, the selected GNN layer must be applicable to both transductive and inductive settings. The GATv2 layer proposed by Brody *et al.* [150] meets these criteria and was therefore selected. As discussed in Section 2.3.5.2, GATv2 is a spatial-based approach capable of deployment in inductive learning. Furthermore, it enables automatic learning of the importance of node u to a node v, which is particularly crucial for datasets that do not inherently conform to a graph structure such as the ABIDE dataset. The choice of GATv2 over the original GAT is based on findings from Brody and his colleagues' work, demonstrating that the newest version consistently outperforms the original.

4.3.1 Data splitting

Upon selecting the GNN layer for implementation, and before starting to optimize the model architecture and learning hyperparameters, the dataset has to be partitioned into a development set and an independent test set. This framework mirrors the common supervised learning, where a model trained on a training set undergoes evaluation on new data points derived from a similar distribution to ensure model generalization. Ensuring a similar distribution between the two sets involves considering various characteristics such as MRI scanner parameters and/or phenotypic information. In this study, greater emphasis was placed on aligning the distributions in terms of phenotypic information, while distributions of MRI scanners and acquisition parameters were not taken into account for dataset splitting.

Age emerges as a crucial feature during dataset splitting due to considerable FC changes observed from early development to old age [230, 231]. Therefore, ensuring a similar age distribution between the two sets minimizes potential age-related biases, preventing the model from learning specificities from data with differing age distributions in the independent test set. Moreover, significant FC differences have been reported between female and male individuals both in TD subjects [232, 233, 234] and in

Table 4.2: Dataset split into development and independent test sets. The development set comprises 680 subjects utilized for model development and optimization. The independent test set comprises 191 subjects employed to assess the generalization capability of the developed model.

Set	Sites	Sample Size	ASD (%) (ASD/TD)	Age mean (range)	Sex proportion (per group, and total)
Development set	CALTECH CMU KKI MAX_MUN NYU OHSU OLIN SBL SDSU STANFORD UCLA UM USM	680	45.88% (312/368)	16.93 (6.47 - 58.00)	ASD: 12.82% TD: 20.11% Total: 16.76%
Independent test set	LEUVEN PITT TRINITY YALE	191	47.64% (91/100)	16.99 (7.00 - 35.02)	ASD: 15.38% TD: 16.00% Total: 15.71%

individuals with ASD [235, 236, 237]. In this way, sex is another crucial aspect to consider during dataset partitioning.

As previously mentioned the ABIDE dataset is composed of 17 distinct acquisition sites with varying sample sizes. Considering this, to split the dataset as illustrated in Table 4.2, three steps were considered. Firstly, the two-sample Kolmogorov-Smirnov (KS2) test statistics was used to measure the maximum distance between the age distribution of each acquisition site-specific subset with the subset composed of all other acquisition sites. Secondly, sites with lower KS2 statistics, indicating lesser differences, were chosen to constitute the independent test set, until it reached 20% of the entire dataset. Lastly, the *sex* proportion was evaluated using the Chi-Squared test to ensure that the proportion between the development and independent test sets did not exhibit significant differences.

4.3.2 Initial graph construction

The success of GNN models depends on the quality of the graph data [144]. As mentioned the data from ABIDE is not inherently conformed to a graph structure. Therefore, there is a need to conform this data in a graph. GAT layers, as detailed in Section 2.3.5.2, automatically compute the attention weights that model the importance of a node u to a node v based on their respective node features. These attention weights, instrumental in generating node embeddings by weighing each neighbor's influence during the aggregation step from each node v neighborhood, effectively act as edge weights. Nevertheless, despite this functionality, such layers still require an initial graph as input, as they are unable to deduce the graph structure from disconnected instances (i.e. feature vectors). Therefore, an initial graph had to be built.

All the subjects in the dataset will be utilized to construct the initial population graph, where each node represents a subject and each edge denotes the similarity between the corresponding subjects. As discussed in Section 2.3.2, it is necessary to define both a feature matrix $\mathbf{X}$ with the respective feature vectors $\mathbf{x}(v)$ of each subject v (i.e., node features), and the graph connectivity, represented by the adjacency matrix $\mathbf{A}$. The adjacency matrix encapsulates the existing edges and their respective weights, which characterize the relationship between every pair of connected subjects. Moreover, no directional/causal relationship exists between the subjects; instead, the connections represent only similarity. Therefore,

the generated graph is an undirected weighted graph.

4.3.2.1 Node features

Each node (i.e, subject) is represented by the respective functional connectivity data. Following the approach introduced by Parisot *et al.*[17], each 111×111 ROI-to-ROI FC matrix was converted into an upper triangular matrix. This transformation involved discarding the lower half of the matrix along with its main diagonal, resulting in a vectorized FC matrix comprising 6105 elements ($(111 \times 111 - 111)/2 = 6105$). Due to the high dimensionality of these feature vectors, feature selection was carried out. Parisot and colleagues [17] compared various feature selection methods and dimensionalities for their GCN model. They found the best performance of their GCN using recursive feature elimination (RFE) with a ridge classifier for feature selection, reducing the dimensionality of the feature vector to 2000. The same RFE approach with a ridge classifier was employed in this work to reduce the dimensionality of the vectorized FC matrices to 2000 features. It is important to note that only subjects belonging to the respective training set were utilized for feature selection. Subsequently, all feature vectors were transformed considering the 2000 features selected.

4.3.2.2 Graph edges

The graph connectivity models the relationships among subjects and forms the basis for the message-passing mechanism of GNN, which disseminates messages across the graph. Consequently, it must be meticulously crafted to faithfully capture the interactions between feature vectors.

To address the heterogeneity inherent in ASD, integrating phenotypic information alongside FC data within the graph has shown promising outcomes (refer to Section 3 for further details). When establishing the initial graph connectivity, the aim is to delineate a precise neighborhood for each node to optimize the performance of the initial message-passing operation. Therefore, the computation of edge weights must consider information that effectively elucidates similarities within FC data and/or similarities among subjects' labels. Given the significant heterogeneity within ABIDE stemming from variations in scanners and acquisition protocols, subjects from the same acquisition site are inherently more comparable. Consequently, the acquisition site was included as non-imaging information in the computation of the edge weights. Moreover, considering the reported sex and age-group differences in individuals with ASD, subjects of the same sex and age-group are more akin, and these two factors were incorporated into the computation of edge weights. Beyond that, to effectively delineate the similarity within FC data, a measure of the similarity of FC was also used to compute the edge weights, as detailed by the following equation:

$$A(v,u) = Sim(S_v, S_u) \sum_{h=1}^{H} \gamma(F_h(v), F_h(u)) \tag{4.1}$$

where A is the adjacency matrix, $Sim(S_v, S_u)$ is the similarity between the FC feature vectors of every pair of subjects (S_v, S_u), and γ is a measure of distance between the two subjects considering the non-imaging phenotypic factors F_h, where h represents each phenotypic information in the set $H = \{site, age, sex\}$. The similarity between FC feature vectors was given by the gaussian kernel as follows:

$$Sim(S_v, S_u) = exp\left(-\frac{[\rho(\mathbf{x}(v), \mathbf{x}(u))]^2}{2\sigma^2}\right) \tag{4.2}$$

where ρ is the correlation distance, and σ is the kernel width. On the other hand, γ is defined differently for categorical (e.g., subjects' sex and acquisition site) and quantitative variables (e.g., age). For the

Table 4.3: The features of the generated graphs are outlined, providing information on the number of nodes and edges for both the full graph, which encompasses the entire dataset, and the development set, a restricted subgraph derived from the development set. Specifically, for the full graph, the counts of edges before and after edge pruning, where edges with weights below 1.1 were removed, are presented.

Graph	Number of nodes	Number of edges
Full	871	before pruning: 378885 after pruning: 59847
Development	680	39834

former, an edge weight is increased when the two subjects being compared have the same sex and/or share the same acquisition site. For the latter, the following function was considered:

$$\gamma(F_h(v), F_h(u)) = \begin{cases} 1 & if \quad |F_h(v) - F_h(u)| < \theta \\ 0 & otherwise. \end{cases} \tag{4.3}$$

where, if the difference between the age of the two subjects in comparison is lower than a specified threshold θ (set to 2), the edge weight is increased by one unit. This strategy follows the approach of [17]. However, instead of solely considering factors such as *sex* and *site*, *age* was also incorporated as additional phenotypic information to compute the initial edge weights. The final edge weights range from zero to three, reflecting the variability of $Sim(S_v, S_u)$ between zero and one, and $\gamma(F_h(v), F_h(u))$ between zero (indicating no shared phenotypic factors between a pair of subjects) and three (indicating shared age, sex, and site between a pair of subjects). To ensure that only pairs of subjects exhibiting a significant degree of similarity were linked, edge pruning was applied. This process involved removing edges with weights lower than 1.1, ensuring that connected subjects shared at least two phenotypic features and had FC similarity (as given by equation 4.2) higher than 0.55. This process increases the graph sparsity, known to improve the overall performance of GNNs while reducing the computational costs of training a GNN on a highly connected graph [144]. The number of nodes and edges for the full graph before and after the edge pruning are presented in Table 4.3.

In the following section, the optimization procedure undertaken will be discussed. As the initial graph has both the development and the independent test sets, and the model development and optimization are supposed to be conducted within the development set, a constrained subgraph of the initial graph relative to the development set was used in this step. The constrained subgraph is from now on referred to as the development subgraph and its features are outlined in Table 4.3.

4.3.3 GNN optimization

The prediction task undertaken in this study was framed as a binary node classification problem, with the aim of diagnosing ASD and TD individuals. To address this task, a GNN architecture was developed by employing GAT layers for graph representation learning (GRL) (as introduced in Section 2.3.3), i.e. to learn the representation of each subject node, which are then utilized for mapping to the target labels (ASD, and TD). Model development and optimization were conducted in the development subgraph, within a transductive learning setting. In this setting, the entire development subgraph was fed into the model, where all nodes and respective edges were used to generate new node embeddings, yet only a subset of nodes was labeled (the training set of the development subgraph). The GNN was implemented in the open source ML framework PyTorch [238] and using the PyTorch Geometric (PyG) [239] library, a specified library for deep learning on graphs, that has several methods available, such as the GATv2

layers.

To evaluate model performance and select the optimal configuration, stratified 10-fold cross-validation was employed on the development subgraph. This technique involved dividing the dataset into 10 sets of equal size, with nine sets utilized for training and one for validation during each iteration. This process was repeated 10 times, with each set serving as the validation set once. Stratification ensures that each fold maintains consistent proportions of observations across specific categorical values, such as class outcome labels. Through rigorous model comparison across different hyperparameter combinations, 10-fold cross-validation facilitates the identification of the model with superior generalization ability.

The hyperparameters under optimization included hyperparameters related to model architecture and model learning. Initially, emphasis was placed on refining the model architecture. The devised architecture was inspired by current GNN models, notably the EV-GCN discussed previously [218]. The developed GNN (depicted in Figure 4.5) comprises stacked GATv2 layers, followed by linear layers serving as a classifier for the prediction in the two classes. Moreover, JK connections [162] were included to avoid the over-smoothing problem, when the architecture becomes deeper (see Section 2.3.6.2 for details on skip connections). The number of GATv2 layers was fixed to four, considering a balance between the depth of the model to learn abstract node representations and the complexity of the model to avoid overfitting problems, while avoiding large computation costs. Each one of these GAT layers was preceded by a dropout layer. The number of hidden units per layer and the activation functions used were considered hyperparameters to be optimized, while the configuration of the classifier was fixed. The classifier comprised two linear layers. The first layer, which contains 256 hidden units, was followed by a ReLU activation, a dropout layer, and a batch normalization layer. Subsequently, an output linear layer with two units was employed, activated by a softmax function, which assigns labels to the respective classes (0: ASD, 1: TD). The use of multiple attention heads was tested, with various combinations. Nevertheless, it was found to not enhance the performance of the developed model, and therefore this strategy was abandoned.

The training process involved several key decisions. Firstly, the loss function selected was the cross-entropy function, augmented with an L2 penalty term whose value was tuned during model optimization. Initialization of weights and biases followed the Kaiming He initialization method [240]. The ADAM optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1 \times 10^{-8}$) was utilized for learning rate optimization, with its initial value fine-tuned along the number of epochs and dropout rate. Furthermore, mini-batch training was undertaken by the neighbor sampling scheme as outlined in Section 2.3.6.3. Since the architecture consisted of only four GATv2 layers, the computational graphs generated by this method were not excessively large. To guarantee this, the number of neighbors for each layer was set to {25, 10, 8, 4} for the first, second, third, and fourth layers, respectively. Additionally, this method aided in graph regularization by pruning the edges that linked nodes that were not included in the neighborhood when constructing computation graphs for each node. The batch size remained constant at 64. For a more detailed theoretical discussion and description of the training parameters and methods employed, please refer to Section 2.3.

Each architectural hyperparameter (i.e., number of hidden units per GAT layer, and activation function) was individually fine-tuned, altering one of the hyperparameters while keeping the other hyperparameters constant. In contrast, learning hyperparameters (i.e., learning rate, L2 parameter, number of epochs, and dropout rate) were tested with various combinations, involving changes to more than one hyperparameter while keeping the others fixed, such as the learning rate and the L2 parameter that influence each other. The initial model's hyperparameters, as outlined in Table 4.4, were established based on

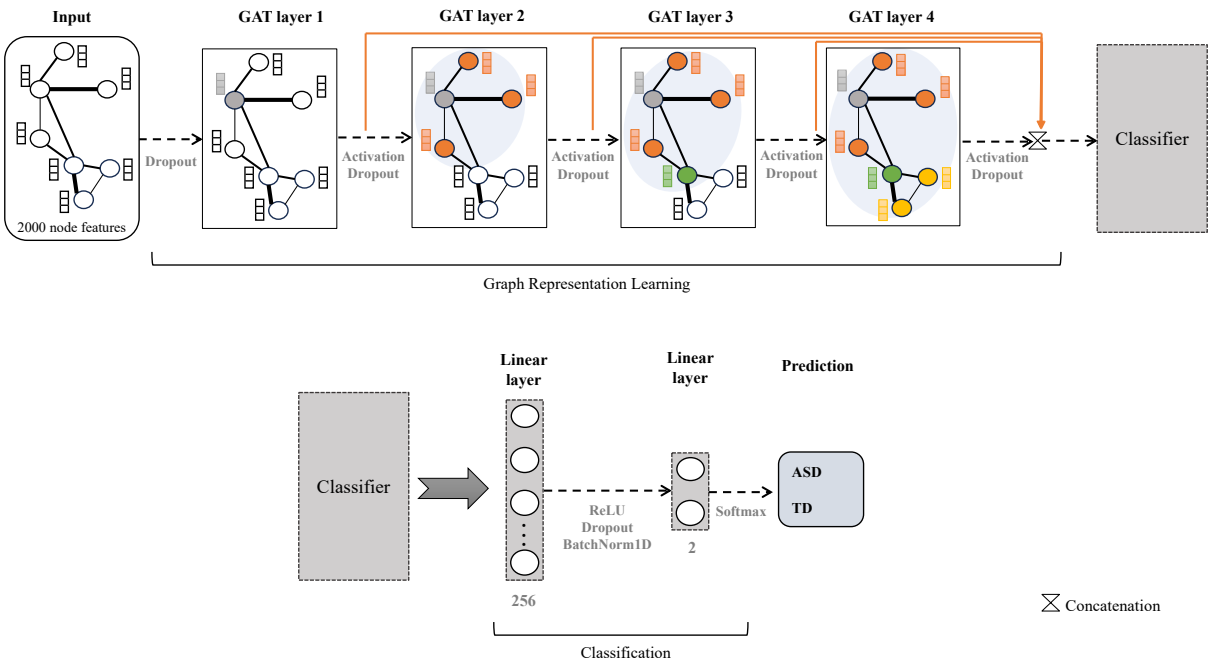


Figure 4.5: Architecture of the model developed. The model receives a graph as input in which each node has 2000 node features, and consists of four GATv2 layers. The node embeddings generated by these layers serve as inputs to a classifier, which comprises one linear layer activated by a ReLU, followed by an output (linear) layer activated by a softmax function for class prediction. Note that while the number of hidden units in the GATv2 layers remains undetermined, those in the classifier are fixed.

commonly employed values in the relevant literature. Several models were tested during hyperparameter optimization as detailed in Table 4.4. To assess model performance, various metrics were employed, including accuracy, AUC-ROC, F1-score, sensitivity/recall, and specificity. These metrics are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.4}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{4.5}$$

$$Specificity = \frac{TN}{TN+FP} \tag{4.6}$$

$$Precision = \frac{TP}{TP+FP} \tag{4.7}$$

$$F1\text{-}score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{4.8}$$

where, true positives (TP) represent correctly identified positive instances, while true negatives (TN) indicate accurate negative classifications. False positives (FP) denote negative instances incorrectly classified as positive, and false negatives (FN) capture positive instances misclassified as negative. Accuracy measures the percentage of correctly classified instances but can be biased in imbalanced datasets. Sensitivity and specificity assess the model's performance in positive and negative instances, respectively. Precision indicates the ratio of correctly identified positive instances. The F1-score represents the harmonic mean between recall and precision. The ROC curve graphically represents the trade-off between recall and specificity and is summarized by the area under the receiver operating characteristic curve

Table 4.4: Tested models during hyperparameter optimization. The hyperparameters under optimization include the activation function, the number of hidden units per GAT layer, the learning rate, the L2 parameter, the number of epochs, the dropout rate, and the use or not of early stopping.

Model	Activation function	Number of hidden units	Learning rate	L2 parameter	Epochs	Dropout
1	ReLU	64	0.001	5.0×10^{-5}	500	0.2
2	PReLU ($\alpha = 0.25$)	64	0.001	5.0×10^{-5}	500	0.2
3	PReLU ($\alpha = 0.25$)	32	0.001	5.0×10^{-5}	500	0.2
4	PReLU ($\alpha = 0.25$)	128	0.001	5.0×10^{-5}	500	0.2
5	PReLU ($\alpha = 0.25$)	64	0.0001	5.0×10^{-4}	500	0.2
6	PReLU ($\alpha = 0.25$)	64	0.0001	5.0×10^{-3}	500	0.2
7	PReLU ($\alpha = 0.25$)	64	0.0001	5.0×10^{-3}	1000	0.2
8	PReLU ($\alpha = 0.25$)	64	0.0001	5.0×10^{-2}	1000	0.2
9	PReLU ($\alpha = 0.25$)	64	0.0001	5.0×10^{-2}	1000	0.4
10	PReLU ($\alpha = 0.25$)	64	0.0001	5.0×10^{-2}	1000	0.6

(AUC), which measures the classifier's ability to distinguish between positive and negative class.

When assessing each model to identify the optimized version, priority was given to metrics that elucidate the model's performance in discriminating between the two classes. These primary measures include accuracy, F1-score, and AUC-ROC. Additionally, secondary metrics evaluating the model's performance in specific aspects of the data, such as sensitivity, and specificity, were also considered. These secondary measures served to resolve ties between models, ensuring the selection of models that strike a balance between sensitivity and specificity. This approach aims to achieve a final model capable of confidently identifying both ASD and TD individuals. For all metrics the average score across the 10 models resulting from the 10-fold cross-validation were reported. To assess overfitting, both the average training accuracy and validation accuracy across the 10 folds were reported. In this way, the model with the best generalization capacity was selected.

Throughout the hyperparameter optimization phase, the loss and accuracy metrics computed for the validation set exhibited considerable oscillations, persisting even after testing numerous models with diverse hyperparameter combinations (a subset of which is outlined in Table 4.4). Although these oscillations diminished notably, they did not cease entirely. To address this issue, early stopping was introduced as a final optimization step within each fold of the cross-validation process to the best-performing model of Table 4.4. Unlike the naive approach detailed in Section 2.3.1, a refined criterion was adopted to preserve the best-performing model based on validation accuracy (see Algorithm 1). Specifically, the criterion mandates that the model's validation accuracy must not deviate by more than 0.05 from the validation accuracy of the preceding five epochs, measured in terms of normalized difference. This process, however, is initiated only after the $400th$ epoch, allowing sufficient time for the model parameters to capture relevant features for the prediction task. The objective of this approach is to stabilize the learning process, ensuring that the model achieved in each fold has converged as effectively as possible. By employing early stopping, the performance of the model is independent of the specified number of epochs. By setting a suitable large number of epochs (e.g., 1000) in tandem with the early stopping algorithm, a balance is struck between achieving model convergence and preventing performance degradation due to excessive training.

After optimizing the model, its performance in the given prediction task was compared to state-of-

Algorithm 1 Early Stopping Algorithm

```
best_validation_accuracy ← 0
validation_accuracy ← accuracy on the validation set for the current epoch.
if validation_accuracy > best_validation_accuracy and epoch > 400 then
                                ▷ Calculate changes in validation accuracy over the last five epochs.
    validation_accuracy_changes ← [abs(validation_accuracy − acc)/acc
        for acc in validation_accuracies[epoch − 5 : epoch]].
    maximum_change ← maximum(validation_accuracy_changes)
    if maximum_change ≤ 0.05 then
                                ▷ Update best validation accuracy and other metrics.
        best_validation_accuracy ← validation_accuracy
                                ▷ Save other performance metrics.
    end if
end if
```

the-art models discussed in Section 3.3. This comparison provided crucial insights into the effectiveness of the developed model. Comparison models were chosen based on the availability of their code and their performance. Specifically, the comparison included the EV-GCN [218] and the LG-GNN [221], both recognized as top-performing models. However, there are several critical considerations to address. Both models adopt different approaches to model training. In the EV-GCN [241], the best-performing model in terms of validation accuracy is selected from each fold of the 10-fold cross-validation starting from the $9th$ epoch onwards. Conversely, the LG-GNN [242] employs a similar approach but selects the best model from the $50th$ epoch onwards. The main issue arises when the model with the highest validation accuracy is selected early in the training process, often during the initial epochs when the loss function and validation accuracy values are still fluctuating significantly. This indicates that the model's performance has not yet stabilized, resulting in the learning of unstable weights and biases, potentially leading to erratic model behavior and unpredictable performance [132]. As previously mentioned, both loss and accuracy metrics computed during model training exhibited considerable oscillations, a phenomenon also noted in the training of the state-of-the-art models used for comparison. Furthermore, for several folds, the models were selected from the early epochs, a practice we believe may not be ideal. However, to ensure a fair comparison, the models were evaluated using both the framework employed in this study, which includes the early stopping criteria described (Algorithm 1), and the framework utilized by them. In both comparison frameworks, all models utilized the subjects from the development set, and the performance metrics mentioned earlier were assessed using a 10-fold cross-validation. Excluding the number of epochs, all other hyperparameters were consistent with those proposed in the respective papers.

Our optimized model, initially developed and fine-tuned within a transductive learning setting, was subsequently subjected to training and evaluation within an inductive setting. While these settings exhibit distinct characteristics and present unique optimization challenges, we chose to conduct the initial optimization process within a transductive setting due to its prevalence in current literature, as highlighted previously. By subsequently training and evaluating the model in an inductive learning environment, we aimed to demonstrate its adaptability and generalization capabilities. We acknowledge that conducting a comprehensive optimization process within an inductive learning framework could potentially yield performance improvements when evaluating the model within the same context. However, this task has been deferred to future work. Instead, we employed the meticulously optimized model derived from the transductive setting for training and evaluation within the inductive setting, which will be explained in

the next subsubsection.

4.3.4 GNN testing

The top-performing model, identified as the one with hyperparameters yielding the most generalizable performance on the development data, was further assessed in an inductive learning setting to gauge its ability to generalize to unseen data. An inductive learning setting was devised to individually evaluate the model's performance on each test subject, simulating the conditions expected in real-world (clinical) applications. The model was trained on the entire development subgraph, referred to as the training graph in this context. After training, the model was tested on the independent test set. In this framework, instead of integrating all subjects from the independent test set into the graph, each test subject was individually incorporated into the training graph. This approach enabled the model to exploit the relational information with the training subjects to whom the test subject being predicted is linked while excluding connections with other test subjects. Note that no two test subjects are included together in the graph. In this way, the model received a distinct graph for each subject composed of the training graph augmented with the node respective to the test subject being predicted. Figure 4.6 is a representation of a graph like this. Subsequently, the results of the evaluation on the test set were then used to study the explainability of the predictions of the devised model.

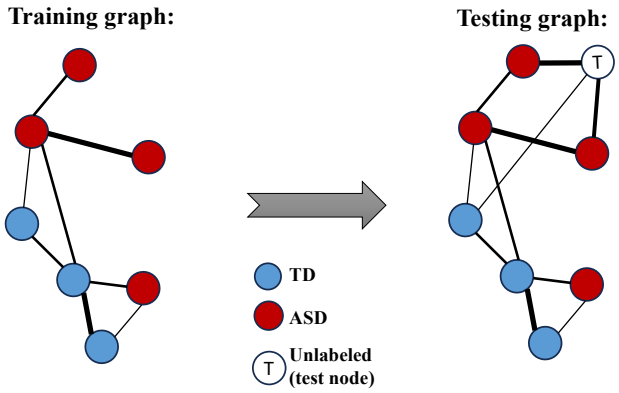


Figure 4.6: Illustration of the evaluation procedure. The model undergoes training on the training graph and is subsequently assessed on a test graph formed by augmenting the training graph with subject T, which is the subject being predicted. The model's performance is evaluated across 191 test graphs, each corresponding to one subject within the independent test set.

4.3.5 Prediction interpretation

After developing and testing the GNN model, an analysis was conducted to explain its predictions, focusing on how effectively it integrates both structural and node feature information while scrutinizing the interpretability of the explanations produced by the XAI method. Ultimately, the XAI results were compared with FC findings in the ASD literature.

In exploring FC abnormalities in individuals with ASD using GNNs, it is essential to utilize a graph XAI method that provides explanations with node feature information, especially considering that each subject is represented by their vectorized FC matrix in the node classification task. Accordingly, the chosen method for this purpose was the GNNExplainer [172] (refer to section 2.3.7 to a theoretical description), which is post-hoc and model-agnostic, being directly applicable to the developed GNN model.

Explanations were generated for the predictions made by the model on the independent test set, as this represents the external validation of the model's performance, ensuring that the insights derived from the explanations are applicable to unseen data and generalizable beyond the training set. Furthermore, the focus was on the true positives (TP), where FC features identified by the method were those utilized by the GNN model to distinguish ASD individuals correctly, potentially serving as biomarkers for the disorder.

Given that the GNNExplainer utilizes gradient descent for optimization, determining appropriate hyperparameter values is essential, such as learning rate and number of epochs. These values were chosen based on PyTorch Geometric tutorials and documentation, to a learning rate of 0.001 and number of epochs of 300. The explanations of the GNNExplainer implementation provided by the PyTorch Geometric library consist of visual representations of the subgraph identified in the computation graph of the target node as most important for explaining the prediction of the respective individual, and a bar plot illustrating the importance values of the k most pertinent node features for that particular prediction. Notice that this method delineates the most influential node features not only for the specific node under scrutiny but also across all nodes in the selected subgraph. Subsequently, the selected node features may vary for each of these nodes, and the displayed values on the bar plot are computed by the sum of the mutual information attributed to the node features across all nodes in the subgraph, including the target node.

A global analysis was conducted to identify consistently influential FC connections across all the TP subjects. The global analysis included computation of the mean node feature importance, and extraction of 30 most important features. Additionally, to determine if specific features were consistently among the 30 most important for predicting each individual, the frequency with which each feature appeared among the top 30 most important features across all TP subjects was assessed.

During the extraction of the most important node features, the disparity in mean importance among the features (i.e. FC pairs) was found to be minimal. Therefore, a two-step criterion was established to identify the most crucial node features. First, features were required to be among the top 10 based on their mean importance score. Second, they needed to demonstrate consistency by appearing among the 30 most important features for at least 20 of the true positive instances. This approach ensured that the selected features possessed high overall importance while also exhibiting consistent relevance across the correctly classified ASD subjects. The identified features were scrutinized and compared with existing literature on alterations in FC associated with ASD. Additionally, to assess the impact of these features on the predictions, an ablation study was conducted. This included removing groups of features and observing the resulting changes in model performance.

Chapter 5

Results and Discussion

This chapter presents and discusses the findings from the experiments detailed in Chapter 4. Firstly the development and optimization of the GNN model is presented and discussed, along with a comparative analysis with current proposed models. Secondly, to study the model generalization and robustness, the GNN model is assessed on the independent test set. Lastly, an analysis and interpretation of the predictions generated by the developed GNN are provided.

5.1 GNN optimization

The development and optimization of the GNN model involved fine-tuning hyperparameters associated with both its architecture and learning process. This optimization was conducted using stratified 10-fold cross-validation to identify the model with the highest generalization capacity. As previously mentioned some hyperparameters values were predefined, including setting the number of GAT layers to four, with a subsequent classifier composed by two linear layers, with 256 hidden units and two output units, respectively, tailored for the binary classification task as illustrated in Figure 4.5. Consequently, the hyperparameters under scrutiny for determining the optimal model architecture included the number of hidden units in the GAT layers and their corresponding activation functions. Subsequently, hyperparameters associated with the learning process of the selected model architecture, including the learning rate, L2 parameter, number of epochs, and dropout rate were optimized. Various combinations of the mentioned hyperparameters were tested, with 10 specific configurations detailed in Table 4.4. The results of this hyperparameter tuning process are summarized in Table 5.1.

Model 1 serves as the baseline, employing hyperparameters drawn from existing literature. Model 2 introduces a modification in the activation function of the GAT layers, transitioning from ReLU in model 1 to PReLU in model 2, with the initial value of the slope of the function for values less than 0 (α) set to the default value (0.25). Despite both models exhibiting signs of overfitting, as evidenced by the disparity between mean train accuracy and mean validation accuracy, model 2 has a higher validation accuracy. Moreover, all evaluation metrics have improved with the employment of the PReLU function. Beyond the validation accuracy, the AUC, and F1-score have improved considerably, and the values of sensitivity and specificity are more balanced. The latter reveals that model 2 is capturing patterns that identify both classes (ASD, and TD), and not favoring one of them, or neither always reporting a positive or negative prediction. Consequently, PReLU was included as the activation of the GAT layers.

In models 3 and 4 the number of hidden units in the GAT layers was lowered from 64 to 32, and increased to 128, respectively, while maintaining the rest of the hyperparameters. In both cases, the overall performance decreased, with smaller values of validation accuracy, AUC and F1-score. As a

Table 5.1: Results of performance metrics obtained during the optimization process for the binary node classification task (ASD vs TD). The reported values represent the mean across a stratified 10-fold cross-validation. The model highlighted in orange bold denotes the best-performing model (9), while the model highlighted in bold represents the performance of that model after applying the early stopping.

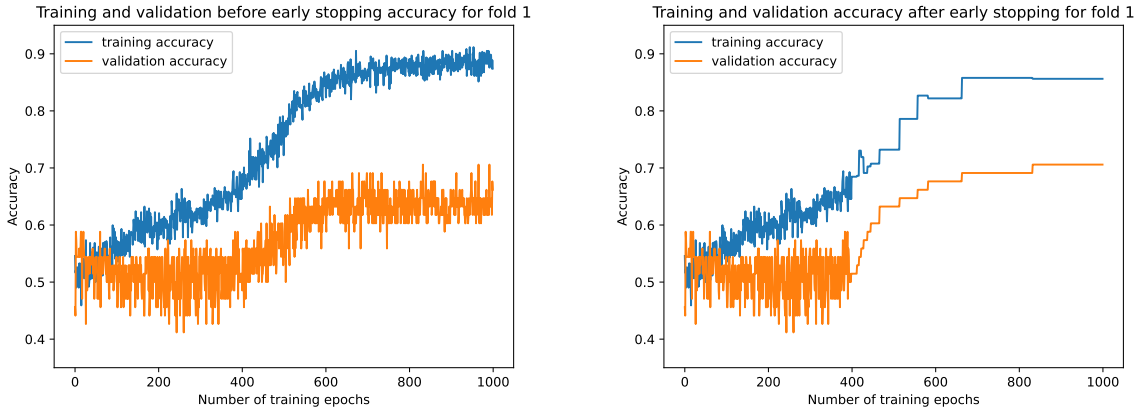
Model	Train Accuracy	Validation Accuracy	AUC	F1-score	Sensitivity	Specificity
1	97.42	62.94	62.23	54.71	53.26	71.21
2	97.65	66.18	70.61	62.41	63.18	68.75
3	97.58	64.26	63.40	56.34	52.89	73.91
4	97.78	65.88	65.07	58.50	54.85	75.29
5	95.59	64.56	71.34	59.11	58.68	69.53
6	95.98	65.59	71.69	61.92	62.87	67.94
7	97.61	66.91	71.20	62.71	63.20	70.11
8	94.71	67.35	72.15	64.58	66.41	68.21
9	**88.66**	**67.35**	**72.15**	**61.47**	**60.62**	**73.09**
10	79.77	66.03	72.68	59.46	57.87	73.21
			Early stopping			
9*	**86.78**	**69.85**	**73.11**	**64.92**	**66.48**	**72.90**

consequence, the number of hidden units was kept at 64.

The best combination of the learning rate, L2 parameter, and number of epochs, for the rest of the hyperparameters set, was achieved in model 8, specifically with the learning rate equal to 0.0001, L2 parameter set to 5.00×10^{-2}, for 1000 epochs. Comparing model 2 with model 8, the training accuracy declined, while validation accuracy, AUC, and F1-score exhibited improvements. Furthermore, the difference between the sensitivity and specificity decreased. Despite these advancements, model 8 still displayed signs of overfitting, evident in the substantial discrepancy between train and validation accuracies. To address this, regularization was enhanced by elevating the dropout rate in models 9 and 10. Although a dropout rate of 0.6 achieved the goal of approximating the values of accuracy in the training and validation sets, it affected considerably other metrics, such as the F1-score and the balance between sensitivity and specificity. Comparing models 8 and 9, with dropout rates of 0.2 and 0.4, respectively, model 9 demonstrated superior overall performance. Although exhibiting a slightly lower F1-score and a wider gap between sensitivity and specificity compared to model 8, model 9 showcased reduced overfitting tendencies. Given the detrimental consequences of an overfitted model, particularly the possible learning of noise from the training data that impacts the generalizability of the model, model 9 was chosen for further analysis.

Throughout the hyperparameter optimization process, the loss and accuracy for both the training and validation sets were monitored by plotting them. During the optimization process for all the hyperparameter combinations tested, significant fluctuations in accuracy and loss were observed throughout the learning process, particularly evident in the validation set. This behavior is illustrated in Figure 5.1a, respective to the first fold of the 10-fold cross-validation process, showing the training and validation accuracy values plotted over the course of 1000 epochs for model 9. As observed, the model is learning, with overall increases of the train and validation accuracies, while oscillating considerably from epoch to epoch of training. Therefore, to mitigate the instability in the learning process, early stopping was implemented using Algorithm 1. This criterion ensures that the saved model is obtained after a predefined sufficient large number of epochs (400) and that the best validation accuracy achieved is not significantly different from the validation accuracy achieved in its vicinity, reducing the likelihood of the saved model capturing noise from the training data. As a result, the learning process became more stable, as demonstrated in Figure 5.1b, from epoch 400 onwards. Beyond stabilizing the learning process, the

application of early stopping enhanced the overall performance of the model, including lower training accuracy and higher mean validation accuracy, overfitting less, and increased AUC and F1-score with a smaller difference between sensitivity and specificity.



(a) Before early stopping (Model 9)

(b) After early stopping (Model 9*)

Figure 5.1: Illustration of the training and validation accuracies before and after the early stopping for model 9. In a) significant oscillations in both training and validation accuracies during the learning process are evident. However, in b), the impact of early stopping on stabilizing the learning process becomes apparent, particularly after epoch 400. Despite the fluctuations observed, there is an overall improvement in both metrics throughout the training process, a phenomenon observed consistently across all 10 folds.

Once the learning process is stabilized, it is more likely that the model has acquired reliable parameters capable of generalizing to new data points. Before testing the generalization capability of the developed model, to evaluate its effectiveness for the binary node classification task (ASD vs TD), a comparison was conducted with two state-of-the-art models (EV-GCN and LG-GNN) and discussed in the following subsection.

5.1.1 Comparison with literature

As outlined in the Methods Section, discrepancies exist in the learning and subsequent evaluation methodologies between the proposed approach and the compared models. Specifically, both compared models adopted an early stopping strategy based uniquely on validation accuracy, with EV-GCN stopping as early as $9th$ epoch and LG-GNN after the $50th$ epoch. While this approach could be appropriate if the model's performance consistently improved during training, it might not be suitable for models exhibiting training oscillations. The mentioned accuracy oscillations observed in the training of the developed model were also present in the training of the compared state-of-the-art models. This phenomenon can lead to models being saved during early epochs where the oscillations are more pronounced. Models in this situation have not yet fully adapted to the training data and may learn unstable weights and biases, potentially leading to erratic behavior and unpredictable performance when applied to new data. Figure 5.2 exemplifies these training oscillations in the EV-GCN, along with the early epoch (epoch 37) when the best model was saved for subsequent computation of mean performance across the 10-fold cross-validation. Moreover, the learning process of the EV-GCN presents other problem. The overall learning curve of the EV-GCN model does not exhibit improvement throughout the training process, as the validation accuracy consistently oscillates around similar values, while the training accuracy is nearly optimal. It is patent a scenario of low bias and high variance (overfitting), where the model excessively fits the training data but struggles to generalize effectively to the validation set data. Such a situation

may arise from factors including noisy data, reduced data sample, or suboptimal hyperparameters. In scenarios of overfitting, early stopping may potentially yield stable parameters, particularly in cases where validation accuracy initially improves but subsequently declines. However, by epoch 37, the learned parameters likely capture noise from the training data, given its early stage in the training process. It is highly unlikely that the model has learned parameters reflecting essential features of the data for the classification task. This pattern is consistent across all the 10 folds. It is worth noting how misleading the mean performance metrics provided as the outcome of the evaluation process for such a model can be. Considering from the validation accuracy curve that the validation accuracy oscillates around similar values, an early stopping criterion should aim to achieve stable performance during the training process.

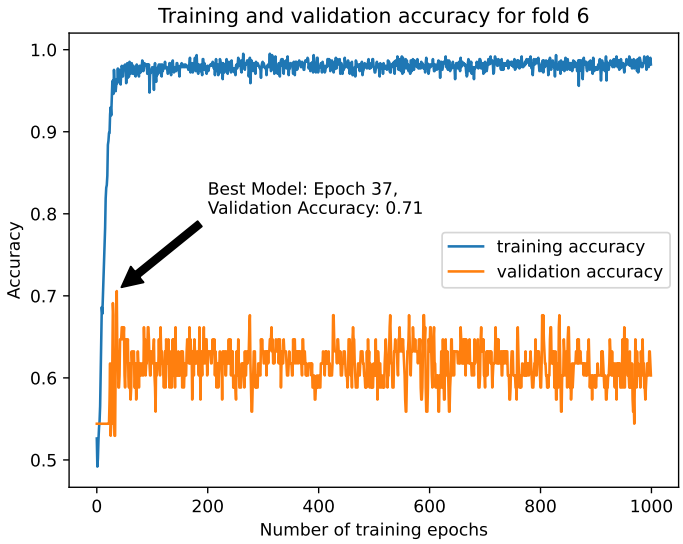


Figure 5.2: Illustration of the EV-GCN model learning process. The oscillations of the validation accuracy across training are observed, and, importantly, it is showcased the epoch and the corresponding validation accuracy value representing the best model achieved based on the framework employed by the authors.

Figure 5.3 illustrates the training and validation accuracy fluctuations observed during the learning process of the LG-GNN model. Compared to the proposed GNN and EV-GCN models, LG-GNN exhibits significantly larger oscillations. Validation accuracy values fluctuate between near 0.9 and as low as 0.5 within consecutive epochs, even after extensive training for almost 800 epochs. In fact the best-performing model is selected at epoch 131, with a validation accuracy of 1.0, despite having a validation accuracy below 0.4 just a few epochs earlier. The validation accuracy value of the selected model is subsequently employed to compute the final mean evaluation metrics across the 10-fold cross-validation.

Considering the differences between the proposed approach and the strategies employed by the compared models, two separate comparisons were conducted: one adhering to our approach and other aligning with each of their respective approaches. Both comparisons were conducted with the same data for all models. The development subgraph composed of 680 subjects described in the Methods section (Section 4.3.2) was used, and the performance metrics values reported correspond to the mean values across the 10-fold cross-validation.

The results from the comparison between the proposed model and the EV-GCN following the framework used by the authors of the EV-GCN, where the best-performing model is chosen solely based on validation accuracy from epoch 9 onwards, are detailed in Table 5.2. Similarly, the performance values relative to the comparison between the developed GNN and LG-GNN network, employing the framework utilized by the authors of LG-GNN, selecting the best-performing model solely based on validation

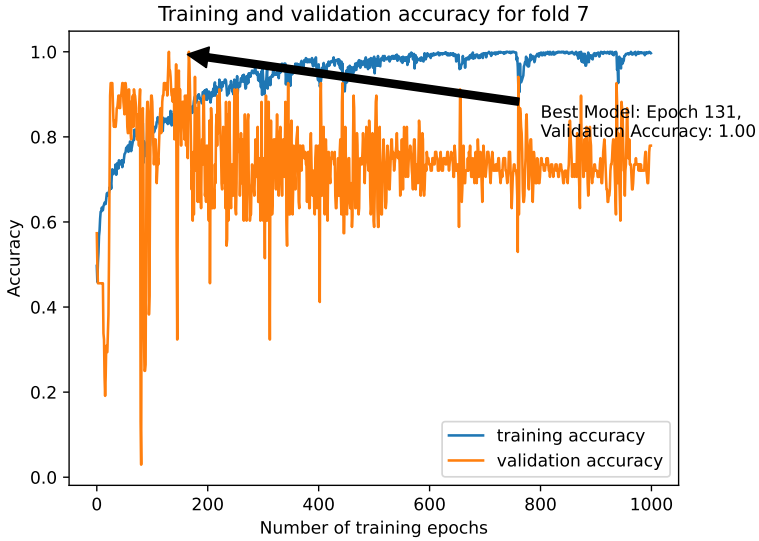


Figure 5.3: Illustration of the LG-GNN model learning process. The oscillations of the validation accuracy across training are observed, and, importantly, it is showcased the epoch and the corresponding validation accuracy value representing the best model achieved based on the framework employed by the authors.

accuracy from epoch 50 onwards, are illustrated in Table 5.3. Furthermore, Table 5.4 presents the comparison results among the three models resorting to the proposed approach where the best-performing model is selected based on validation accuracy, ensuring that it does not deviate by more than 0.05 from the validation accuracy of the preceding five epochs, measured in terms of normalized difference, beginning from epoch 400.

Comparing Tables 5.2 and 5.4 allows for an assessment of how the performance of the developed model and EV-GCN changed between the approach employed in the EV-GCN study and the methodology proposed in this work. The majority of metrics present a general trend of decreasing values from the former approach to the latter, observed for both models, i.e. the implementation of a more restrictive early stopping criterion yielded lower performance metrics, as was expected. This trend underscores the impact discussed above regarding the selection of best-performing models from initial epochs, particularly for models experiencing significant oscillations in validation accuracy during training. Importantly, these are potentially more stable results that better reflect the model's true capacity. On the other hand, certain patterns are common, such as the fact that EV-GCN overfit more in both scenarios and that both models present higher specificity than sensitivity, which suggests that both can more easily capture relevant features for the negative class (i.e. TD) than for the positive class (i.e. ASD). The fact that the developed model outperforms the EV-GCN for the approach used by the EV-GCN authors highlights how the validation accuracy for the developed model oscillates through the model. However, it is crucial to note that even with the more restrictive early stopping criteria, the developed model continues to outperform the EV-GCN model.

On the other hand, LG-GNN surpasses the developed GNN when following the LG-GNN approach. This is potentially due to the selection of a model that will not generalize correctly to new data. This concern arises from the observation that, when the approach proposed in this work was applied, the validation accuracy fluctuated so much in at least one fold of the LG-GNN cross-validation that no model that met the early stopping criteria could be found, and subsequently the final evaluation could not be computed. This is not surprising when considering the oscillations in validation accuracy in Figure 5.3. To explore if with less restrictive early stopping criteria the model could run, the early stopping criterion

Table 5.2: Comparison of performance metrics between the developed GNN model and the literature EV-GCN model in the binary node classification task (ASD vs TD), adopting the methodology proposed by the EV-GCN authors, i.e. the best performing model based on validation accuracy from the 9th epoch onwards was selected. The reported values denote the mean performance across a stratified 10-fold cross-validation, using the development subgraph (680 subjects).

Model	Training Accuracy	Validation Accuracy	AUC	F1-score	Sensitivity	Specificity
Ours	83.79	73.97	73.81	68.63	65.10	81.53
EV-GCN	98.01	72.50	72.88	68.87	67.27	76.85

Table 5.3: Comparison of performance metrics between the developed GNN model and the literature LG-GNN model in the binary node classification task (ASD vs TD), adopting the methodology proposed by the LG-GNN authors, i.e. the best-performing model based on validation accuracy from the 50th epoch onwards was selected. The reported values denote the mean performance across a stratified 10-fold cross-validation, using the development subgraph (680 subjects).

Model	Training Accuracy	Validation Accuracy	AUC	F1-score	Sensitivity	Specificity
Ours	82.32	73.09	73.68	70.54	70.82	74.95
LG-GNN	85.10	82.79	85.39	83.55	80.67	85.31

was adjusted to allow the validation accuracy of the best-performing model to deviate by up to 0.1 (as opposed to 0.05) from the validation accuracy of the preceding five epochs, starting at the same epoch (i.e. 400). Despite this adjustment, there remained at least one fold in which no model satisfying the criteria could be identified. On the other hand, if the normalized difference criterion is disregarded, with the best-performing model being selected for each fold from epoch 400 onwards, the model still presents high-performance values, which was expected considering that the model is significantly oscillating between low and high validation accuracy values even after 400 epochs. The outcomes of both attempts are presented in Table 5.5.

In this way, the developed GNN appears to be more capable than both comparison models in generating stable and reliable performance metrics when using the proposed early stopping criteria. Specifically, this early stopping strategy mitigated the issue of oscillations, resulting in more consistent and stable results, almost akin to disregarding them altogether. Nevertheless, the oscillations reported should be avoided altogether in future works to ensure that the models learn the most stable parameters possible, thus enabling robust generalization to unseen data.

The top-performing model identified through this approach underwent evaluation in the independent test set to assess its generalizability. However, it is important to note that this model was identified using a transductive learning setting, which may not be optimal for clinical applications due to the need for retraining every time a new instance is added to the graph. Therefore, an inductive setting is adopted for evaluation, which is discussed in the following section.

Table 5.4: Comparison of performance metrics among the developed GNN model, the EV-GCN, and the LG-GNN models in the binary node classification task (ASD vs TD). The evaluation adopts the proposed approach, where the best-performing model is selected based on its validation accuracy, that must not deviate by more than 0.05 from the validation accuracy of the preceding five epochs, measured in terms of normalized difference, starting on epoch 400. The reported values represent the mean performance across a stratified 10-fold cross-validation, using the development subgraph (680 subjects).

Model	Training Accuracy	Validation Accuracy	AUC	F1-score	Sensitivity	Specificity
Ours	86.78	69.85	73.11	64.92	66.48	72.90
EV-GCN	98.48	67.79	70.87	62.57	61.30	73.41
LG-GNN	No stable value	No stable value	No stable value	No stable value	No stable value	No stable value

Table 5.5: Results of performance metrics for the LG-GNN model under two different attempts to ameliorate the early stopping criteria. In the first attempt, the criterion allowed the validation accuracy of the best-performing model to deviate by up to 0.1 (compared to the original 0.05) from the validation accuracy of the preceding five epochs, starting at epoch 400. In the second attempt, the criterion based on the normalized difference was disregarded, and the best-performing model was chosen for each fold from epoch 400 onwards. The reported values represent the mean performance across a stratified 10-fold cross-validation, using the development subgraph (680 subjects).

Attempt	Training Accuracy	Validation Accuracy	AUC	F1-score	Sensitivity	Specificity
1	No stable value	No stable value	No stable value	No stable value	No stable value	No stable value
2	93.97	81.18	81.30	81.55	78.78	84.08

5.2 GNN testing

In this part of the work, the goal was to test how the developed model generalized to new unseen data. For this, the developed GNN model (Model 9 in Table 5.1) was trained on the entire development subgraph (680 subjects), and tested in an independent test set (191 subjects) in an inductive learning setting, as described in Section 4.3.4. For comparative analysis with other methods in the literature, the model with early stopping (Model 9* in Table 5.1) was utilized due to its superior performance during validation. However, for the evaluation on the independent test set, the entire development subgraph (from now on referred to as training subgraph) was used for training, leaving no separate validation set to apply the early stopping criterion. Consequently, the second-best performing model (Model 9 in Table 5.1), trained without early stopping, was employed for the final evaluation on the independent test set.

Table 5.6 displays the values of the performance metrics achieved, including the training accuracy in the training subgraph, and the test accuracy, AUC, F1-score, sensitivity, and specificity values achieved in the classification of the subjects included in the independent test set. The respective confusion matrix is also depicted in Figure 5.4 for a direct analysis of the distribution of true positives, true negatives, false positives and false negatives instances.

Table 5.6: Performance metrics attained by the final model in the binary node classification (ASD vs TD). Training accuracy reflects accuracy within the training subgraph, while test accuracy, AUC, F1-score, sensitivity, and specificity denote values acquired from the independent test set, under an inductive setting.

Training Accuracy	Test Accuracy	AUC	F1-score	Sensitivity	Specificity
85.59	62.30	69.98	52.00	42.86	80.00

The training and test accuracies values reveal a significant gap between them. While the training accuracy of 85.59% suggests that the model performs relatively well on the training data, the drop in the accuracy value when evaluated on unseen data (test accuracy) suggests that the model is overfitting the training data — a concerning indication of diminished performance on unseen data. Despite a moderate AUC value indicating a reasonable discriminative ability in classifying both classes, the F1-score, representing a balance between precision and sensitivity, is notably low at 52.00%. This underscores the model's struggle in accurately classifying positive instances (ASD individuals), corroborated by the low sensitivity achieved. Conversely, the model exhibits a high specificity value, indicating relatively better performance in identifying TD individuals, which may explain the significant difference between AUC and F1-score, given that F1-score is solely based on true positives and false positives, without considering true negatives. The observations from the confusion matrix further support these findings, indicating that while the model effectively captures patterns for classifying the negative class, it encounters challenges in capturing features for the positive class.

The observed performance metrics and training behavior of the developed model suggest that the

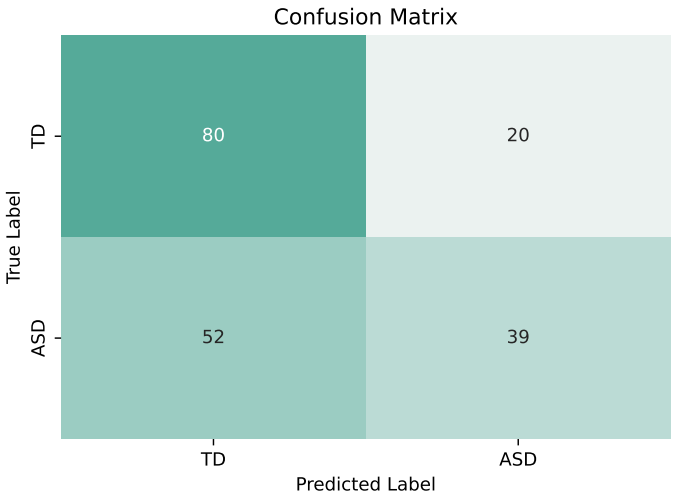


Figure 5.4: Confusion matrix attained by the final model for the binary classification (ASD vs TD), under an inductive setting.

model may be suffering from the curse of dimensionality. The curse of dimensionality refers to a group of phenomena that arise when analyzing data in high-dimensional spaces, where the volume of the space increases exponentially with the number of dimensions, making the available data sparse [243]. This sparsity makes it difficult for ML models to find meaningful patterns and can lead to issues such as overfitting and poor generalization.

Specifically, the model was trained on a graph comprising 680 subjects, each represented by a 2000-dimensional feature vector that corresponds to the subject's vectorized FC matrix. In this scenario, the feature dimensionality surpasses the number of training examples. Dealing with high-dimensional node features increases the complexity of GNNs' aggregations from neighboring nodes, making the learning process inefficient and prone to capturing noise rather than meaningful information. This is exacerbated when the graph has a complex structure with several edges, as in this case (39834 edges). Consequently, the GNN needs to perform various computations, each involving high-dimensional vectors, which further compounds the problem. This can lead to the considerable fluctuations observed in validation accuracy during the 10-fold cross-validation, indicating model sensitivity to noise in the data, with the model struggling to find a stable solution. Moreover, in high-dimensional spaces, if the model has enough capacity to fit the noise instead of the underlying patterns in the training data the model will overfit the training data. This overfitting is notorious in the disparity between the mean training accuracy (88.66%) and the mean validation accuracy (67.35%). Overfitting results in poor generalization, as the model captures noise and specific patterns in the training data that has difficulties in generalizing effectively to unseen data. This was evident in the model's evaluation on the independent test set, where the performance significantly declined, with a training accuracy of 85.59% on the training subgraph (680 subjects) but a substantially lower test accuracy of 62.30% on the independent test (191 subjects).

Addressing the challenge of high dimensionality can involve solutions such as further reducing feature dimensionality or employing regularization techniques such as sampling techniques like neighbor sampling, or early stopping, and edge pruning, which have also been applied. Additionally, increasing the dataset size could be beneficial. This could be achieved by acquiring more data or incorporating data from other databases, such as ABIDE II. However, in the latter should be ensured data variability is minimized by selecting acquisitions compatible in terms of acquisition parameters and applying consistent pre-processing steps. Expanding the dataset in this manner would help cover the entire feature space,

allowing the model to capture underlying patterns more effectively and achieve better performance in classification tasks.

On the other hand, the high heterogeneity of the disorder and of the ABIDE dataset have also contributed to the difficulty of the model to perform well on new data, as the model has difficulties to accommodate all the ASD complexities that come from its heterogeneity. This is particularly noticeable in the existent difference between the (low) sensitivity and the (high) specificity observed in both the 10-fold cross-validation and the evaluation of the model on the independent test set. Regarding the disorder, atypical disconnectivities in the whole brain have been associated with the multiple ASD phenotypes, considering the entire ASD spectrum, which is further influenced by co-occurring comorbidities, such as ADHD. Additionally, sex differences also play a role, as the manifestation of ASD can vary significantly between female and male individuals, potentially contributing to the observed heterogeneity. Moreover, some subjects are taking medication that can alter brain FC [244]. The model appears to struggle in capturing this high variance, thus obtaining a low sensitivity value. Additionally, the ABIDE dataset comprises data from multiple sites with diverse acquisition protocols. The training set and the independent test set devised are composed of data from different sites. Consequently, the model may struggle to accommodate the high variance arising from the ABIDE's heterogeneity, as the captured features during model training may not represent the entirety of the variance of the dataset, which leads to the model performance to decrease when applied to the independent test set. Notably, prior literature has reported significant differences in classification performance metrics across the sites comprising the ABIDE dataset [189, 192]. By failing to capture the full intricacies of data and subsequently, of the disorder, the model struggles to accurately classify new ASD subjects, resulting in lower sensitivity during the evaluation stage. Additionally, it is important to recognize that the optimization measures are derived from a learning setting that differs from the one used during model evaluation. All these processes may have influenced the disparities observed in performance metrics between values computed via cross-validation and those derived from evaluation on the test set.

The disparity between sensitivity and specificity values achieved by DL models for the classification of ASD individuals resorting to the ABIDE dataset is frequently reported in the literature [189, 192]. While a high specificity value is crucial for ensuring accurate diagnosis of healthy individuals given their prevalence in the population, the low sensitivity is not ideal for the subsequent analysis, as the goal was to evaluate true positive instances to uncover FC features inherent to ASD through XAI analysis. Nevertheless, given the inherent limitations and challenges in this work, a perfect model is difficult to obtain. This reflects broader trends in DL methods applied to ASD diagnosis using fMRI, particularly within the ABIDE, where specificity often outweighs sensitivity, reflecting the dataset's inherent heterogeneity and the complexity of the disorder.

5.3 Prediction interpretation

After testing the final model developed on the independent test set, an XAI analysis was performed to understand the decision-making process of the developed model for classifying individuals with ASD. The focus on explaining the true positive instances allows for assessing which inter-brain ROI functional connections are the most relevant features for ASD prediction, which can potentially be considered biomarkers for the disorder. The explanations provided in this type of analysis enhance the transparency of the model's decisions, which is particularly important when the model's predictions impact critical decisions, such as in the case of a DL system applied in clinical diagnosis. Since the developed model is not inherently interpretable, it requires the application of an external method. By resorting to the post-hoc

GNNExplainer method, a global analysis was performed to extract relevant features for the classification of ASD across all the individuals correctly identified as having the disorder.

The global analysis started with the extraction of the 30 most important features for the ASD classification, derived from the computation of the mean node feature importance, as depicted in Figure 5.5. The indices correspond to specific FC connections between brain ROIs. Table 5.7 summarizes the top 10 connections that are most relevant for ASD prediction, including their corresponding brain regions and importance scores. The detailed mapping of these features to their corresponding brain ROIs is provided in Table A.1.

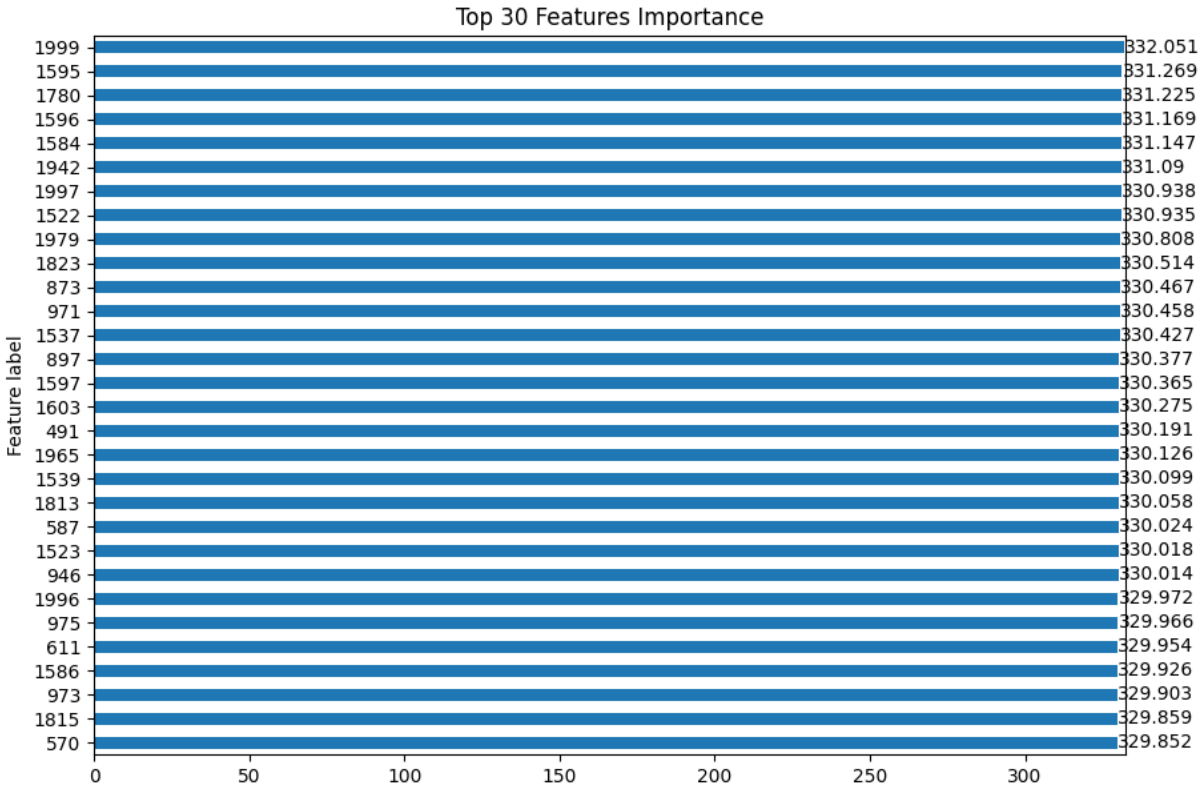


Figure 5.5: Top 30 most important FC connections and their respective importance score for ASD classification with the developed GNN model. Each feature label corresponds to one specific functional connection. The importance score for each connection represents the average feature importance across all subjects correctly predicted as ASD by the model. Feature importance was derived from the GNNExplainer, which quantifies the contribution of each FC connection to the model's predictions.

Given the minimal difference in mean importance scores across features, the consistency of specific FC pairs in contributing to predictions for each correctly classified ASD individual was assessed. Therefore, the frequency of each FC connection appearing among the 30 most important features across all TP instances was computed. Considering that there are 39 TP instances, the focus was on features with a frequency of at least 20. The respective results are presented in Figure 5.6.

From Figure 5.6 one observes that there are at least 11 features that were consistently among the 30 most important features for the identification of ASD individuals by the developed GNN model. However, not all of them are also in the top 10 features with higher importance scores.

With the criterion devised to identify the key functional connections contributing to the classification of ASD, features 1999, 1596, 1780, 1942, and 1979 were marked. Remember that the devised criterion ensured that these connections had high overall importance and demonstrated consistent relevance across all correctly predicted ASD individuals. Essentially, it extracted features common to both Table 5.7 (top

Table 5.7: Top 10 most important FC connections, and respective brain ROIs for ASD classification with the developed GNN model. Each row shows the functional connection label, the corresponding brain ROIs involved, and the associated importance score. The importance score reflects the average contribution of each connection to the model's predictions across all subjects correctly classified as ASD by the model.

Feature Label	Brain ROI 1	Brain ROI 2	Importance Score
1999	Left Supracalcarine Cortex	Left Occipital Pole	332.051
1595	Right Supracalcarine Cortex	Left Heschl's Gyrus	331.269
1780	Left Middle Temporal Gyrus; temporooccipital part	Left Inferior Temporal Gyrus; posterior division	331.225
1596	Right Supracalcarine Cortex	Left Supracalcarine Cortex	331.169
1584	Right Supracalcarine Cortex	Left Inferior Temporal Gyrus; posterior division	331.147
1942	Left Precuneous Cortex	Left Occipital Pole	331.09
1997	Left Heschl's Gyrus	Left Occipital Pole	330.938
1522	Right Parietal Operculum Cortex	Left Middle Temporal Gyrus; temporooccipital part	330.935
1979	Left Temporal Occipital Fusiform Cortex	Left Planum Polare	330.808
1823	Left Inferior Temporal Gyrus; temporooccipital part	Left Heschl's Gyrus	330.514

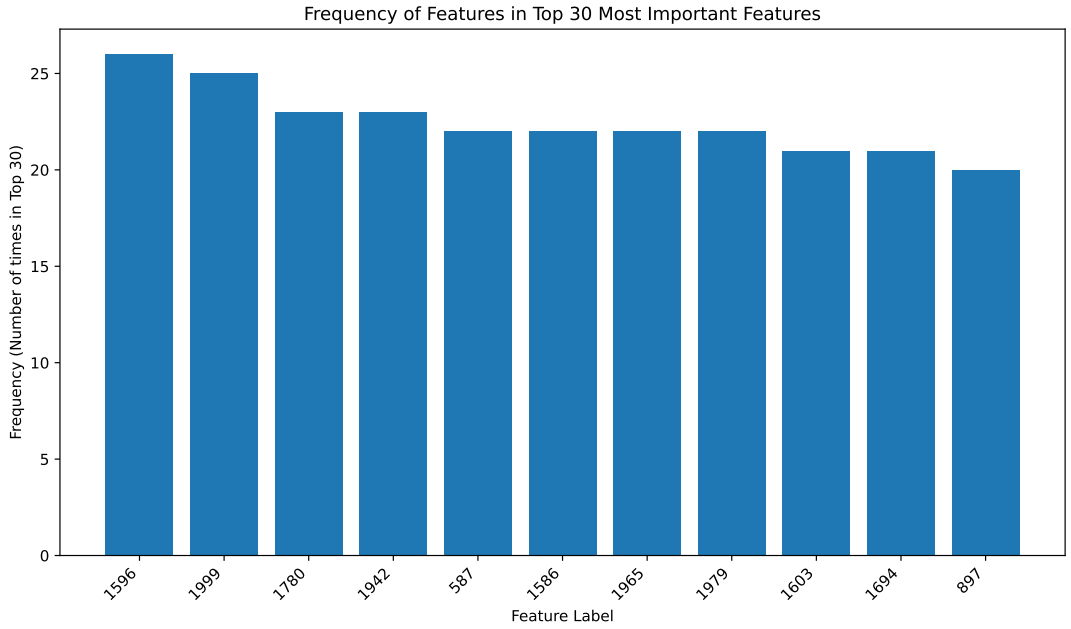


Figure 5.6: The FC connections identified as being among the 30 most important features for at least 20 out of the 39 subjects correctly classified as having ASD by the developed GNN model. These features represent the most consistent FC connections contributing to the model's predictions across the correctly classified ASD cases.

10 most important features) and Figure 5.6 (consistently important features) (refer to Table 5.7 or Table A.1 for the associated brain ROIs). Figure 5.7 depicts the connectogram highlighting these key FC pairs and their corresponding importance scores (refer to Table A.2 for a complete list of the corresponding brain ROIs).

To assess the impact of the identified key functional connections on the model's predictions an ablation study was conducted. In this analysis, the features corresponding to these key connections were removed from the node features for all subjects and the developed GNN model was evaluated on the subjects in the independent test set. It is important to note that while the GNN model is evaluated on the

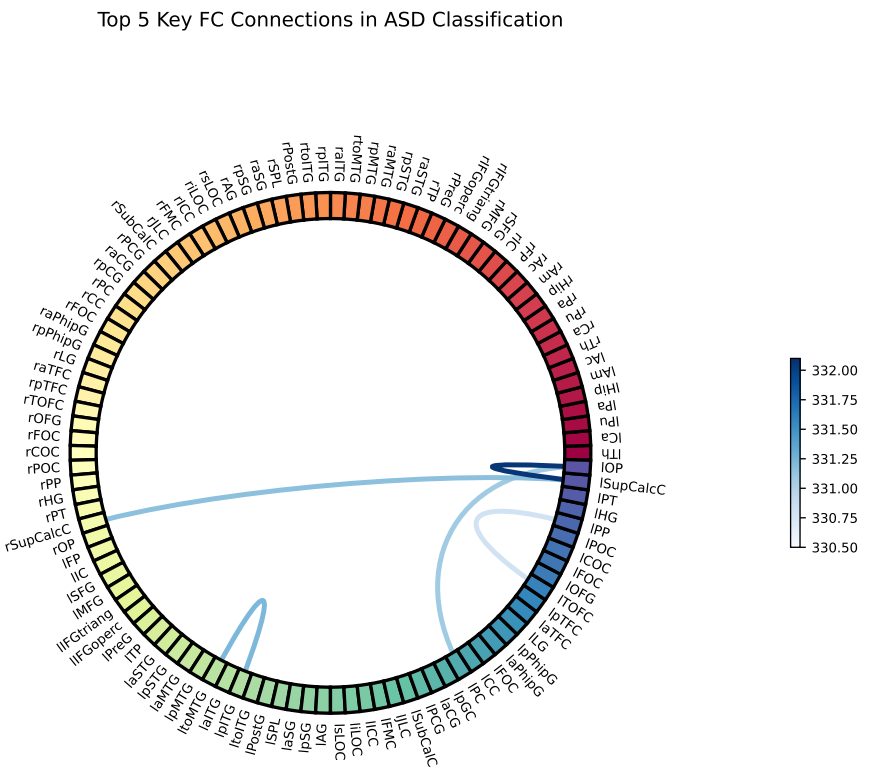


Figure 5.7: Connectogram of the top key FC connections for the classification of ASD. These key features possess high overall importance and exhibit consistent relevance across all correctly classified ASD individuals. Each node in the connectogram represents a brain ROI from the HO anatomical atlas. The color intensity corresponds to the feature importance score, with darker blue indicating higher importance.

independent test set for performance assessment, it still utilizes information from the training subgraph during the prediction process. Therefore, it is essential to remove them from the node features of all subjects. Figure 5.8 presents the obtained performance compared to the final performance achieved by the developed model with all features included. The ablation of key features results in a slight overall decrease in performance, particularly in the specificity value. The reduction in specificity associated with constant sensitivity value suggests that the removed features were not crucial for correctly classifying ASD individuals, contrary to what was expected, but were significant for accurately identifying TD individuals. The model may be using other sets of features that are robust and sufficient for the identification of ASD subjects, or it might have multiple pathways to correctly classify ASD individuals, hence removing a few features does not impact the sensitivity. It is reasonable to infer this given the minimal variation among the importance scores of the different features, as depicted in Figure 5.5. This phenomenon is also related to the high dimensionality of the node features discussed earlier. In high-dimensional spaces, certain features can become redundant or irrelevant, making it challenging for the GNNExplainer to identify the features that are genuinely crucial for the model's performance. On the other hand, removing just five features, compared to the 2000 features present for each subject, may not significantly impact the model performance.

Subsequently, instead of removing only the five key features, two other experiments were performed where 10 features were ablated. In the first experiment were removed the 10 most important features (refer to Table 5.7), which included the five key features along with five additional features. In the

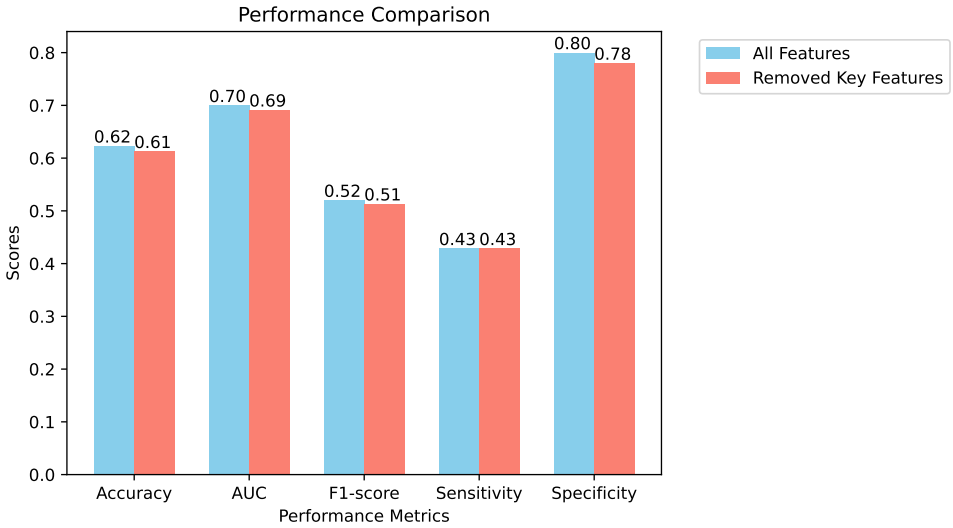


Figure 5.8: Ablation study: performance impact of key FC connections. Comparison between the performance metrics of the final GNN model with all features to the performance after removing the key FC connections identified in the analysis. Both evaluations were conducted on the independent test set.

second experiment the 10 most frequently occurring features among the 30 most important features were removed. Figure 5.9 presents the performance results for both analyses, compared to the final performance achieved by the developed model with all features included. With the ablation of the 10 most important features, the performance decreased, particularly the sensitivity. This indicates that these 10 functional connections are important for capturing ASD specific patterns, as their removal led to a decrease in the model's ability to correctly identify ASD individuals (decrease in sensitivity). Conversely, when the 10 most frequently found among the most important features were removed the sensitivity increased, which suggests that these features are not as critical for identifying ASD specific patterns, and in fact seem to have been capturing noise or redundant information that was affecting sensitivity negatively.

Although the difference in importance scores among the most important features, as computed by GNNExplainer, is minimal, the sensitivity experienced a considerable drop. This suggests that the feature importance calculated by the GNNExplainer more accurately reflects the significance of features for classifying ASD individuals compared to the frequency with which these features appear among the top 30 most important features. These findings suggest that the model relies heavily on specific features for accurate classification, likely capturing essential patterns or characteristics unique to ASD individuals.

Although the XAI analysis successfully identified relevant FC connections for ASD classification, some limitations have to be considered. Firstly, the sample size was small, particularly as the focus was on the true positive instances, which numbered only 39. It is crucial to focus on the correctly identified ASD individuals to pinpoint the FC connections that better characterize the ASD individuals. Secondly, GNNExplainer requires the setting of various parameters associated with gradient descent optimization, such as the learning rate and number of epochs, as well as the regularization of the explanations provided, including compactness parameters like "edge_size" in the PyG library implementation that ensures the most relevant subgraph compactness. While this work does not directly assess the subgraph explanations, the identified subgraph influences the importance values reported by GNNExplainer, as these values are computed by summing the feature importance attributed to the node features across all nodes in the subgraph.

Despite the mentioned limitations, the GNNExplainer provided significant and relevant explanations

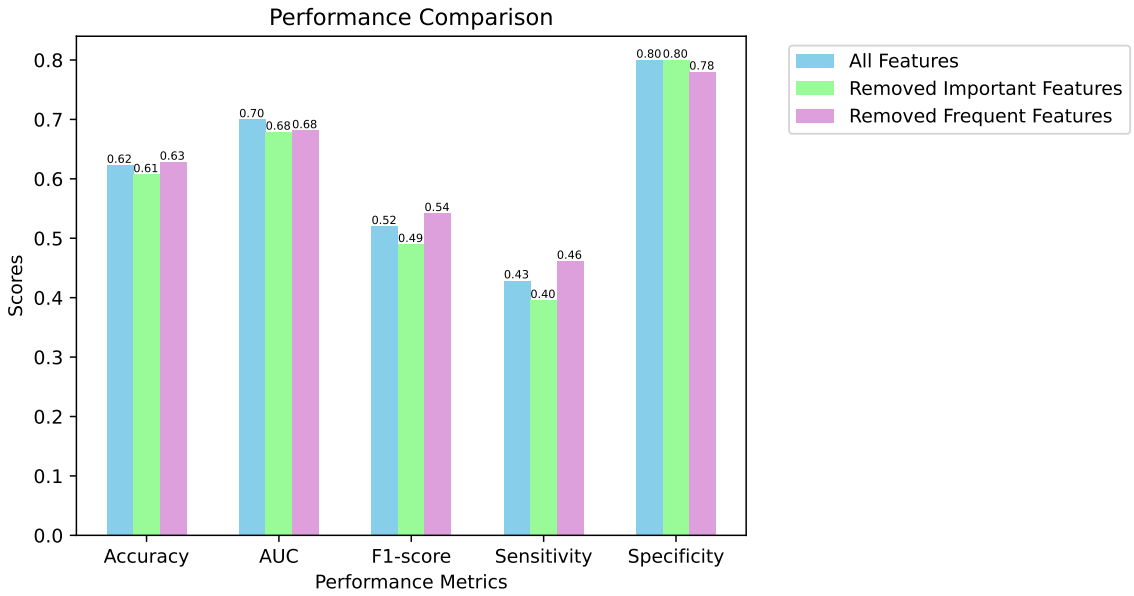


Figure 5.9: Ablation study: the importance of feature strength versus frequency in the model's predictions. Comparison between the performance metrics of the final GNN model with all features to its performance after removing two sets of features: 1. top 10 most important FC connections identified in the analysis; 2. top 10 features that appeared most frequently among the top 30 most important for correctly predicted ASD subjects. All evaluations were conducted on the independent test set.

for the classification of ASD by the developed model, thereby enhancing the transparency of the model's decision-making process. The relevance of the identified FC connections can be assessed by evaluating how they align with broader ASD research and their potential clinical implications. In the next subsection, these connections will be compared against known ASD cognitive impairments (e.g. social communication and interaction), findings from the literature, and differences between ASD individuals and TD controls.

5.3.1 Relevance of Identified FC in ASD

The top 10 FC connections identified by the GNNExplainer as having the highest importance demonstrated the greatest impact on model performance in the ablation studies. Consequently, this subsection will focus on evaluating their significance within the broader context of ASD research and their potential clinical implications.

To determine if the FC values differ between the groups of individuals with ASD and TD subjects, boxplots of the FC values for the 10 most important FC pairs were generated, as depicted in Figure 5.10. For the majority of the FC pairs, the respective FC values are distributed in lower values among ASD individuals than in the TD controls, specifically the pairs: Right Parietal Operculum Cortex - Left Middle Temporal Gyrus (temporooccipital part) (feature 1522); Right Supracalcarine Cortex - Left Inferior Temporal Gyrus (posterior division) (feature 1584); Left Middle Temporal Gyrus (temporooccipital part) - Left Inferior Temporal Gyrus (posterior division) (feature 1780); Left Precuneous Cortex - Left Occipital Pole (feature 1942); Left Temporal Occipital Fusiform Cortex - Left Planum Polare (feature 1979); Left Heschl's Gyrus (includes H1 and H2) - Left Occipital Pole (feature 1997); Left Supracalcarine Cortex - Left Occipital Pole (feature 1999). Conversely, in the following pairs, the FC values appear distributed among higher connectivity values in ASD individuals: Right Supracalcarine Cortex - Left Heschl's Gyrus (feature 1595); Right Supracalcarine Cortex - Left Supracalcarine Cortex (feature 1596); Left Inferior Temporal Gyrus (temporooccipital part) - Left Heschl's Gyrus (includes H1 and

H2) (feature 1823). To test the significance of these observations a one-tailed Mann-Whitney U-test for independent samples was conducted between the distributions of the FC values for each FC pair across the clinical groups. The results of these tests are presented in Table 5.8. Among the observations, only the finding of under-connectivity between the left temporal occipital fusiform cortex and the left planum polare (feature 1979) in ASD individuals relative to TD controls was found statistically significant.

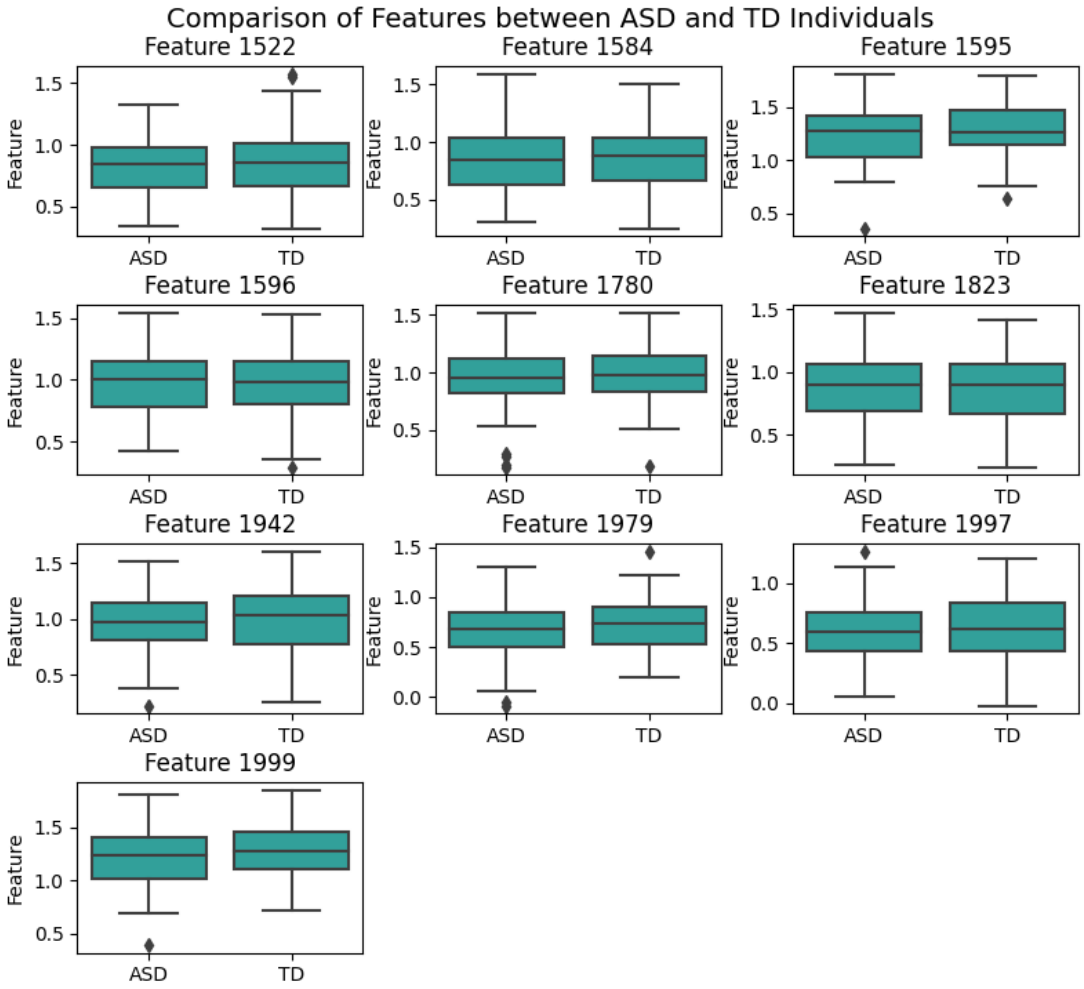


Figure 5.10: Boxplots comparing the distribution of FC values between ASD and TD individuals. The plots focus on the top 10 most important FC pairs identified in the XAI analysis, as indicated by their corresponding feature labels.

The left planum polare is one of three anatomical subregions of the left superior temporal gyrus [245]. The superior temporal gyrus is involved in processing auditory information, including language, and also plays a role in social cognition [246]. Specifically, Boddaert and colleagues found that ASD individuals exhibited reduced activation of the left superior temporal gyrus in response to auditory stimuli [247]. The temporal occipital fusiform cortex is involved in high-level visual processing, such as object recognition, face perception, and reading [248]. One of the most significant and early impairments in ASD is the abnormal visual processing of human faces. Under-connectivity in the lateral area of the fusiform gyrus, known as the fusiform facial area, is a consistent finding in ASD research [249]. This region is considered a crucial node for face processing. Therefore, this functional connection supports tasks requiring the integration of auditory and visual information, such as language processing and the recognition and naming of faces. These functions are particularly important in social communication, one of the core symptoms of ASD. Moreover, this functional connection may play roles in tasks like

Table 5.8: Results of the Mann-Whitney U tests performed to assess the significance of differences in FC values between the ASD subjects and the TD controls. The tests focus on the top 10 most important FC pairs identified in the figure by their corresponding feature labels. The significance level was set to $\alpha = 0.05$ for all statistical tests.

Feature	Hypothesis	U statistic	p-value	Result
1522	H0: ASD $\geq$ TD H1: ASD < TD	4286.0	0.2449	Fail to reject H0
1584	H0: ASD $\geq$ TD H1: ASD < TD	4328.0	0.2808	Fail to reject H0
1595	H0: ASD $\leq$ TD H1: ASD > TD	4251.0	0.7837	Fail to reject H0
1596	H0: ASD $\leq$ TD H1: ASD > TD	4723.0	0.3256	Fail to reject H0
1780	H0: ASD $\geq$ TD H1: ASD < TD	4265.0	0.2280	Fail to reject H0
1823	H0: ASD $\leq$ TD H1: ASD > TD	4623.0	0.4247	Fail to reject H0
1942	H0: ASD $\geq$ TD H1: ASD < TD	4220.0	0.1939	Fail to reject H0
1979	H0: ASD $\geq$ TD H1: ASD < TD	3909.0	0.04662	Reject H0
1997	H0: ASD $\geq$ TD H1: ASD < TD	4234.0	0.2042	Fail to reject H0
1999	H0: ASD $\geq$ TD H1: ASD < TD	4063.0	0.1012	Fail to reject H0

reading and the recognition and naming of objects.

Although the differences in connectivity values for the other nine FC pairs between ASD individuals and TD controls were not statistically significant, their relevance should be evaluated in the context of current ASD knowledge. It is noteworthy that many of the brain ROIs involved in the identified FC pairs are associated with visual processing. These regions include the left supracalcarine cortex, the right supracalcarine cortex, the left middle temporal gyrus, and the left inferior temporal gyrus. These areas are crucial for visual processing, as well as face and object recognition. Individuals with autism often exhibit abnormalities in visual processing, including a heightened ability to detect fine details, known as local processing, which is frequently associated with difficulty in integrating these details into a cohesive global picture [250]. Beyond visual processing, individuals with ASD often face challenges in integrating sensory information from multiple modalities. The right parietal operculum cortex which plays a key role in integrating sensory inputs to form a coherent perception of the environment is part of one of the top 10 FC pairs identified. Impairments in this area can lead to deficits in spatial awareness and the ability to direct attention to relevant stimuli, further contributing to the sensory processing difficulties observed in ASD [251]. Moreover, the left middle temporal gyrus and the left inferior temporal gyrus are critical regions involved in language processing, specifically in semantic processing and the comprehension of written and spoken language [252]. Language impairments are common among ASD individuals, with some being non-verbal or minimally verbal. The Heschl's gyrus, which anatomically includes the primary auditory cortex, is part of three of the ten functional connections identified as important for classifying ASD individuals. This region is crucial for speech processing. Kaku and colleagues [253] have identified a relationship between language processing deficits and abnormal clustering in the Heschl's gyrus. Alterations in this brain ROI can disrupt the flow of language information through the auditory network, further impacting language abilities. Furthermore, as discussed in Section 3.2 of the

state-of-the-art, the precuneus cortex is a critical node of the default mode network and is linked to Theory of Mind [199]. Wang *et al.* reported an association between under-connectivity in the precuneus and the difficulty ASD individuals experience in inferring the perspectives of others. Beyond its connection to ToM, the precuneus is involved in self-referential thinking, including autobiographical memory and self-reflection [254]. Additionally, it plays a role in episodic memory retrieval (information about unique personal experiences) and visuo-spatial imagery. Impairments in self-referential thinking, as well as autobiographical and episodic memory, have been observed in autistic individuals, contributing to their social impairments [255, 256, 257].

Despite the majority of the functional connections identified as most important for predicting ASD did not differ significantly between ASD and TD subjects, a comparison with the literature on functional brain alterations associated with ASD reveals that all identified brain ROIs that compose these functional connections have been extensively studied in ASD research and are involved in the functional disruptions that lead to core cognitive impairments of this neurodevelopmental condition. In this way, attesting the overall relevance of the ASD functional disruptions found by the GNNExplainer.

Chapter 6

Conclusions and Future Work

The multifaceted nature of ASD presents a significant challenge for diagnosis. Current methods heavily rely on behavioral observations, which can be subjective and fail to capture the various symptom domains and presentations associated with the spectrum of the disorder. ML models also struggle to successfully capture the inherent heterogeneity of ASD. GNNs offer a promising approach by easily incorporating relational information. This dissertation explored the potential of these networks to address the multifaceted landscape of the disorder. A GNN model was developed to classify ASD individuals and a XAI approach was applied to highlight functional connections critical for the detection of autism.

An initial literature review was carried out at the beginning of the project. This review included studies on the use of AI methods applied to ASD using neuroimaging data, research on the atypical FC observed in ASD individuals, and ML methods applied to ASD diagnosis using FC data, including GNN approaches. This comprehensive review was crucial in defining an appropriate workflow to achieve the project's objectives.

Given that the rs-fMRI pre-processing is outside the scope of this work and the primary goal was to develop a GNN model, the data used had already been pre-processed using a standardized pipeline. This approach ensures a fair comparison with state-of-the-art models. For the initial population graph, each node was represented by the vectorized FC matrix of each individual. The relationships between nodes were encoded based on similarities in age, sex, and acquisition site, while also considering the correlation between the FC data of the individuals.

The GNN was subsequently developed and optimized to achieve good performance and generalization capacity through a process involving 10-fold cross-validation. The optimization results revealed that the model may be suffering from the curse of dimensionality, evidenced by significant fluctuations in validation accuracy during training. While the use of early stopping mitigated some of these effects, issues such as overfitting persisted. The effects of the high feature dimensionality were found to be transversal to the state-of-the-art models. Despite these challenges, the developed GNN demonstrated superior overall performance. The learning curve indicated that the model was effectively learning and possessed good capacity for application to data beyond the training set.

Subsequently, the model was tested on an independent test set, in a setting that leveraged relational information with the training subjects. Given the overfitting present during model optimization, the model could not generalize perfectly to new data, which can be considered an indirect consequence of the high dimensionality of the data. Notoriously the model demonstrated difficulties in predicting ASD, despite yielding a good performance when classifying TD controls. This indicates that the model could not capture the entire data variance, that arises from the inherent heterogeneity of the disorder and of the ABIDE dataset. The combination of high variance and high dimensionality exacerbated the

model's challenge in accurately predicting ASD. Considering that these challenges are common in state-of-the-art models, subsequent studies should prioritize refining graph data representation, as the model performance is as good as the quality of the data given as input.

Lastly, the XAI analysis was conducted utilizing the GNNExplainer method. Ablation studies were conducted to validate the significance of the features identified by the GNNExplainer in classifying ASD. Discoveries linked to visual processing, particularly in face recognition, as well as language processing and self-referential thinking, were identified. These findings are pertinent to the social and language impairments observed in individuals with ASD.

Considering the exploratory nature of this work, some limitations need to be acknowledged. First, the small dataset size hinders the model's ability to effectively capture the full range of variability present in the data, limiting its ability to generalize to unseen data and increasing the risk of overfitting. Second, the dimensionality reduction performed on the features was insufficient considering the number of data examples. Third, although GAT layers automatically learn edge weights, they are unable to deduce the graph structure from disconnected instances, requiring the construction of an initial graph, which would impair its application in a clinical setting. Fourth, the model was optimized in a transductive setting but evaluated in an inductive setting. While it is unclear whether optimization using inductive learning would result in significant performance improvements, implementing this approach would require substantially more computational resources.

Overall, this work successfully achieved the objectives outlined at the beginning of this dissertation. While it may not have improved the overall ASD diagnosis performance of DL models, it demonstrated an inherent issue in population-based approaches using GNNs for ASD diagnosis, where the high dimensionality of node features hampers their performance. The developed model exhibited improvements in the learning process compared to some state-of-the-art models. Additionally, this work represents one important step towards a computer-aided diagnosis method for ASD by applying an inductive learning setting to these approaches, enabling the model to be applied to entirely new data while effectively leveraging relational information. The XAI analysis demonstrated to be promising by identifying common FC patterns associated with ASD that are well-documented in the literature, attesting the potential to use XAI methods alongside graph-based approaches to uncover ASD biomarkers, while also contributing to the future translation of these models into clinical practice. It is worth noting that while this work focused on ASD diagnosis, the proposed GNN-based approach has the potential to be adapted and applied to other neurological disorders that can be represented using connectivity matrices derived from structural or functional neuroimaging data.

This work raises new challenges that could be addressed in future research. Reducing feature dimensionality to a value lower than the number of training examples is crucial for improving model performance. This can be achieved using techniques like autoencoders, which compress data into lower dimensions. On the other hand, employing a brain atlas based on RSNs rather than an anatomical atlas can help to address not only the high dimensionality but also the challenges posed by the inherent heterogeneity of ASD. Usually, an RSN-based atlas splits the brain into fewer regions compared to an anatomical atlas, significantly reducing the number of features and thereby mitigating the curse of dimensionality. Moreover, RSNs represent coherent patterns of brain activity, providing biologically relevant features that are more consistent across individuals. Additionally, including the cerebellum in the analysis holds promise for a more comprehensive understanding of the neural correlates underlying ASD. In addition to be involved in motor function, the cerebellum has been found to be involved in executive function, visuospatial processing, and emotional regulation [258]. A growing body of research suggests that dysfunction in this region may contribute to the cognitive impairments observed in ASD [258]. Fur-

thermore, the integration of complementary functional metrics such as partial correlation, coherence, and entropy, would be able to capture different aspects of brain connectivity, enhancing the model's sensitivity to the subtle and varied manifestations of ASD. However, it is important to manage the increased dimensionality this approach might introduce, making feature reduction techniques essential in this context. Finding optimal graph similarity measures to compute the relation between individuals could reduce the data sparsity in high-dimensional spaces and enhance model performance. This optimization may encompass integrating additional clinical and biological features, such as cognitive test results, handedness, and medication status. Furthermore, developing an adaptive approach for graph structure learning that can compute the graph connectivity from disconnected instances would significantly contribute to a future ASD classification model, both in terms of its performance and its transition to the clinic, as the model would be able to receive new subjects, compute their connectivity to existing subjects, and perform diagnoses based on both individual and relational information.

References

[1] A. P. Association, ed., *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Washington, D.C: American Psychiatric Association, 5th ed., 2013.

[2] L. Francés, J. Quintero, A. Fernández, A. Ruiz, J. Caules, G. Fillon, A. Hervás, and C. V. Soler, "Current state of knowledge on the prevalence of neurodevelopmental disorders in childhood according to the DSM-5: A systematic review in accordance with the PRISMA criteria," *Child and Adolescent Psychiatry and Mental Health*, vol. 16, p. 27, Mar. 2022.

[3] I. Parenti, L. G. Rabaneda, H. Schoen, and G. Novarino, "Neurodevelopmental Disorders: From Genetics to Functional Pathways," *Trends in Neurosciences*, vol. 43, pp. 608–621, Aug. 2020.

[4] R. Bosch, M. Pagerols, C. Rivas, L. Sixto, L. Bricollé, G. Español-Martín, R. Prat, J. A. Ramos-Quiroga, and M. Casas, "Neurodevelopmental disorders among Spanish school-age children: Prevalence and sociodemographic correlates," *Psychological Medicine*, vol. 52, pp. 3062–3072, Oct. 2022.

[5] S. Bölte, S. Girdler, and P. B. Marschik, "The contribution of environmental exposure to the etiology of autism spectrum disorder," *Cellular and Molecular Life Sciences*, vol. 76, pp. 1275–1297, Apr. 2019.

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

[7] L. Zwaigenbaum and M. Penner, "Autism spectrum disorder: Advances in diagnosis and evaluation," *BMJ*, p. k1674, May 2018.

[8] N. Salari, S. Rasoulpoor, S. Rasoulpoor, S. Shohaimi, S. Jafarpour, N. Abdoli, B. Khaledi-Paveh, and M. Mohammadi, "The global prevalence of autism spectrum disorder: A comprehensive systematic review and meta-analysis," *Italian Journal of Pediatrics*, vol. 48, p. 112, July 2022.

[9] A. Masi, M. M. DeMayo, N. Glozier, and A. J. Guastella, "An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options," *Neuroscience Bulletin*, vol. 33, pp. 183–193, Apr. 2017.

[10] M. Stamou, K. M. Streifel, P. E. Goines, and P. J. Lein, "Neuronal connectivity as a convergent target of gene × environment interactions that confer risk for Autism Spectrum Disorders," *Neurotoxicology and Teratology*, vol. 36, pp. 3–16, Mar. 2013.

[11] L. Shen, X. Liu, H. Zhang, J. Lin, C. Feng, and J. Iqbal, "Biomarkers in autism spectrum disorders: Current progress," *Clinica Chimica Acta*, vol. 502, pp. 41–54, Mar. 2020.

[12] J. C. McPartland, "Considerations in biomarker development for neurodevelopmental disorders," *Current opinion in neurology*, vol. 29, pp. 118–122, Apr. 2016.

[13] A. V. S. Buescher, Z. Cidav, M. Knapp, and D. S. Mandell, "Costs of Autism Spectrum Disorders in the United Kingdom and the United States," *JAMA Pediatrics*, vol. 168, pp. 721–728, Aug. 2014.

[14] O. Sporns, "The human connectome: A complex network," *Annals of the New York Academy of Sciences*, vol. 1224, no. 1, pp. 109–125, 2011.

[15] G. S. Dichter, "Functional magnetic resonance imaging of autism spectrum disorders," *Dialogues in Clinical Neuroscience*, vol. 14, pp. 319–351, Sept. 2012.

[16] S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neuroscience & Biobehavioral Reviews*, vol. 74, pp. 58–75, Mar. 2017.

[17] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, "Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease," *Medical Image Analysis*, vol. 48, pp. 117–130, Aug. 2018.

[18] C. Lord, T. S. Brugha, T. Charman, J. Cusack, G. Dumas, T. Frazier, E. J. H. Jones, R. M. Jones, A. Pickles, M. W. State, J. L. Taylor, and J. Veenstra-VanderWeele, "Autism spectrum disorder," *Nature Reviews Disease Primers*, vol. 6, p. 5, Jan. 2020.

[19] D. H. Geschwind, "Autism: Many Genes, Common Pathways?," *Cell*, vol. 135, pp. 391–395, Oct. 2008.

[20] L. Kanner, "Autistic disturbances of affective contact," *Nervous Child*, vol. 2, pp. 217–250, 1943.

[21] H. Asperger, "Die „Autistischen Psychopathen" im Kindesalter," *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 117, pp. 76–136, June 1944.

[22] N. E. Rosen, C. Lord, and F. R. Volkmar, "The Diagnosis of Autism: From Kanner to DSM-III to DSM-5 and Beyond," *Journal of Autism and Developmental Disorders*, vol. 51, no. 12, pp. 4253–4270, 2021.

[23] J. B. Barahona-Corrêa and C. N. Filipe, "A Concise History of Asperger Syndrome: The Short Reign of a Troublesome Diagnosis," *Frontiers in Psychology*, vol. 6, p. 2024, Jan. 2016.

[24] R. Lordan, C. Storni, and C. A. De Benedictis, "Autism Spectrum Disorders: Diagnosis and Treatment," in *Autism Spectrum Disorders* (A. M. Grabrucker, ed.), Brisbane (AU): Exon Publications, 2021.

[25] L. Wing, "Asperger's syndrome: A clinical account," *Psychological Medicine*, vol. 11, pp. 115–129, Feb. 1981.

[26] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM-III*. American Psychiatric Association, 3rd ed., 1980.

[27] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM IV*. 1000 Wilson Boulevard, Arlington, VA 22209: American Psychiatric Association, 4th ed., Jan. 1994.

[28] A. K. Sauer, J. E. Stanton, S. Hans, and A. M. Grabrucker, "Autism Spectrum Disorders: Etiology and Pathology," in *Autism Spectrum Disorders* (A. M. Grabrucker, ed.), Brisbane (AU): Exon Publications, 2021.

[29] R. H. Wozniak, N. B. Leezenbaum, J. B. Northrup, K. L. West, and J. M. Iverson, "The development of autism spectrum disorders: Variability and causal complexity," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 8, p. e1426, Jan. 2017.

[30] M. Fakhoury, "Autistic spectrum disorders: A review of clinical features, theories and diagnosis," *International Journal of Developmental Neuroscience*, vol. 43, pp. 70–77, June 2015.

[31] D. H. Geschwind, "Genetics of autism spectrum disorders," *Trends in Cognitive Sciences*, vol. 15, pp. 409–416, Sept. 2011.

[32] I. Mohammad-Rezazadeh, J. Frohlich, S. K. Loo, and S. S. Jeste, "Brain Connectivity in Autism Spectrum Disorder," *Current opinion in neurology*, vol. 29, pp. 137–147, Apr. 2016.

[33] J. A. Chen, O. Peñagarikano, T. G. Belgard, V. Swarup, and D. H. Geschwind, "The Emerging Picture of Autism Spectrum Disorder: Genetics and Pathology," *Annual Review of Pathology: Mechanisms of Disease*, vol. 10, no. 1, pp. 111–144, 2015.

[34] C. A. Pardo and C. G. Eberhart, "The Neurobiology of Autism," *Brain Pathology*, vol. 17, no. 4, pp. 434–447, 2007.

[35] K. Yenkoyan, A. Grigoryan, K. Fereshetyan, and D. Yepremyan, "Advances in understanding the pathophysiology of autism spectrum disorders," *Behavioural Brain Research*, vol. 331, pp. 92–101, July 2017.

[36] M. K. Belmonte, G. Allen, A. Beckel-Mitchener, L. M. Boulanger, R. A. Carper, and S. J. Webb, "Autism and Abnormal Development of Brain Connectivity," *The Journal of Neuroscience*, vol. 24, pp. 9228–9231, Oct. 2004.

[37] D. Ben Bashat, V. Kronfeld-Duenias, D. A. Zachor, P. M. Ekstein, T. Hendler, R. Tarrasch, A. Even, Y. Levy, and L. Ben Sira, "Accelerated maturation of white matter in young children with autism: A high b value DWI study," *NeuroImage*, vol. 37, pp. 40–47, Aug. 2007.

[38] B. J. Hwang, M. A. Mohamed, and J. R. Brašić, "Molecular imaging of autism spectrum disorder," *International Review of Psychiatry*, vol. 29, pp. 530–554, Nov. 2017.

[39] Y.-H. Pan, N. Wu, and X.-B. Yuan, "Toward a Better Understanding of Neuronal Migration Deficits in Autism Spectrum Disorders," *Frontiers in Cell and Developmental Biology*, vol. 7, p. 205, Sept. 2019.

[40] L. D. Yankowitz, J. D. Herrington, B. E. Yerys, J. A. Pereira, J. Pandey, and R. T. Schultz, "Evidence against the "normalization" prediction of the early brain overgrowth hypothesis of autism," *Molecular Autism*, vol. 11, p. 51, June 2020.

[41] L. Lorenzini, G. van Wingen, and L. Cerliani, "Atypically high influence of subcortical activity on primary sensory regions in autism," *NeuroImage : Clinical*, vol. 32, p. 102839, Oct. 2021.

[42] H. C. Hazlett, M. Poe, G. Gerig, R. G. Smith, J. Provenzale, A. Ross, J. Gilmore, and J. Piven, "Magnetic Resonance Imaging and Head Circumference Study of Brain Size in Autism: Birth Through Age 2 Years," *Archives of General Psychiatry*, vol. 62, pp. 1366–1376, Dec. 2005.

[43] E. Courchesne, K. Pierce, C. M. Schumann, E. Redcay, J. A. Buckwalter, D. P. Kennedy, and J. Morgan, "Mapping Early Brain Development in Autism," *Neuron*, vol. 56, pp. 399–413, Oct. 2007.

[44] E. Courchesne, P. R. Mouton, M. E. Calhoun, K. Semendeferi, C. Ahrens-Barbeau, M. J. Hallet, C. C. Barnes, and K. Pierce, "Neuron Number and Size in Prefrontal Cortex of Children With Autism," *JAMA*, vol. 306, pp. 2001–2010, Nov. 2011.

[45] J. Wegiel, I. Kuchna, K. Nowicki, H. Imaki, J. Wegiel, E. Marchi, S. Y. Ma, A. Chauhan, V. Chauhan, T. W. Bobrowicz, M. de Leon, L. A. S. Louis, I. L. Cohen, E. London, W. T. Brown, and T. Wisniewski, "The neuropathology of autism: Defects of neurogenesis and neuronal migration, and dysplastic changes," *Acta Neuropathologica*, vol. 119, pp. 755–770, June 2010.

[46] S. H. Fatemi, T. D. Folsom, T. J. Reutiman, and P. D. Thuras, "Expression of GABAB Receptors Is Altered in Brains of Subjects with Autism," *The Cerebellum*, vol. 8, pp. 64–69, Mar. 2009.

[47] J. L. R. Rubenstein and M. M. Merzenich, "Model of autism: Increased ratio of excitation/inhibition in key neural systems," *Genes, Brain and Behavior*, vol. 2, no. 5, pp. 255–267, 2003.

[48] A. F. d. C. Hamilton, "Emulation and Mimicry for Social Interaction: A Theoretical Approach to Imitation in Autism," *Quarterly Journal of Experimental Psychology*, vol. 61, pp. 101–115, Jan. 2008.

[49] X. Duan and H. Chen, "Mapping brain functional and structural abnormities in autism spectrum disorder: Moving toward precision treatment," *Psychoradiology*, vol. 2, pp. 78–85, Sept. 2022.

[50] I. Dziobek, S. Fleck, K. Rogers, O. T. Wolf, and A. Convit, "The 'amygdala theory of autism' revisited: Linking structure to behavior," *Neuropsychologia*, vol. 44, pp. 1891–1899, Jan. 2006.

[51] S. Baron-Cohen, H. A. Ring, E. T. Bullmore, S. Wheelwright, C. Ashwin, and S. C. R. Williams, "The amygdala theory of autism," *Neuroscience & Biobehavioral Reviews*, vol. 24, pp. 355–364, May 2000.

[52] S. Wang and X. Li, "A revisit of the amygdala theory of autism: Twenty years after," *Neuropsychologia*, vol. 183, p. 108519, May 2023.

[53] F. R. Volkmar and J. M. Wolf, "When children with autism become adults," *World Psychiatry*, vol. 12, no. 1, pp. 79–80, 2013.

[54] D. Fein, M. Barton, I.-M. Eigsti, E. Kelley, L. Naigles, R. T. Schultz, M. Stevens, M. Helt, A. Orinstein, M. Rosenthal, E. Troyb, and K. Tyson, "Optimal Outcome in Individuals with a History of Autism," *Journal of child psychology and psychiatry, and allied disciplines*, vol. 54, pp. 195–205, Feb. 2013.

[55] M. Careaga, J. Van de Water, and P. Ashwood, "Immune dysfunction in autism: A pathway to treatment," *Neurotherapeutics*, vol. 7, pp. 283–292, July 2010.

[56] R. Grzadzinski, M. Huerta, and C. Lord, "DSM-5 and autism spectrum disorders (ASDs): An opportunity for identifying ASD subtypes," *Molecular Autism*, vol. 4, p. 12, May 2013.

[57] T. Hirvikoski, E. Mittendorfer-Rutz, M. Boman, H. Larsson, P. Lichtenstein, and S. Bölte, "Premature mortality in autism spectrum disorder," *The British Journal of Psychiatry*, vol. 208, pp. 232–238, Mar. 2016.

[58] Y. I. J. Hwang, P. Srasuebkul, K.-R. Foley, S. Arnold, and J. N. Trollor, "Mortality and cause of death of Australians on the autism spectrum," *Autism Research*, vol. 12, no. 5, pp. 806–815, 2019.

[59] O. Le Meur, A. Nebout, M. Cherel, and E. Etchamendy, "From Kanner Austim to Asperger Syndromes, the Difficult Task to Predict Where ASD People Look at," *IEEE Access*, vol. 8, pp. 162132–162140, 2020.

[60] S. M. Myers, C. P. Johnson, and the Council on Children With Disabilities, "Management of Children With Autism Spectrum Disorders," *Pediatrics*, vol. 120, pp. 1162–1182, Nov. 2007.

[61] P. McCarty and R. E. Frye, "Early Detection and Diagnosis of Autism Spectrum Disorder: Why Is It So Difficult?," *Seminars in Pediatric Neurology*, vol. 35, p. 100831, Oct. 2020.

[62] M. Kohli, A. K. Kar, and S. Sinha, "The Role of Intelligent Technologies in Early Detection of Autism Spectrum Disorder (ASD): A Scoping Review," *IEEE Access*, vol. 10, pp. 104887–104913, 2022.

[63] Council on Children With Disabilities, Section on Developmental Behavioral Pediatrics, Bright Futures Steering Committee, and Medical Home Initiatives for Children With Special Needs Project Advisory Committee, "Identifying Infants and Young Children With Developmental Disorders in the Medical Home: An Algorithm for Developmental Surveillance and Screening," *Pediatrics*, vol. 118, pp. 405–420, July 2006.

[64] "(ADOS®-2) Autism Diagnostic Observation Schedule™, Second Edition." https://www.wpspublish.com/ados-2-autism-diagnostic-observation-schedule-second-edition.

[65] "(ADI®-R) Autism Diagnostic Interview–Revised." https://www.wpspublish.com/adi-r-autism-diagnostic-interviewrevised.html.

[66] C. Ecker, S. Y. Bookheimer, and D. G. M. Murphy, "Neuroimaging in autism spectrum disorder: Brain structure and function across the lifespan," *The Lancet Neurology*, vol. 14, pp. 1121–1134, Nov. 2015.

[67] L. Fusar-Poli, N. Brondino, P. Politi, and E. Aguglia, "Missed diagnoses and misdiagnoses of adults with autism spectrum disorder," *European Archives of Psychiatry and Clinical Neuroscience*, vol. 272, pp. 187–198, Mar. 2022.

[68] CDC, "Data and Statistics on Autism Spectrum Disorder | CDC." https://www.cdc.gov/ncbddd/autism/data.html, Jan. 2023.

[69] S. L. Ferri, T. Abel, and E. S. Brodkin, "Sex Differences in Autism Spectrum Disorder: A Review," *Current Psychiatry Reports*, vol. 20, p. 9, Mar. 2018.

[70] R. E. Frye, S. Vassall, G. Kaur, C. Lewis, M. Karim, and D. Rossignol, "Emerging biomarkers in autism spectrum disorder: A systematic review," *Annals of Translational Medicine*, vol. 7, p. 792, Dec. 2019.

[71] R. W. Emerson, C. Adams, T. Nishino, H. C. Hazlett, J. J. Wolff, L. Zwaigenbaum, J. N. Constantino, M. D. Shen, M. R. Swanson, J. T. Elison, S. Kandala, A. M. Estes, K. N. Botteron, L. Collins, S. R. Dager, A. C. Evans, G. Gerig, H. Gu, R. C. McKinstry, S. Paterson, R. T. Schultz, M. Styner, IBIS. Network, B. L. Schlaggar, J. R. Pruett, and J. Piven, "Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age," *Science translational medicine*, vol. 9, p. eaag2882, June 2017.

[72] M. Khodatars, A. Shoeibi, D. Sadeghi, N. Ghaasemi, M. Jafari, P. Moridian, A. Khadem, R. Alizadehsani, A. Zare, Y. Kong, A. Khosravi, S. Nahavandi, S. Hussain, U. R. Acharya, and M. Berk, "Deep learning for neuroimaging-based diagnosis and rehabilitation of Autism Spectrum Disorder: A review," *Computers in Biology and Medicine*, vol. 139, p. 104949, Dec. 2021.

[73] S. A. Mitelman, M.-C. Bralet, M. Mehmet Haznedar, E. Hollander, L. Shihabuddin, E. A. Hazlett, and M. S. Buchsbaum, "Positron emission tomography assessment of cerebral glucose metabolic rates in autism spectrum disorder and schizophrenia," *Brain Imaging and Behavior*, vol. 12, pp. 532–546, Apr. 2018.

[74] T. Mori, K. Mori, E. Fujii, Y. Toda, M. Miyazaki, M. Harada, T. Hashimoto, and S. Kagami, "Evaluation of the GABAergic nervous system in autistic brain: 123I-iomazenil SPECT study," *Brain and Development*, vol. 34, pp. 648–654, Sept. 2012.

[75] A. Del Casale, S. Ferracuti, A. Alcibiade, S. Simone, M. N. Modesti, and M. Pompili, "Neuroanatomical correlates of autism spectrum disorders: A meta-analysis of structural magnetic resonance imaging (MRI) studies," *Psychiatry Research: Neuroimaging*, vol. 325, p. 111516, Sept. 2022.

[76] A. D. Nijhof, L. Bardi, M. Brass, and J. R. Wiersema, "Brain activity for spontaneous and explicit mentalizing in adults with autism spectrum disorder: An fMRI study," *NeuroImage: Clinical*, vol. 18, pp. 475–484, Jan. 2018.

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

[78] T. P. DeRamus and R. K. Kana, "Anatomical likelihood estimation meta-analysis of grey and white matter anomalies in autism spectrum disorders," *NeuroImage: Clinical*, vol. 7, pp. 525–536, Jan. 2015.

[79] J. S. Kohli, M. K. Kinnear, C. H. Fong, I. Fishman, R. A. Carper, and R.-A. Müller, "Local Cortical Gyrification is Increased in Children With Autism Spectrum Disorders, but Decreases Rapidly in Adolescents," *Cerebral Cortex*, vol. 29, pp. 2412–2423, June 2019.

[80] C. M. Schumann, C. C. Barnes, C. Lord, and E. Courchesne, "Amygdala Enlargement in Toddlers with Autism Related to Severity of Social and Communication Impairments," *Biological Psychiatry*, vol. 66, pp. 942–949, Nov. 2009.

[81] N. E. V. Foster, K. A. R. Doyle-Thomas, A. Tryfon, T. Ouimet, E. Anagnostou, A. C. Evans, L. Zwaigenbaum, J. P. Lerch, J. D. Lewis, and K. L. Hyde, "Structural Gray Matter Differences During Childhood Development in Autism Spectrum Disorder: A Multimetric Approach," *Pediatric Neurology*, vol. 53, pp. 350–359, Oct. 2015.

[82] J. J. Wolff, G. Gerig, J. D. Lewis, T. Soda, M. A. Styner, C. Vachet, K. N. Botteron, J. T. Elison, S. R. Dager, A. M. Estes, H. C. Hazlett, R. T. Schultz, L. Zwaigenbaum, and J. Piven, "Altered corpus callosum morphology associated with autism over the first 2 years of life," *Brain*, vol. 138, pp. 2046–2058, July 2015.

[83] T. Qiu, C. Chang, Y. Li, L. Qian, C. Y. Xiao, T. Xiao, X. Xiao, Y. H. Xiao, K. K. Chu, M. H. Lewis, and X. Ke, "Two years changes in the development of caudate nucleus are involved in restricted repetitive behaviors in 2–5-year-old children with autism spectrum disorder," *Developmental Cognitive Neuroscience*, vol. 19, pp. 137–143, June 2016.

[84] M. D. Shen, C. W. Nordahl, D. D. Li, A. Lee, K. Angkustsiri, R. W. Emerson, S. J. Rogers, S. Ozonoff, and D. G. Amaral, "Extra-axial cerebrospinal fluid in high-risk and normal-risk children with autism aged 2–4 years: A case-control study," *The Lancet Psychiatry*, vol. 5, pp. 895–904, Nov. 2018.

[85] F. Rafiee, R. Rezvani Habibabadi, M. Motaghi, D. M. Yousem, and I. J. Yousem, "Brain MRI in Autism Spectrum Disorder: Narrative Review and Recent Advances," *Journal of Magnetic Resonance Imaging*, vol. 55, no. 6, pp. 1613–1624, 2022.

[86] X. Li, K. Zhang, X. He, J. Zhou, C. Jin, L. Shen, Y. Gao, M. Tian, and H. Zhang, "Structural, Functional, and Molecular Imaging of Autism Spectrum Disorder," *Neuroscience Bulletin*, vol. 37, pp. 1051–1071, July 2021.

[87] M. W. Mosconi and J. A. Sweeney, "Sensorimotor dysfunctions as primary features of autism spectrum disorders," *Science China Life Sciences*, vol. 58, pp. 1016–1023, Oct. 2015.

[88] B. R. Morgan, G. M. Ibrahim, V. M. Vogan, R. C. Leung, W. Lee, and M. J. Taylor, "Characterization of Autism Spectrum Disorder across the Age Span by Intrinsic Network Patterns," *Brain Topography*, vol. 32, pp. 461–471, May 2019.

[89] L. K. Fung, R. E. Flores, M. Gu, K. L. Sun, D. James, R. K. Schuck, B. Jo, J. H. Park, B. C. Lee, J. H. Jung, S. E. Kim, M. Saggar, M. D. Sacchet, G. Warnock, M. M. Khalighi, D. Spielman, F. T. Chin, and A. Y. Hardan, "Thalamic and prefrontal GABA concentrations but not GABAA receptor densities are altered in high-functioning adults with autism spectrum disorder," *Molecular Psychiatry*, vol. 26, pp. 1634–1646, May 2021.

[90] D. C. Chugani, "Neuroimaging and Neurochemistry of Autism," *Pediatric Clinics of North America*, vol. 59, pp. 63–73, Feb. 2012.

[91] R. B. Buxton, "The physics of functional magnetic resonance imaging (fMRI)," *Reports on Progress in Physics*, vol. 76, p. 096601, Sept. 2013.

[92] G. H. Glover, "Overview of Functional Magnetic Resonance Imaging," *Neurosurgery Clinics of North America*, vol. 22, pp. 133–139, Apr. 2011.

[93] J. M. Soares, R. Magalhães, P. S. Moreira, A. Sousa, E. Ganz, A. Sampaio, V. Alves, P. Marques, and N. Sousa, "A Hitchhiker's Guide to Functional Magnetic Resonance Imaging," *Frontiers in Neuroscience*, vol. 10, Nov. 2016.

[94] M. P. Van Den Heuvel and H. E. Hulshoff Pol, "Exploring the brain network: A review on resting-state fMRI functional connectivity," *European Neuropsychopharmacology*, vol. 20, pp. 519–534, Aug. 2010.

[95] F. Bloch, "Nuclear Induction," *Physical Review*, vol. 70, pp. 460–474, Oct. 1946.

[96] E. M. Purcell, H. C. Torrey, and R. V. Pound, "Resonance Absorption by Nuclear Magnetic Moments in a Solid," *Physical Review*, vol. 69, pp. 37–38, Jan. 1946.

[97] V. P. Grover, J. M. Tognarelli, M. M. Crossey, I. J. Cox, S. D. Taylor-Robinson, and M. J. McPhail, "Magnetic Resonance Imaging: Principles and Techniques: Lessons for Clinicians," *Journal of Clinical and Experimental Hepatology*, vol. 5, pp. 246–255, Sept. 2015.

[98] S. Currie, N. Hoggard, I. J. Craven, M. Hadjivassiliou, and I. D. Wilkinson, "Understanding MRI: Basic MR physics for physicians," *Postgraduate Medical Journal*, vol. 89, pp. 209–223, Apr. 2013.

[99] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan, *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley, 1 ed., Apr. 2014.

[100] J. P. Lerch, A. J. W. Van Der Kouwe, A. Raznahan, T. Paus, H. Johansen-Berg, K. L. Miller, S. M. Smith, B. Fischl, and S. N. Sotiropoulos, "Studying neuroanatomy using MRI," *Nature Neuroscience*, vol. 20, pp. 314–326, Mar. 2017.

[101] B. H. Brown, ed., *Medical Physics and Biomedical Engineering*. Medical Science Series, Bristol ; Philadelphia: Institute of Physics Pub, 1999.

[102] L. Raimondo, Ĺ. A. Oliveira, J. Heij, N. Priovoulos, P. Kundu, R. F. Leoni, and W. Van Der Zwaag, "Advances in resting state fMRI acquisitions for functional connectomics," *NeuroImage*, vol. 243, p. 118503, Nov. 2021.

[103] K. Smitha, K. Akhil Raja, K. Arun, P. Rajesh, B. Thomas, T. Kapilamoorthy, and C. Kesavadas, "Resting state fMRI: A review on methods in resting state connectivity analysis and resting state networks," *The Neuroradiology Journal*, vol. 30, pp. 305–317, Aug. 2017.

[104] J. C. Gore, "Principles and practice of functional MRI of the human brain," *Journal of Clinical Investigation*, vol. 112, pp. 4–9, July 2003.

[105] C. Gauthier and A. Fan, "BOLD signal physiology: Models and applications," *NeuroImage*, vol. 187, pp. 116–127, Feb. 2019.

[106] J. C. Siero, A. Bhogal, and J. M. Jansma, "Blood Oxygenation Level–dependent/Functional Magnetic Resonance Imaging," *PET Clinics*, vol. 8, pp. 329–344, July 2013.

[107] M. Barth and B. A. Poser, "Advances in High-Field BOLD fMRI," *Materials*, vol. 4, pp. 1941–1955, Nov. 2011.

[108] B. A. Seitzman, A. Z. Snyder, E. C. Leuthardt, and J. S. Shimony, "The State of Resting State Networks," *Topics in Magnetic Resonance Imaging*, vol. 28, pp. 189–196, Aug. 2019.

[109] B. T. Thomas Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, B. Fischl, H. Liu, and R. L. Buckner, "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *Journal of Neurophysiology*, vol. 106, pp. 1125–1165, Sept. 2011.

[110] T. F. Boerger, P. Pahapill, A. M. Butts, E. Arocho-Quinones, M. Raghavan, and M. O. Krucoff, "Large-scale brain networks and intra-axial tumor surgery: A narrative review of functional mapping techniques, critical needs, and scientific opportunities," *Frontiers in Human Neuroscience*, vol. 17, p. 1170419, July 2023.

[111] E. Canario, D. Chen, and B. Biswal, "A review of resting-state fMRI and its use to examine psychiatric disorders," *Psychoradiology*, vol. 1, pp. 42–53, May 2021.

[112] J. Caspers, C. Rubbert, S. B. Eickhoff, F. Hoffstaedter, M. Südmeyer, C. J. Hartmann, B. Sigl, N. Teichert, J. Aissa, B. Turowski, A. Schnitzler, and C. Mathys, "Within- and across-network alterations of the sensorimotor network in Parkinson's disease," *Neuroradiology*, vol. 63, pp. 2073–2085, Dec. 2021.

[113] M. Catani, F. Dell'Acqua, and M. Thiebaut De Schotten, "A revised limbic system model for memory, emotion and behaviour," *Neuroscience & Biobehavioral Reviews*, vol. 37, pp. 1724–1737, Sept. 2013.

[114] S. Vossel, J. J. Geng, and G. R. Fink, "Dorsal and Ventral Attention Systems: Distinct Neural Circuits but Collaborative Roles," *The Neuroscientist*, vol. 20, pp. 150–159, Apr. 2014.

[115] J. Schimmelpfennig, J. Topczewski, W. Zajkowski, and K. Jankowiak-Siuda, "The role of the salience network in cognitive and affective deficits," *Frontiers in Human Neuroscience*, vol. 17, p. 1133367, Mar. 2023.

[116] E. W. Lang, A. M. Tomé, I. R. Keck, J. M. Górriz-Sáez, and C. G. Puntonet, "Brain Connectivity Analysis: A Short Survey," *Computational Intelligence and Neuroscience*, vol. 2012, pp. 1–21, 2012.

[117] K. Li, L. Guo, J. Nie, G. Li, and T. Liu, "Review of methods for functional brain connectivity detection using fMRI," *Computerized Medical Imaging and Graphics*, vol. 33, pp. 131–139, Mar. 2009.

[118] G. Deco and M. L. Kringelbach, "Great Expectations: Using Whole-Brain Computational Connectomics for Understanding Neuropsychiatric Disorders," *Neuron*, vol. 84, pp. 892–905, Dec. 2014.

[119] A. Fornito, A. Zalesky, and M. Breakspear, "The connectomics of brain disorders," *Nature Reviews Neuroscience*, vol. 16, pp. 159–172, Mar. 2015.

[120] M. Venkatesh, J. Jaja, and L. Pessoa, "Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification," *NeuroImage*, vol. 207, p. 116398, Feb. 2020.

[121] J. Zhang, A. Kucyi, J. Raya, A. N. Nielsen, J. S. Nomi, J. S. Damoiseaux, D. J. Greene, S. G. Horovitz, L. Q. Uddin, and S. Whitfield-Gabrieli, "What have we really learned from functional connectivity in clinical populations?," *NeuroImage*, vol. 242, p. 118466, Nov. 2021.

[122] J. E. Chen and G. H. Glover, "Functional Magnetic Resonance Imaging Methods," *Neuropsychology Review*, vol. 25, pp. 289–313, Sept. 2015.

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

[124] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain," *NeuroImage*, vol. 15, pp. 273–289, Jan. 2002.

[125] A. K. Azeez and B. B. Biswal, "A Review of Resting-State Analysis Methods," *Neuroimaging Clinics of North America*, vol. 27, pp. 581–592, Nov. 2017.

[126] S. Eickhoff and V. Müller, "Functional Connectivity," in *Brain Mapping*, pp. 187–201, Elsevier, 2015.

[127] M. Lee, C. Smyser, and J. Shimony, "Resting-State fMRI: A Review of Methods and Clinical Applications," *American Journal of Neuroradiology*, vol. 34, pp. 1866–1872, Oct. 2013.

[128] S. E. Joel, B. S. Caffo, P. C. M. Van Zijl, and J. J. Pekar, "On the relationship between seed-based and ICA-based measures of functional connectivity," *Magnetic Resonance in Medicine*, vol. 66, pp. 644–657, Sept. 2011.

[129] A. Fornito, A. Zalesky, and M. Breakspear, "Graph analysis of the human connectome: Promise, progress, and pitfalls," *NeuroImage*, vol. 80, pp. 426–444, Oct. 2013.

[130] T. M. Mitchell, J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, *Machine Learning: A Guide to Current Research*, vol. 12 of *The Kluwer International Series in Engineering and Computer Science*. Boston, MA: Springer US, 1986.

[131] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[132] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Massachusetts, USA: MIT Press, 2017.

[133] L. Waikhom and R. Patgiri, "A survey of graph neural networks in various learning paradigms: Methods, applications, and challenges," *Artificial Intelligence Review*, vol. 56, pp. 6295–6364, July 2023.

[134] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric Deep Learning: Going beyond Euclidean data," *IEEE Signal Processing Magazine*, vol. 34, pp. 18–42, July 2017.

[135] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, "Graph-Based Deep Learning for Medical Diagnosis and Analysis: Past, Present and Future," *Sensors*, vol. 21, p. 4758, July 2021.

[136] S. Xiao, S. Wang, Y. Dai, and W. Guo, "Graph neural networks in node classification: Survey and evaluation," *Machine Vision and Applications*, vol. 33, p. 4, Nov. 2021.

[137] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 4–24, Jan. 2021.

[138] S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," *Computer Science Review*, vol. 40, p. 100379, May 2021.

[139] A. Shrestha and A. Mahmood, "Review of Deep Learning Algorithms and Architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.

[140] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, and M. Dehmer, "An Introductory Review of Deep Learning for Prediction Models With Big Data," *Frontiers in Artificial Intelligence*, vol. 3, 2020.

[141] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference for Learning Representations*, (San Diego), arXiv preprint arXiv:1412.6980, 2015.

[142] W. L. Hamilton, "Graph Representation Learning," vol. 14, no. 3, pp. 1–159, 2020.

[143] P. Veličković, "Everything is connected: Graph neural networks," *Current Opinion in Structural Biology*, vol. 79, p. 102538, Apr. 2023.

[144] Y. Chen and L. Wu, "Graph Neural Networks: Graph Structure Learning," in *Graph Neural Networks: Foundations, Frontiers, and Applications* (L. Wu, P. Cui, J. Pei, and L. Zhao, eds.), pp. 297–321, Singapore: Springer Nature Singapore, 2022.

[145] P. Cui, L. Wu, J. Pei, L. Zhao, and X. Wang, "Graph Representation Learning," in *Graph Neural Networks: Foundations, Frontiers, and Applications*, pp. 17–26, Singapore: Springer Nature Singapore, 2022.

[146] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," in *Neural Information Processing Systems*, arXiv, Sept. 2018.

[147] R. Tripodi and M. Pelillo, "6 - Transductive Learning Games for Word Sense Disambiguation," in *Cognitive Approach to Natural Language Processing* (B. Sharp, F. Sèdes, and W. Lubaszewski, eds.), pp. 109–128, Elsevier, Jan. 2017.

[148] W. Lingfei, P. Cui, J. Pei, L. Zhao, and L. Song, "Graph Neural Networks," in *Graph Neural Networks: Foundations, Frontiers, and Applications*, pp. 27–37, Singapore: Springer Singapore, 2022.

[149] P. Li and J. Leskovec, "The Expressive Power of Graph Neural Networks," in *Graph Neural Networks: Foundations, Frontiers, and Applications* (L. Wu, P. Cui, J. Pei, and L. Zhao, eds.), pp. 63–98, Singapore: Springer Nature Singapore, 2022.

[150] S. Brody, U. Alon, and E. Yahav, "How Attentive are Graph Attention Networks?," Jan. 2022.

[151] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *International Conference on Learning Representations*, Feb. 2018.

[152] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *5th International Conference on Learning Representations*, arXiv, Feb. 2017.

[153] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," Feb. 2017.

[154] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral Networks and Locally Connected Networks on Graphs," *arXiv preprint arXiv:1312.6203*, p. 14, May 2014.

[155] J. Tang and R. Liao, "Graph Neural Networks for Node Classification," in *Graph Neural Networks: Foundations, Frontiers, and Applications* (L. Wu, P. Cui, J. Pei, and L. Zhao, eds.), pp. 41–61, Singapore: Springer Nature Singapore, 2022.

[156] Q. Li, Z. Han, and X.-M. Wu, "Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning," in *AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, p. 8, AAAI Press, Jan. 2018.

[157] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, 2016.

[158] T. Chen, K. Zhou, K. Duan, W. Zheng, P. Wang, X. Hu, and Z. Wang, "Bag of Tricks for Training Deeper Graph Neural Networks: A Comprehensive Benchmark Study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 2769–2781, Mar. 2023.

[159] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, pp. 157–166, Mar. 1994.

[160] G. Li, M. Müller, G. Qian, I. C. Delgadillo, A. Abualshour, A. Thabet, and B. Ghanem, "DeepGCNs: Making GCNs Go as Deep as CNNs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 6923–6939, June 2023.

[161] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, IEEE, 2017.

[162] K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka, "Representation Learning on Graphs with Jumping Knowledge Networks," in *35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018*, June 2018.

[163] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[164] H. Ma, Y. Rong, and J. Huang, "Graph Neural Networks: Scalability," in *Graph Neural Networks: Foundations, Frontiers, and Applications* (L. Wu, P. Cui, J. Pei, and L. Zhao, eds.), pp. 99–119, Singapore: Springer Nature Singapore, 2022.

[165] W. Huang, T. Zhang, Y. Rong, and J. Huang, "Adaptive Sampling Towards Fast Graph Representation Learning," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, (Montreal, Canada), pp. 4563–4572, Curran Associates Inc., Nov. 2018.

[166] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, "Cluster-GCN: An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 257–266, July 2019.

[167] H. Zeng, H. Zhou, A. Srivastava, R. Kannan, and V. Prasanna, "GraphSAINT: Graph Sampling Based Inductive Learning Method," in *arXiv.Org*, July 2019.

[168] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in Graph Neural Networks: A Taxonomic Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–19, 2022.

[169] T. Hulsen, "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare," *AI*, vol. 4, pp. 652–666, Aug. 2023.

[170] K. Amara, R. Ying, Z. Zhang, Z. Han, Y. Shan, U. Brandes, S. Schemm, and C. Zhang, "GraphFramEx: Towards Systematic Evaluation of Explainability Methods for Graph Neural Networks," in *NeurIPS 2022 GLFrontiers Workshop*, arXiv, Oct. 2022.

[171] N. Liu, Q. Feng, and X. Hu, "Interpretability in Graph Neural Networks," in *Graph Neural Networks: Foundations, Frontiers, and Applications* (L. Wu, P. Cui, J. Pei, and L. Zhao, eds.), pp. 121–147, Singapore: Springer Nature Singapore, 2022.

[172] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating Explanations for Graph Neural Networks," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019.

[173] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized Explainer for Graph Neural Network," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, (Vancouver, BC, Canada), Curran Associates Inc., Nov. 2020.

[174] L. Zhang, M. Wang, M. Liu, and D. Zhang, "A Survey on Deep Learning for Neuroimaging-Based Brain Disorder Analysis," *Frontiers in Neuroscience*, vol. 14, p. 779, Oct. 2020.

[175] D.-Y. Song, S. Y. Kim, G. Bong, J. M. Kim, and H. J. Yoo, "The Use of Artificial Intelligence in Screening and Diagnosis of Autism Spectrum Disorder: A Literature Review," *Journal of the Korean Academy of Child and Adolescent Psychiatry*, vol. 30, pp. 145–152, Oct. 2019.

[176] D.-Y. Song, C.-C. Topriceanu, D. C. Ilie-Ablachim, M. Kinali, and S. Bisdas, "Machine learning with neuroimaging data to identify autism spectrum disorder: A systematic review and meta-analysis," *Neuroradiology*, vol. 63, pp. 2057–2072, Dec. 2021.

[177] M. Xu, V. Calhoun, R. Jiang, W. Yan, and J. Sui, "Brain imaging-based machine learning in autism spectrum disorder: Methods and applications," *Journal of Neuroscience Methods*, vol. 361, p. 109271, Sept. 2021.

[178] A. Abraham, M. P. Milham, A. Di Martino, R. C. Craddock, D. Samaras, B. Thirion, and G. Varoquaux, "Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example," *NeuroImage*, vol. 147, pp. 736–745, Feb. 2017.

[179] A. Brahim and N. Farrugia, "Graph Fourier transform of fMRI temporal signals based on an averaged structural connectome for the classification of neuroimaging," *Artificial Intelligence in Medicine*, vol. 106, p. 101870, June 2020.

[180] O. Dekhil, H. Hajjdiab, B. Ayinde, A. Shalaby, A. Switala, D. Sosnin, A. Elshamekh, M. Ghazal, R. Keynton, G. Barnes, and A. El-Baz, "Using resting state functional MRI to build a personalized autism diagnosis system," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 1381–1385, Apr. 2018.

[181] A. Irimia, X. Lei, C. M. Torgerson, Z. J. Jacokes, S. Abe, and J. D. Van Horn, "Support Vector Machines, Multidimensional Scaling and Magnetic Resonance Imaging Reveal Structural Brain Abnormalities Associated With the Interaction Between Autism Spectrum Disorder and Sex," *Frontiers in Computational Neuroscience*, vol. 12, p. 93, Nov. 2018.

[182] T.-E. Kam, H.-I. Suk, and S.-W. Lee, "Multiple functional networks modeling for autism spectrum disorder diagnosis," *Human Brain Mapping*, vol. 38, no. 11, pp. 5804–5821, 2017.

[183] G. Spera, A. Retico, P. Bosco, E. Ferrari, L. Palumbo, P. Oliva, F. Muratori, and S. Calderoni, "Evaluation of Altered Functional Connections in Male Children With Autism Spectrum Disorders on Multiple-Site Data Optimized With Machine Learning," *Frontiers in Psychiatry*, vol. 10, 2019.

[184] C. P. Chen, C. L. Keown, A. Jahedi, A. Nair, M. E. Pflieger, B. A. Bailey, and R.-A. Müller, "Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default mode, and visual regions in autism," *NeuroImage: Clinical*, vol. 8, pp. 238–245, Jan. 2015.

[185] A. Kazeminejad and R. C. Sotero, "Topological Properties of Resting-State fMRI Functional Networks Improve Machine Learning-Based Autism Classification," *Frontiers in Neuroscience*, vol. 12, 2019.

[186] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. Minshew, C. S. Monk, S. Mueller, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, and M. P. Milham, "The Autism Brain Imaging Data Exchange: Towards Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism," *Molecular psychiatry*, vol. 19, pp. 659–667, June 2014.

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

autism using the autism brain imaging data exchange II," *Scientific Data*, vol. 4, p. 170010, Mar. 2017.

[188] Z. Sherkatghanad, M. Akhondzadeh, S. Salari, M. Zomorodi-Moghadam, M. Abdar, U. R. Acharya, R. Khosrowabadi, and V. Salari, "Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network," *Frontiers in Neuroscience*, vol. 13, 2020.

[189] X. Yang, N. Zhang, and P. Schrader, "A study of brain networks for autism spectrum disorder classification using resting-state functional connectivity," *Machine Learning with Applications*, vol. 8, p. 100290, June 2022.

[190] P. Elliott, T. C. Peakman, and on behalf of UK Biobank, "The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine," *International Journal of Epidemiology*, vol. 37, pp. 234–244, Apr. 2008.

[191] F. Z. Subah, K. Deb, P. K. Dhar, and T. Koshiba, "A Deep Learning Approach to Predict Autism Spectrum Disorder Using Multisite Resting-State fMRI," *Applied Sciences*, vol. 11, p. 3636, Jan. 2021.

[192] F. Almuqhim and F. Saeed, "ASD-SAENet: A Sparse Autoencoder, and Deep-Neural Network Model for Detecting Autism Spectrum Disorder (ASD) Using fMRI Data," *Frontiers in Computational Neuroscience*, vol. 15, 2021.

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

[194] D. P. Kennedy and E. Courchesne, "The intrinsic functional organization of the brain is altered in autism," *NeuroImage*, vol. 39, pp. 1877–1885, Feb. 2008.

[195] S. Baron-Cohen, H. Ring, J. Moriarty, B. Schmitz, D. Costa, and P. Ell, "Recognition of Mental State Terms: Clinical Findings in Children with Autism and a Functional Neuroimaging Study of Normal Adults," *British Journal of Psychiatry*, vol. 165, pp. 640–649, Nov. 1994.

[196] V. L. Cherkassky, R. K. Kana, T. A. Keller, and M. A. Just, "Functional connectivity in a baseline resting-state network in autism," *NeuroReport*, vol. 17, pp. 1687–1690, Nov. 2006.

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

[198] K. A. R. Doyle-Thomas, W. Lee, N. E. V. Foster, A. Tryfon, T. Ouimet, K. L. Hyde, A. C. Evans, J. Lewis, L. Zwaigenbaum, E. Anagnostou, and f. t. N. A. I. Group, "Atypical functional brain connectivity during rest in autism spectrum disorders," *Annals of Neurology*, vol. 77, no. 5, pp. 866–876, 2015.

[199] Q. Wang, H.-Y. Li, Y.-D. Li, Y.-T. Lv, H.-B. Ma, A.-F. Xiang, X.-Z. Jia, and D.-Q. Liu, "Resting-state abnormalities in functional connectivity of the default mode network in autism spectrum disorder: A meta-analysis," *Brain Imaging and Behavior*, vol. 15, pp. 2583–2592, Oct. 2021.

[200] J. M. Lee, S. Kyeong, E. Kim, and K.-A. Cheon, "Abnormalities of Inter- and Intra-Hemispheric Functional Connectivity in Autism Spectrum Disorders: A Study Using the Autism Brain Imaging Data Exchange Database," *Frontiers in Neuroscience*, vol. 10, p. 191, May 2016.

[201] X. Liu and H. Huang, "Alterations of functional connectivities associated with autism spectrum disorder symptom severity: A multi-site study using multivariate pattern analysis," *Scientific Reports*, vol. 10, p. 4330, Mar. 2020.

[202] P. Hannant, S. Cassidy, T. Tavassoli, and F. Mann, "Sensorimotor Difficulties Are Associated with the Severity of Autism Spectrum Conditions," *Frontiers in Integrative Neuroscience*, vol. 10, 2016.

[203] B. E. Yerys, E. M. Gordon, D. N. Abrams, T. D. Satterthwaite, R. Weinblatt, K. F. Jankowski, J. Strang, L. Kenworthy, W. D. Gaillard, and C. J. Vaidya, "Default mode network segregation and social deficits in autism spectrum disorder: Evidence from non-medicated children," *NeuroImage: Clinical*, vol. 9, pp. 223–232, Jan. 2015.

[204] A. Caria and S. de Falco, "Anterior insular cortex regulation in autism spectrum disorders," *Frontiers in Behavioral Neuroscience*, vol. 9, p. 38, Mar. 2015.

[205] E. A. H. von dem Hagen, R. S. Stoyanova, S. Baron-Cohen, and A. J. Calder, "Reduced functional connectivity within and between 'social' resting state networks in autism spectrum conditions," *Social Cognitive and Affective Neuroscience*, vol. 8, pp. 694–701, Aug. 2013.

[206] Y.-Y. Chen, M. Uljarevic, J. Neal, S. Greening, H. Yim, and T.-H. Lee, "Excessive Functional Coupling With Less Variability Between Salience and Default Mode Networks in Autism Spectrum Disorder," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 7, pp. 876–884, Sept. 2022.

[207] J. V. Hull, L. B. Dokovna, Z. J. Jacokes, C. M. Torgerson, A. Irimia, and J. D. Van Horn, "Resting-State Functional Connectivity in Autism Spectrum Disorders: A Review," *Frontiers in Psychiatry*, vol. 7, 2017.

[208] K. Dadi, M. Rahim, A. Abraham, D. Chyzhyk, M. Milham, B. Thirion, and G. Varoquaux, "Benchmarking functional connectome-based predictive models for resting-state fMRI," *NeuroImage*, vol. 192, pp. 115–134, May 2019.

[209] T. Eslami, V. Mirjalili, A. Fong, A. R. Laird, and F. Saeed, "ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data," *Frontiers in Neuroinformatics*, vol. 13, 2019.

[210] M. Rakić, M. Cabezas, K. Kushibar, A. Oliver, and X. Lladó, "Improving the detection of autism spectrum disorder by combining structural and functional MRI information," *NeuroImage: Clinical*, vol. 25, p. 102181, Jan. 2020.

[211] N. C. Dvornek, P. Ventola, and J. S. Duncan, "Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 725–728, Apr. 2018.

[212] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 20, pp. 61–80, Jan. 2009.

[213] A. Kazi, S. Shekarforoush, S. Arvind Krishna, H. Burwinkel, G. Vivar, K. Kortüm, S.-A. Ahmadi, S. Albarqouni, and N. Navab, "InceptionGCN: Receptive Field Aware Graph Convolutional Network for Disease Prediction," in *Information Processing in Medical Imaging* (A. C. S. Chung, J. C. Gee, P. A. Yushkevich, and S. Bao, eds.), vol. 11492, pp. 73–85, Cham: Springer International Publishing, 2019.

[214] R. Anirudh and J. J. Thiagarajan, "Bootstrapping Graph Convolutional Neural Networks for Autism Spectrum Disorder Classification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3197–3201, May 2019.

[215] Z. Rakhimberdina, X. Liu, and T. Murata, "Population Graph-Based Multi-Model Ensemble Method for Diagnosing Autism Spectrum Disorder," *Sensors*, vol. 20, p. 6001, Jan. 2020.

[216] H. Jiang, P. Cao, M. Xu, J. Yang, and O. Zaiane, "Hi-GCN: A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction," *Computers in Biology and Medicine*, vol. 127, p. 104096, Dec. 2020.

[217] C. Yang, P. Wang, J. Tan, Q. Liu, and X. Li, "Autism spectrum disorder diagnosis using graph attention network based on spatial-constrained sparse functional brain networks," *Computers in Biology and Medicine*, vol. 139, p. 104963, Dec. 2021.

[218] Y. Huang and A. C. S. Chung, "Disease prediction with edge-variational graph convolutional networks," *Medical Image Analysis*, vol. 77, p. 102375, Apr. 2022.

[219] L. Li, H. Jiang, G. Wen, P. Cao, M. Xu, X. Liu, J. Yang, and O. Zaiane, "TE-HI-GCN: An Ensemble of Transfer Hierarchical Graph Convolutional Networks for Disorder Diagnosis," *Neuroinformatics*, vol. 20, pp. 353–375, Apr. 2022.

[220] J. Pan, H. Lin, Y. Dong, Y. Wang, and Y. Ji, "MAMF-GCN: Multi-scale adaptive multi-channel fusion deep graph convolutional network for predicting mental disorder," *Computers in Biology and Medicine*, vol. 148, p. 105823, Sept. 2022.

[221] H. Zhang, R. Song, L. Wang, L. Zhang, D. Wang, C. Wang, and W. Zhang, "Classification of Brain Disorders in rs-fMRI via Local-to-Global Graph Neural Networks," *IEEE Transactions on Medical Imaging*, vol. 42, pp. 444–455, Feb. 2023.

[222] J. Mao, J. Liu, H. Lin, H. Kuang, and Y. Pan, "Multi-modal Multi-kernel Graph Learning for Autism Prediction and Biomarker Discovery," *IEEE Transactions on Medical Imaging*, vol. 41, pp. 2207–2216, Mar. 2023.

[223] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. G. Moreno, B. Glocker, and D. Rueckert, "Spectral Graph Convolutions for Population-Based Disease Prediction," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017* (M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, eds.), vol. 10435, pp. 177–185, Cham: Springer International Publishing, 2017.

[224] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, (Red Hook, NY, USA), pp. 3844–3852, Curran Associates Inc., Dec. 2016.

[225] Y. Rong, W. Huang, T. Xu, and J. Huang, "DropEdge: Towards Deep Graph Convolutional Networks on Node Classification," *arXiv preprint arXiv:1907.10903*, Mar. 2020.

[226] S. Luan, M. Zhao, X.-W. Chang, and D. Precup, "Break the Ceiling: Stronger Multi-scale Deep Graph Convolutional Networks," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, NeurIPS Proceedings, Sept. 2019.

[227] D. Li, H.-O. Karnath, and X. Xu, "Candidate Biomarkers in Children with Autism Spectrum Disorder: A Review of MRI Studies," *Neuroscience Bulletin*, vol. 33, pp. 219–237, Apr. 2017.

[228] C. Press, N. Weiskopf, and J. M. Kilner, "Dissociable roles of human inferior frontal gyrus during action execution and observation," *NeuroImage*, vol. 60, pp. 1671–1677, Apr. 2012.

[229] "Preprocessed Connectomes Project." http://preprocessed-connectomes-project.org/.

[230] X. Wen, H. He, L. Dong, J. Chen, J. Yang, H. Guo, C. Luo, and D. Yao, "Alterations of local functional connectivity in lifespan: A resting-state fMRI study," *Brain and Behavior*, vol. 10, no. 7, p. e01652, 2020.

[231] M. Edde, G. Leroux, E. Altena, and S. Chanraud, "Functional brain connectivity changes across the human life span: From fetal development to old age," *Journal of Neuroscience Research*, vol. 99, no. 1, pp. 236–262, 2021.

[232] D. Tomasi and N. D. Volkow, "Gender differences in brain functional connectivity density," *Human Brain Mapping*, vol. 33, pp. 849–860, Mar. 2011.

[233] R. Casanova, C. Whitlow, B. Wagner, M. Espeland, and J. Maldjian, "Combining Graph and Machine Learning Methods to Analyze Differences in Functional Connectivity Across Sex," *The Open Neuroimaging Journal*, vol. 6, pp. 1–9, Jan. 2012.

[234] S. J. Fenske, J. Liu, H. Chen, M. A. Diniz, R. L. Stephens, E. Cornea, J. H. Gilmore, and W. Gao, "Sex differences in resting state functional connectivity across the first two years of life," *Developmental Cognitive Neuroscience*, vol. 60, p. 101235, Apr. 2023.

[235] K. Alaerts, S. P. Swinnen, and N. Wenderoth, "Sex differences in autism: A resting-state fMRI investigation of functional brain connectivity in males and females," *Social Cognitive and Affective Neuroscience*, vol. 11, pp. 1002–1016, June 2016.

[236] R. E. W. Smith, J. A. Avery, G. L. Wallace, L. Kenworthy, S. J. Gotts, and A. Martin, "Sex Differences in Resting-State Functional Connectivity of the Cerebellum in Autism Spectrum Disorder," *Frontiers in Human Neuroscience*, vol. 13, 2019.

[237] K. E. Lawrence, L. M. Hernandez, H. C. Bowman, N. T. Padgaonkar, E. Fuster, A. Jack, E. Aylward, N. Gaab, J. D. Van Horn, R. A. Bernier, D. H. Geschwind, J. C. McPartland, C. A. Nelson, S. J. Webb, K. A. Pelphrey, S. A. Green, S. Y. Bookheimer, and M. Dapretto, "Sex Differences in Functional Connectivity of the Salience, Default Mode, and Central Executive Networks in Youth with ASD," *Cerebral Cortex (New York, NY)*, vol. 30, pp. 5107–5120, July 2020.

[238] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance

Deep Learning Library," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[239] M. Fey and J. E. Lenssen, "Fast Graph Representation Learning with PyTorch Geometric," *ArXiv*, vol. abs/1903.02428, Apr. 2019.

[240] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, (Santiago, Chile), pp. 1026–1034, IEEE, Dec. 2015.

[241] Y. Huang, "Edge Variational Graph Convolutional Networks for Disease Prediction," Jan. 2024.

[242] cnuzh, "LG-GNN," Mar. 2024.

[243] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges," *ArXiv*, vol. abs/2104.13478, 2021.

[244] A. C. Linke, L. Olson, Y. Gao, I. Fishman, and R.-A. Müller, "Psychotropic Medication Use in Autism Spectrum Disorders May Affect Functional Brain Connectivity," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 2, pp. 518–527, Sept. 2017.

[245] Y. B. Yoon, J.-Y. Yun, W. H. Jung, K. I. K. Cho, S. N. Kim, T. Y. Lee, H. Y. Park, and J. S. Kwon, "Altered Fronto-Temporal Functional Connectivity in Individuals at Ultra-High-Risk of Developing Psychosis," *PLOS ONE*, vol. 10, p. e0135347, Aug. 2015.

[246] E. D. Bigler, S. Mortensen, E. S. Neeley, S. Ozonoff, L. Krasny, M. Johnson, J. Lu, S. L. Provencal, W. McMahon, and J. E. Lainhart, "Superior Temporal Gyrus, Language Function, and Autism," *Developmental Neuropsychology*, vol. 31, pp. 217–238, Mar. 2007.

[247] N. Boddaert, N. Chabane, P. Belin, M. Bourgeois, V. Royer, C. Barthelemy, M.-C. Mouren-Simeoni, A. Philippe, F. Brunelle, Y. Samson, and M. Zilbovicius, "Perception of Complex Sounds in Autism: Abnormal Auditory Cortical Processing in Children," *American Journal of Psychiatry*, vol. 161, pp. 2117–2120, Nov. 2004.

[248] K. S. Weiner and K. Zilles, "The anatomical and functional specialization of the fusiform gyrus," *Neuropsychologia*, vol. 83, pp. 48–62, Mar. 2016.

[249] J. A. Pereira, P. Sepulveda, M. Rana, C. Montalba, C. Tejos, R. Torres, R. Sitaram, and S. Ruiz, "Self-Regulation of the Fusiform Face Area in Autism Spectrum: A Feasibility Study With Real-Time fMRI Neurofeedback," *Frontiers in Human Neuroscience*, vol. 13, p. 446, Dec. 2019.

[250] P. H. J. M. Vlamings, L. M. Jonkman, E. Van Daalen, R. J. Van Der Gaag, and C. Kemner, "Basic Abnormalities in Visual Processing Affect Face Processing at an Early Age in Autism Spectrum Disorder," *Biological Psychiatry*, vol. 68, pp. 1107–1113, Dec. 2010.

[251] A. D. Smith, "Spatial navigation in autism spectrum disorders: A critical review," *Frontiers in Psychology*, vol. 6, Jan. 2015.

[252] A. J. Herringshaw, C. J. Ammons, T. P. DeRamus, and R. K. Kana, "Hemispheric differences in language processing in autism spectrum disorders: A meta-analysis of neuroimaging studies," *Autism Research*, vol. 9, pp. 1046–1057, Oct. 2016.

[253] S. M. Kaku, A. Jayashankar, S. C. Girimaji, S. Bansal, S. Gohel, R. D. Bharath, and S. Srinath, "Early childhood network alterations in severe autism," *Asian Journal of Psychiatry*, vol. 39, pp. 114–119, Jan. 2019.

[254] N. B. Dadario and M. E. Sughrue, "The functional role of the precuneus," *Brain*, vol. 146, pp. 3598–3607, Sept. 2023.

[255] P. Wantzen, A. Boursette, E. Zante, J. Mioche, F. Eustache, F. Guénolé, J.-M. Baleyte, and B. Guillery-Girard, "Autobiographical Memory and Social Identity in Autism: Preliminary Results of Social Positioning and Cognitive Intervention," *Frontiers in Psychology*, vol. 12, p. 641765, Mar. 2021.

[256] J. E. Norris and K. Maras, "Supporting autistic adults' episodic memory recall in interviews: The role of executive functions, theory of mind, and language abilities," *Autism*, vol. 26, pp. 513–524, Feb. 2022.

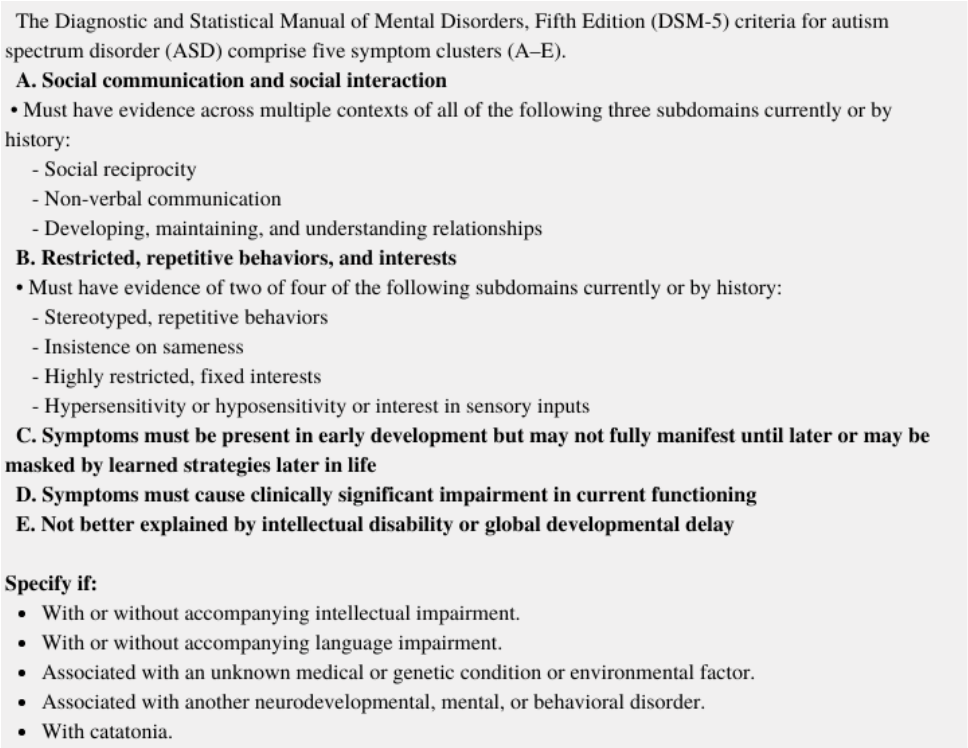
[257] M. V. Lombardo, J. L. Barnes, S. J. Wheelwright, and S. Baron-Cohen, "Self-Referential Cognition and Empathy in Autism," *PLoS ONE*, vol. 2, p. e883, Sept. 2007.

[258] X. Shan, L. Q. Uddin, R. Ma, P. Xu, J. Xiao, L. Li, X. Huang, Y. Feng, C. He, H. Chen, and X. Duan, "Disentangling the Individual-Shared and Individual-Specific Subspace of Altered Brain Functional Connectivity in Autism Spectrum Disorder," *Biological Psychiatry*, vol. 95, pp. 870–880, May 2024.

Appendix A

Appendix

A.1 Additional Information



The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) criteria for autism spectrum disorder (ASD) comprise five symptom clusters (A–E).

A. Social communication and social interaction

• Must have evidence across multiple contexts of all of the following three subdomains currently or by history:

- Social reciprocity
- Non-verbal communication
- Developing, maintaining, and understanding relationships

B. Restricted, repetitive behaviors, and interests

• Must have evidence of two of four of the following subdomains currently or by history:

- Stereotyped, repetitive behaviors
- Insistence on sameness
- Highly restricted, fixed interests
- Hypersensitivity or hyposensitivity or interest in sensory inputs

C. Symptoms must be present in early development but may not fully manifest until later or may be masked by learned strategies later in life

D. Symptoms must cause clinically significant impairment in current functioning

E. Not better explained by intellectual disability or global developmental delay

Specify if:

- With or without accompanying intellectual impairment.
- With or without accompanying language impairment.
- Associated with an unknown medical or genetic condition or environmental factor.
- Associated with another neurodevelopmental, mental, or behavioral disorder.
- With catatonia.

Figure A.1: Diagnostic criteria for ASD as presented in DSM-5 (adapted from [6]).

Table A.1: This table lists the corresponding brain ROIs associated with the FC connections identified by the feature labels in Figure 5.5, which represent the most important FC connections for ASD classification based on the developed GNN model. The feature labels are in the range between 0 and 1999.

Feature Label	Brain ROI 1	Brain ROI 1
491	Right Frontal Pole	Right Inferior Frontal Gyrus; pars triangularis
570	Right Middle Frontal Gyrus	Right Postcentral Gyrus
587	Right Middle Frontal Gyrus	Left Superior Temporal Gyrus; anterior division
611	Right Inferior Frontal Gyrus; pars triangularis	Right Cingulate Gyrus; posterior division
873	Right Middle Temporal Gyrus; temporooccipital part	Left Lingual Gyrus
897	Right Inferior Temporal Gyrus; anterior division	Left Lingual Gyrus
946	Right Postcentral Gyrus	Right Supramarginal Gyrus; anterior division
971	Right Superior Parietal Lobule	Right Supramarginal Gyrus; posterior division
973	Right Superior Parietal Lobule	Right Lateral Occipital Cortex; superior division
975	Right Superior Parietal Lobule	Right Frontal Medial Cortex
1522	Right Parietal Operculum Cortex	Left Middle Temporal Gyrus; temporooccipital part
1523	Right Parietal Operculum Cortex	Left Inferior Temporal Gyrus; posterior division
1537	Right Planum Polare	Left Middle Temporal Gyrus; temporooccipital part
1539	Right Planum Polare	Left Inferior Temporal Gyrus; posterior division
1584	Right Supracalcarine Cortex	Left Inferior Temporal Gyrus; posterior division
1586	Right Supracalcarine Cortex	Left Lateral Occipital Cortex; inferior division
1595	Right Supracalcarine Cortex	Left Heschl's Gyrus (includes H1 and H2)
1596	Right Supracalcarine Cortex	Left Supracalcarine Cortex
1597	Right Supracalcarine Cortex	Left Occipital Pole
1603	Right Occipital Pole	Left Intracalcarine Cortex
1780	Left Middle Temporal Gyrus; temporooccipital part	Left Inferior Temporal Gyrus; posterior division
1813	Left Inferior Temporal Gyrus; posterior division	Left Heschl's Gyrus (includes H1 and H2)
1815	Left Inferior Temporal Gyrus; posterior division	Left Occipital Pole
1823	Left Inferior Temporal Gyrus; temporooccipital part	Left Heschl's Gyrus (includes H1 and H2)
1942	Left Precuneous Cortex	Left Occipital Pole
1965	Left Lingual Gyrus	Left Planum Polare
1979	Left Temporal Occipital Fusiform Cortex	Left Planum Polare
1996	Left Heschl's Gyrus (includes H1 and H2)	Left Supracalcarine Cortex
1997	Left Heschl's Gyrus (includes H1 and H2)	Left Occipital Pole
1999	Left Supracalcarine Cortex	Left Occipital Pole

Table A.2: List of the acronyms used in Figure 5.7. The brain ROIs correspond to the Harvard-Oxford anatomical atlas.

Brain ROIs	Label
Left Thalamus	lTh
Left Caudate	lCa
Left Putamen	lPu
Left Pallidum	lPa
Left Hippocampus	lHip
Left Amygdala	lAm
Left Accumbens	lAc
Right Thalamus	rTh
Right Caudate	rCa
Right Putamen	rPu
Right Pallidum	rPa
Right Hippocampus	rHip
Right Amygdala	rAm
Right Accumbens	rAc
Right Frontal Pole	rFP
Right Insular Cortex	rIC
Right Superior Frontal Gyrus	rSFG
Right Middle Frontal Gyrus	rMFG
Right Inferior Frontal Gyrus; pars triangularis	rIFGtriang
Right Inferior Frontal Gyrus; pars opercularis	rIFGoperc
Right Precentral Gyrus	rPreG
Right Temporal Pole	rTP
Right Superior Temporal Gyrus; anterior division	raSTG
Right Superior Temporal Gyrus; posterior division	rpSTG
Right Middle Temporal Gyrus; anterior division	raMTG
Right Middle Temporal Gyrus; posterior division	rpMTG
Right Middle Temporal Gyrus; temporooccipital part	rtoMTG
Right Inferior Temporal Gyrus; anterior division	raITG
Right Inferior Temporal Gyrus; posterior division	rpITG
Right Inferior Temporal Gyrus; temporooccipital part	rtoITG
Right Postcentral Gyrus	rPostG
Right Superior Parietal Lobule	rSPL
Right Supramarginal Gyrus; anterior division	raSG
Right Supramarginal Gyrus; posterior division	rpSG
Right Angular Gyrus	rAG
Right Lateral Occipital Cortex; superior division	rsLOC
Right Lateral Occipital Cortex; inferior division	riLOC
Right Intracalcarine Cortex	rICC
Right Frontal Medial Cortex	rFMC
Right Juxtapositional Lobule Cortex	rJLC
Right Subcallosal Cortex	rSubCalC
Right Paracingulate Gyrus	rPCG

Table A.2: List of the acronyms used in Figure **??**. The brain ROIs correspond to the Harvard-Oxford anatomical atlas. (Continued)

Brain ROIs	Label
Right Cingulate Gyrus; anterior division	raCG
Right Cingulate Gyrus; posterior division	rpCG
Right Precuneous Cortex	rPC
Right Cuneal Cortex	rCC
Right Frontal Orbital Cortex	rFOC
Right Parahippocampal Gyrus; anterior division	raPhipG
Right Parahippocampal Gyrus; posterior division	rpPhipG
Right Lingual Gyrus	rLG
Right Temporal Fusiform Cortex; anterior division	raTFC
Right Temporal Fusiform Cortex; posterior division	rpTFC
Right Temporal Occipital Fusiform Cortex	rTOFC
Right Occipital Fusiform Gyrus	rOFG
Right Frontal Operculum Cortex	rFOC
Right Central Opercular Cortex	rCOC
Right Parietal Operculum Cortex	rPOC
Right Planum Polare	rPP
Right Heschl's Gyrus (includes H1 and H2)	rHG
Right Planum Temporale	rPT
Right Supracalcarine Cortex	rSupCalcC
Right Occipital Pole	rOP
Left Frontal Pole	lFP
Left Insular Cortex	lIC
Left Superior Frontal Gyrus	lSFG
Left Middle Frontal Gyrus	lMFG
Left Inferior Frontal Gyrus; pars triangularis	lIFGtriang
Left Inferior Frontal Gyrus; pars opercularis	lIFGoperc
Left Precentral Gyrus	lPreG
Left Temporal Pole	lTP
Left Superior Temporal Gyrus; anterior division	laSTG
Left Superior Temporal Gyrus; posterior division	lpSTG
Left Middle Temporal Gyrus; anterior division	laMTG
Left Middle Temporal Gyrus; posterior division	lpMTG
Left Middle Temporal Gyrus; temporooccipital part	ltoMTG
Left Inferior Temporal Gyrus; anterior division	laITG
Left Inferior Temporal Gyrus; posterior division	lpITG
Left Inferior Temporal Gyrus; temporooccipital part	ltoITG
Left Postcentral Gyrus	lPostG
Left Superior Parietal Lobule	lSPL
Left Supramarginal Gyrus; anterior division	laSG
Left Supramarginal Gyrus; posterior division	lpSG
Left Angular Gyrus	lAG
Left Lateral Occipital Cortex; superior division	lsLOC

Table A.2: List of the acronyms used in Figure **??**. The brain ROIs correspond to the Harvard-Oxford anatomical atlas. (Continued)

Brain ROIs	Label
Left Lateral Occipital Cortex; inferior division	liLOC
Left Intracalcarine Cortex	lICC
Left Frontal Medial Cortex	lFMC
Left Juxtapositional Lobule Cortex	lJLC
Left Subcallosal Cortex	lSubCalC
Left Paracingulate Gyrus	lPCG
Left Cingulate Gyrus; anterior division	laCG
Left Cingulate Gyrus; posterior division	lpGC
Left Precuneous Cortex	lPC
Left Cuneal Cortex	lCC
Left Frontal Orbital Cortex	lFOC
Left Parahippocampal Gyrus; anterior division	laPhipG
Left Parahippocampal Gyrus; posterior division	lpPhipG
Left Lingual Gyrus	lLG
Left Temporal Fusiform Cortex; anterior division	laTFC
Left Temporal Fusiform Cortex; posterior division	lpTFC
Left Temporal Occipital Fusiform Cortex	lTOFC
Left Occipital Fusiform Gyrus	lOFG
Left Frontal Operculum Cortex	lFOC
Left Central Opercular Cortex	lCOC
Left Parietal Operculum Cortex	lPOC
Left Planum Polare	lPP
Left Heschl's Gyrus (includes H1 and H2)	lHG
Left Planum Temporale	lPT
Left Supracalcarine Cortex	lSupCalcC
Left Occipital Pole	lOP