Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Deep Learning for Neuroimaging: Advancing Brain-Based Biomarkers of Autism Spectrum Disorder

Thesis submitted to the University of Dublin, Trinity College, for the degree of Doctor in Philosophy, 2023

Mélanie Garcia

Supervisor: Dr. Clare Kelly

School of Medicine, Department of Psychiatry

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Date: 13/09/2023

Signed: Mélanie Garcia

Acknowledgements

I would like to express my sincere gratitude to all the participants and organisers of the ABIDE, ABCD, HBN, and ADHD200 studies, whose contributions made this PhD project possible. I am also very thankful for the successive funding sources for my PhD research: IHC (Hypercube Institute, Paris), the School of Medicine at Trinity College Dublin for the 1252 award, and the Irish Research Council for the Government of Ireland Postgraduate Scholarship award. Thank you for your consideration and interest in this project.

I am extremely grateful to my PhD supervisor, Dr. Clare Kelly, for giving me the opportunity to engage in this research over the past four years and for her encouragement, support, empathy, openness and guidance.

Thank you to my thesis panel members Professor Rhodri Cusack and Dr. Ariel Rokem who generously agreed to provide their valuable review and feedback on this project.

I am also thankful to Professor Louise Gallagher and Professor Robert Whelan for their helpful advice during my PhD.

I feel very fortunate to have met François d'Ormesson, former founder of IHC. Thank you for your mentoring, positivity, and confidence in me.

A huge thank you to Dr. Mihaela Juganaru (one of my previous professors at Ecole des Mines de Saint-Etienne) and Dr. Jean-Philippe Charrier (my previous internship supervisor at bioMérieux) for your support and recommendations for my grant applications.

I have a special thought for Professor Jean-Marc Orgogozo (R.I.P.) who really helped me jump-start my career in the field of psychiatry. Thank you for your listening and everything you taught me, I miss our regular meetings in French Basque Country.

I want to acknowledge all the inspiring people who offered me formal or informal opportunities to grow as a young researcher, in particular: Professor Nicolas Glaichenhaus and Dr. Laetitia Davidovic, the organisers of the Imaging Interest Group at TCIN (Trinity College Institute of Neuroscience), Dr. Clara Moreau and Professor Thomas Bourgeron, Dr. Emily Graves Severance, Thierry Sibieude and Lydia Millot, Sylvie Guerin and Dominique Charme, the Rotary Club Paris Concorde, Xavier De Bernede and Cedric Mavoungou, Dr.

Marina Mischenko, Dr. Axel Orgogozo, Dr. Annabelle Blangero, Dr. Margaux Luck, and others.

My everyday life would not have been the same without my fantastic fellow (Pi-)office-mates Jiv, Xinyue, Andrew and Daniel, and labmates Sara, Emily, Chiara, and Frank. Thank you for always being positive and uplifting, no matter the Irish weather. You were my "rayons de soleil" (rays of sunshine).

I was also lucky to meet all the other thoughtful, engaging people at TCIN, including Fabiana, Camilla, Francesca, Mikey, Anna Marie, Parnian, Sharon, Celine, and others.

Thank you to former colleagues and supervisors, in particular Fred for your patience, sense of humour and guidance at the start of my programming journey, and Professor Laurence Devillers for theoretical and practical knowledge in ethics.

Thank you also to Nicolas Levillain, François Chiron, colleagues from the Bordeaux-ECNP, ENERGHY and RYLA schools, and colleagues I met during conferences.

I want to thank all my friends who have always been so supportive, especially Chloé, Sandra, Jihane, Louise, Guillemette, Mathilde, Laura, Océane, and Sandy.

I am deeply grateful to all my family, most of all my parents, my brother, and Mareling.

Finally, I thank you Géraud-Loup for your unconditional love and support, for listening to me, understanding me, and being patient with me during this PhD journey, and for agreeing to keep listening, understanding, and being patient for the rest of our lives together.

Table of Contents

List of Figures and Tables

Certain titles were reduced for readability.

List of Abbreviations

ABCD: The Adolescent Brain Cognitive Development Study

ABIDE: The Autism Brain Imaging Data Exchange database

ADHD: Attention Deficit Hyperactivity Disorder

ADHD200: The ADHD200 sample

ANN: Artificial Neural Networks

ASD: Autism Spectrum Disorder

CNN: Convolutional Neural Networks, a category of deep learning algorithm

DenseNet161: Densely Connected Convolutional Networks with 161 layers

DenseNet121: Densely Connected Convolutional Networks with 121 layers

DL: deep learning

Epoch: a hyperparameter that defines the number of times that the learning algorithm has optimised the parameters on the entire training dataset.

fMRI: Functional MRI

HBN: The Healthy Brain Network study

LLM: Large Language Models

Med3dNet - Resnet50: pretrained Residual Networks model with 50 layers

ML: machine learning

MLP: Multilayer Perceptron

MRI: Magnetic Resonance Imaging

NLP: Natural Language Processing

ProtoPNet: Prototypical Part Network model

proto-V19: ProtoPNet model with a VGG19 architecture in the CNN part

proto-R152: ProtoPNet model with a ResNet152 architecture in the CNN part

proto-D161: ProtoPNet model with a DenseNet161 architecture in the CNN part

ResNet152: Residual Networks model with 152 layers

rs-fMRI: Resting-state functional MRI

sMRI: Structural MRI

TDIs: Typically Developing Individuals

VGG19: Visual Geometry Group model, a type of very deep convolutional neural network with 19 layers in the model.

Abstract

In the last decade, deep learning (DL) has revolutionised fields like speech and vision through artificial neural networks. Can DL similarly transform biological psychiatry and neuroimaging? This thesis explores that question for Autism Spectrum Disorder (ASD) using MRI data.

ASD involves a mosaic of social, communication, cognitive, and sensorimotor differences. Diagnosis relies on behavioural assessments by highly trained clinicians. This process is challenging and resource-intensive, often involving trial-and-error before optimal interventions are identified. MRI offers promise for improving diagnosis and care by revealing ASD's brain bases. But no reproducible biomarkers have emerged, likely reflecting ASD's heterogeneity, small sample sizes, unimodal data, and limitations of standard tools.

Recent advances in multivariate predictive modelling could overcome these hurdles. In particular, DL methods from other fields now show potential for neuroimaging applications. Leveraging this opportunity, this thesis had three main aims:

1. Build a DL model for rapid, accurate quality control of structural MRI data, enabling analysis of huge datasets.
2. Predict ASD from raw structural MRI scans without standard template registration, preserving sensitivity to anatomical alterations.
3. Analyse functional MRI data with Transformer models that incorporate spatial and temporal patterns, working toward prediction from 4D data.

These projects illustrated successful applications of DL in neuroimaging for ASD, while raising questions around generalisability across confounds like age and gender. Findings emphasise the continued need to refine preprocessing methods for atypical brains and quantify bias from procedural variations.

Overall, this thesis advanced reproducible pipelines for potential MRI-based biomarkers of ASD. By openly sharing code and creating a novel tool, it enabled future DL applications in neuroimaging. Follow-up work on multi-modal prediction, optimal sample sizes, expanded

categorical labels, and new DL architectures will further realise the promise of neuroimaging to improve psychiatric diagnosis and care.

Outputs

Articles

Garcia, M., Dosenbach, N., & Kelly, C. (2022). *BrainQCNet: A Deep Learning attention-based model for multi-scale detection of artefacts in brain structural MRI scans* (p. 2022.03.11.483983). bioRxiv. https://doi.org/10.1101/2022.03.11.483983

Garcia, M., & Kelly, C. (2022). Towards 3D Deep Learning for neuropsychiatry: Predicting Autism diagnosis using an interpretable Deep Learning pipeline applied to minimally processed structural MRI data (p. 2022.10.18.22281196). medRxiv. https://doi.org/10.1101/2022.10.18.22281196

Garcia, M., & Kelly, C. (2023). *Optimising your reproducible neuroimaging workflow with Git*. OSF Preprints. https://doi.org/10.31219/osf.io/jqwpv

Ramduny, J.*, Garcia, M.*, & Kelly, C. (2023). *Establishing a reproducible and sustainable analysis workflow*. OSF Preprints. https://doi.org/10.31219/osf.io/rcxg8

* means equal contribution

Abstracts and Posters

Garcia, M., & Kelly, C. (2023) *Towards modeling brain spatio-temporal activity with Transformers: an application to predict Autism* - School of Psychology Symposium 2023, Trinity College Dublin. The Poster is in **Appendix 7**.

Garcia, M., & Kelly, C. (2022) *Towards building an interpretable predictive tool for ASD with 3D Convolutional Neural Networks* - The Organization for Human Brain Mapping 2022. The Poster is in **Appendix 8**.

Garcia, M., & Kelly, C. (2021) *Deep attention model for local detection of artifacts on brain structural MRI scans* - The Organization for Human Brain Mapping 2021. The Poster is in **Appendix 9**.

Garcia, M., Orgogozo, J.-M., Kelly, C., & Luck, M. (2019) *Towards Autism detection on brain structural MRI scans using deep unsupervised learning models* - Medical Imaging meets NeurIPS 2019. The Poster and Abstract are in **Appendix 10**.

Ethical considerations

Below is a statement detailing the careful consideration I have given to the ethical implications of the PhD research.

In every project, I meticulously considered the Ethics Guidelines for Trustworthy AI presented by the High-Level Expert Group on AI in April 2019. According to the guidelines, trustworthy AI should be: (1) lawful, (2) ethical, (3) robust.

Data Protection:

The three databases used in the project - ABIDE 1 and 2, ADHD200 and HBN - are shared by the International Neuroimaging Data-sharing Initiative (INDI). Prior to being shared by INDI, datasets must meet certain ethical requirements. Each dataset was required to be fully de-identified and anonymized data in accordance with the U. S. Health Insurance Portability and Accountability Act (HIPAA). In addition, all the datasets were required to have been collected following the local (e.g., Institutional Research Board; IRB) regulations on ethics and data protection. Each research group also tailored specific agreements related to data reuse for their participants. For these three datasets, data usage is unrestricted for non-commercial research purposes, it is openly shared with the scientific community under the licence Creative Commons BY-NC-SA. My work with these open data is approved by the School of Psychology Research Ethics Committee (see **Appendix 6**).

Data from the ABCD study (Clark et al., 2018; Volkow et al., 2018) were fully de-identified and anonymized, and each data-collecting site obtained informed consent from participants and their parents/guardians. The ABCD study developed guidelines for ethical considerations to be applied by each data-collecting site, and organised a hierarchy of workgroups who assessed whether each step of the collection process conformed to the ABCD guidelines (Clark et al., 2018). Data from the ABCD study were used under a Data Agreement between Trinity College Dublin and Washington University and is also approved by the School of Psychology Research Ethics Committee (see **Appendix 6**).

AI ethics law:

At the beginning of the PhD, I paid attention to AI regulations in Health in Europe. AI products and services already on the market are regulated by legislation such as the GDPR, the Health Research Regulations 2019, the Data Protection Act 2018, the Copyright and Related Rights Act 2000, and the European Union (Protection of Trade Secrets) Regulations 2018. I followed all the updates regarding AI ethics laws in Europe and in Ireland (https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/ireland). As stated above, however, all data used in this project were irrevocably anonymised prior to my accessing them - no personal data were analysed in this project.

AI robustness:

In all the empirical studies, I intended for my analysis pipelines to be highly rigorous.

It is necessary to build robust models and to be accurately transparent with regard to the data and algorithms used to build every model, and with regard to the strengths and weaknesses the model might have.

At every model-construction step, I was diligent with data preparation, with the interpretability of the algorithm, and with the level of quality of my code, to maximally prevent misinterpretations that could negatively impact our knowledge and understanding of Autism. Therefore, throughout the PhD project I made it my duty to be aware of the current literature on ASD neuromarkers and to integrate new findings into my work. I also ensured that I was highly proficient in MRI processing, deep learning, programming and cloud computing.

My skills in statistics and programming grew dramatically during the PhD, thanks to practice and continuous training. As a result, some differences may be evident across the empirical studies. These differences are evidence of my own evolution and progression as a researcher and demonstrate how valuable the PhD has been for me.

Collaborating and conflict of interests:

At the beginning of every collaborative project, I identified and reported all interests, and subsequently kept an eye out for new ones that might emerge over the course of the project. Prior to beginning work, I organised meetings to define a project management plan, comprising deliverables and milestones, how the tasks will be shared, the authorship and everyone's interests.

Autism:

By presenting this work, I do not intend to stigmatise autistic people. I am fully aware of current debates on neurodiversity and questions about whether, instead of a disorder or a condition, Autism should be considered a divergent branch of neurodevelopment that may confer autistic people with a different way of thinking and atypical sensoriality.

By doing this research, I intend to help to improve global understanding of these differences for better inclusion of Autistic people, and to improve and personalise care for people with specific needs.

I did my best to use the best wording possible through all of my projects, and my vocabulary evolved all along the PhD journey. I apologise in advance if certain words are shocking, stigmatising, or seem inappropriate to people with Autism. I remain open to suggestions about how to adapt definitions and wording as necessary.

1. Introduction

“When you have seen a child with autism, you have seen one child with autism.”

Lorna Wing, cited by Bourgeron (2023).

Autism Spectrum Disorder (ASD) is a prevalent neurodevelopmental divergence characterised by heterogeneous clinical presentations and unclear biological underpinnings. In the US, about 1 in 36 children has been identified with ASD (Maenner, 2023). In Europe, the estimation is of 1 in 71 children (Sacco et al., 2022). While neuroimaging has provided clues into atypical neural patterns, methodological challenges have hindered biomarker discovery and translational insights. Recent trends, however, engender optimism. Open scientific data sharing has enabled unprecedented sample sizes, coinciding with advances in multivariate techniques like machine learning. Realising the full potential of these synergies requires navigating myriad sources of individual variability and thoughtfully applying cutting-edge analytics. This thesis aims to advance efforts to elucidate robust, generalisable MRI-based biomarkers for ASD by developing optimised pipelines integrating DL algorithms.

Dimension reduction of high-dimensional neuroimaging data holds promise for enhancing biological psychiatry. But fully realising the potential of advanced analytics requires ongoing advances in critical areas like standardisation, integration, and interpretability. This thesis contributes methodological optimizations aimed at discovering ASD biomarkers, while promoting open science and reproducibility. More broadly, this work aligns with the evolving zeitgeist (“spirit of the age”) in computational psychiatry - embracing scale, heterogeneity, and cross-disciplinary innovation. The future of the field rests in building synergies to translate signals from diverse methodological noise into clinical insights that tangibly improve patient outcomes.

1.1. ASD Description and Phenotype

1.1.1. Origins and diagnosis

Autism Spectrum Disorder (ASD) is a common neurodevelopmental profile, characterised by social communication challenges and restrictive, repetitive behaviours (APA, Diagnostic and Statistical Manual of Mental Disorders (DSM-5®), 2013; Pierce et al., 2019). The prevalence of ASD diagnoses has risen steadily over recent decades (Buescher et al., 2014; Christensen et al., 2019; Reports on the Prevalence of Autism in Ireland and a Review of the Services for People with Autism, 2018; Zeidan et al., 2022), posing significant challenges for families, educators, and clinicians. While many individuals with ASD are intellectually able, they face disadvantages in social, educational, and vocational outcomes (Heasman, 2017; Heasman & Gillespie, 2018; Milton et al., 2018; Bird & Flint, 2019).

ASD is highly heterogeneous, with a strong genetic basis but unclear biological mechanisms (Bourgeron, 2015; Ecker et al., 2015; Lee et al., 2019; Miller et al., 2019; Nakagawa et al., 2019; Pagani et al., 2019; Ruzzo et al., 2019; Satterstrom et al., 2020; Schork et al., 2019; Silva et al., 2019; The Brainstorm Consortium et al., 2018; Yoon et al., 2020; Zhang et al., 2020). It is likely that the underlying biological mechanisms of ASD are associated with differences in brain development pathways compared to neurotypical development (Dickie et al., 2018; Ecker et al., 2015; Emerson et al., 2017; Fishman et al., 2018; Ha et al., 2015; Haar et al., 2016; Heinsfeld et al., 2018; Kishida et al., 2019; Lake et al., 2019; McKinnon et al., 2019; MRC AIMS Consortium et al., 2020; Pagnozzi et al., 2018; Pereira et al., 2018; Riddle et al., 2017; Sha et al., 2019; Subbaraju et al., 2017; Yang et al., 2016; Zheng et al., 2021). Environmental factors may interact with genetic risks (Ecker et al., 2015; Ha et al., 2015). Phenotypic variation is high, as no behavioural or biological subtypes have been firmly identified (Baker et al., 2019; Elibol et al., 2016; Fishman et al., 2018; Jiang et al., 2018; Lake et al., 2019; McKinnon et al., 2019; Milton et al., 2018; Walbrin et al., 2018; Wolfers et al., 2019). Co-occurring diagnoses like anxiety further complicate the picture (Allsopp et al., 2019; Kushki et al., 2019; Miller et al., 2019a, 2019b; Schork et al., 2019; Sha et al., 2019; Silva et al., 2019; The Brainstorm Consortium et al., 2018).

Gold standard ASD diagnosis relies on specialist behavioural assessments around age 3 years (Lord et al., 1989, 1994; Van ’T Hof et al., 2021). However, limited availability of experts leads to long wait times (O’Regan, 2023). Earlier screening methods exist but are not widely implemented (Emerson et al., 2017; Guthrie et al., 2019; Zuckerman et al., 2017; Zwaigenbaum et al., 2007), despite evidence that early intervention improves outcomes (Clark et al., 2018; Dawson & Burner, 2011).

Factors like gender, comorbidities, and age-related changes challenge efforts to elucidate ASD's neurobiological roots. For example, the marked gender imbalance in diagnosis (3-10:1 male predominance) (Fombonne, 2009; Loomes et al., 2017) hinders the study of gender influences on ASD neurobiology. High comorbidity rates confound attempts to identify ASD-specific neural correlates (Ecker et al., 2015; Kushki et al., 2019). Further, symptom profiles and neural patterns change across development in heterogeneous ways (Sanders, 2015; Van Wijngaarden-Cremers et al., 2014; Wolfers et al., 2019). However, in this thesis project, it was assumed that, these factors notwithstanding, Autistic people share common characteristic patterns in the brain that can be identified using multivariate analytic approaches.

1.1.2. Influence of gender

The significant gender imbalance in ASD diagnosis poses challenges for understanding its neurobiological bases. ASD is diagnosed at a rate of 3:1 (boys:girls) globally, and a rate of 10:1 in those without intellectual disability (Fombonne, 2009; Loomes et al., 2017). Girls meeting diagnostic criteria often go unidentified or misdiagnosed due to differing symptom presentation from boys and increased ability to mask difficulties (Cazalis, 2017; Dean et al., 2017; Kirkovski et al., 2013; Loomes et al., 2017; Zeidan et al., 2022). Current assessment practices may be ill-suited for girls (Beggiato et al., 2017; Van Wijngaarden-Cremers et al., 2014; Zeidan et al., 2022). The scarcity of diagnosed females hinders neuroimaging research on gender influences. However, growing open-science datasets now provide sufficient female samples (~100 with ASD) to enable preliminary investigations (Alexander et al., 2017; Bellec et al., 2017; Di Martino et al., 2014, 2017).

These observations prompt us to ask whether neuroimaging biomarkers differ between males and females with ASD? If so, how does this inform our understanding of gender-dependent neurophenotypes?

1.1.3. Comorbidities

High rates of comorbid diagnoses are common in ASD and further confound the search for ASD biomarkers, including ADHD (~30-50%), anxiety, depression (30-70%), and other neurological or developmental disorders (Ecker et al., 2015; Ghaziuddin et al., 2002; C. Gillberg, 2010; I. C. Gillberg et al., 2016; Simonoff et al., 2008). However, exact comorbidity rates vary across studies (Ghaziuddin et al., 2002; Gillberg et al., 2016; Simonoff et al., 2008).

Extensive symptomatic, genetic, and neural overlap exists between ASD and psychiatric diagnoses like ADHD, OCD, and schizophrenia (Ecker et al., 2015; Kushki et al., 2019; The Brainstorm Consortium et al., 2018). This challenges the current model of discrete psychiatric diagnoses (Ecker et al., 2015; Kushki et al., 2019) and complicates identification of ASD-specific biomarkers. Disentangling disorder-specific neural correlates becomes challenging (Ecker et al., 2015; Kushki et al., 2019) and analyses ignoring comorbidity risk finding non-specific brain patterns (Ecker et al., 2015).

Large datasets with deep phenotyping are required to address comorbidities. What methodological strategy can we build to study the impact of comorbidities on the identification of ASD?

1.1.4. Age and development

Autism is a neurodevelopmental disorder with clinical profiles changing across the lifespan (APA, Diagnostic and Statistical Manual of Mental Disorders (DSM-5®), 2013; Ecker et al., 2015; Ha et al., 2015; Sanders, 2015; Van Wijngaarden-Cremers et al., 2014; Wolfers et al., 2019). For example, an analysis of electronic records identified distinct, age-dependent ASD symptom trajectories from ages 0-15 years (Elibol et al., 2016).

If symptoms map to specific neural patterns, these shifting profiles over development likely complicate our ability to detect robust biomarkers. Each individual also grows up in unique environmental contexts influencing gene expression, brain and behaviour. Such factors include medication, social determinants, and potential traumas (Bourgeron et al., 2023).

While ASD involves atypical neurodevelopment (Emerson et al., 2017; Ecker et al., 2015; Ha et al., 2015), precisely linking symptoms, brain changes, and age remains limited. This once again calls for large datasets with deep phenotyping, including environmental and developmental contexts.

Given the age dependence of ASD, we can wonder how to best study the influence of age on the detection of ASD with deep learning?

The challenges of diagnosing and subtyping ASD repeatedly underscore the need for large, deeply phenotyped datasets. The costs involved in neuroimaging research have typically precluded the generation of datasets on this scale, nor did the methods exist to capitalise on such data. As a result, small sample sizes, limited clinical and phenotypic measures, and a reliance on largely univariate statistical methods have likely hindered efforts to identify reliable neural correlates.

However, the tide is turning with the emergence of open large-scale repositories like ABIDE (Di Martino et al., 2014, 2017), ABCD (Volkow et al., 2018), UK Biobank (Sudlow et al., 2015), and Healthy Brain Network (Alexander et al., 2017). These datasets provide unprecedented sample sizes with extensive clinical, behavioural, and environmental phenotyping beyond conventional neuroimaging resources. Paired with new multivariate methods from machine learning (ML), these mega-datasets engender optimism for reproducible pattern discovery amidst immense heterogeneity. In this project, I hypothesised that these large-scale database projects would provide sufficient data with enough variability to build robust algorithms, enabling advances toward robust, generalisable MRI-based biomarkers for ASD.

The quest for ASD biomarkers underscores a transformative opportunity for neuroimaging. Fulfilling the long-held promise of illuminating brain-behaviour relationships may finally be possible through synergistic advances in resources and techniques.

1.2. ASD detection and characterisation using brain MRI data

1.2.1. Context

Neuroimaging studies reveal an atypical developmental trajectory in ASD (Dickie et al., 2018; Ecker et al., 2015; Emerson et al., 2017; Fishman et al., 2018; Ha et al., 2015; Haar et al., 2016; Heinsfeld et al., 2018; Kishida et al., 2019; Lake et al., 2019; McKinnon et al., 2019; MRC AIMS Consortium et al., 2020; Pagnozzi et al., 2018; Pereira et al., 2018; Riddle et al., 2017; Subbaraju et al., 2017; Yang et al., 2016; Zheng et al., 2021). Autistic toddlers exhibit accelerated brain overgrowth and enlarged cortical surface area (Ecker et al., 2015; Ha et al., 2015; MRC AIMS Consortium et al., 2020; Nakagawa et al., 2019; Pagnozzi et al., 2018; Zhang et al., 2020), which reverses by adulthood with decreased brain volume and accelerated cortical thinning (Ecker et al., 2015; Pereira et al., 2018; Yang et al., 2016; Zheng et al., 2021). This initial overgrowth may disrupt white matter development and contribute to altered morphology and connectivity across the lifespan (Ecker et al., 2015; Pereira et al., 2018). While the evidence to date is compelling, the precise characterisation of this developmental trajectory awaits further longitudinal data (Lee et al., 2021; Raznahan et al., 2013), and interrogation of the influence of factors such as gender and the presence of psychiatric comorbidities, as discussed above. Further, methodological and cohort factors influencing population norms need consideration when interpreting volumetrics (Lee et al., 2021; Raznahan et al., 2013). In the future, integrating genetic and environmental data could elucidate growth dysregulation mechanisms.

Across age groups, divergent structure is consistently found in frontotemporal, frontoparietal, limbic, and midline regions implicated in social, emotional, and behavioural functions affected in ASD (Ecker et al., 2015; Ha et al., 2015; Pereira et al., 2018). Atypical cortical folding, influenced by early overgrowth, is also observed (Ecker et al., 2015; Ha et al., 2015; MRC AIMS Consortium et al., 2020; Nakagawa et al., 2019; Pereira et al., 2018;

Yang et al., 2016; Zheng et al., 2021). Finally, studies using functional MRI reveal differences in activation and connectivity (Ecker et al., 2015; Fishman et al., 2018; Ha et al., 2015; He et al., 2020; King et al., 2019; Pagnozzi et al., 2018; Pereira et al., 2018), though some findings exhibit poor reproducibility, likely due to small samples and methodological inconsistencies (Ecker et al., 2015; Ha et al., 2015; He et al., 2020). Further, as noted above, many of these neural differences are not ASD-specific and are observed in other neurodevelopmental and psychiatric diagnoses such as ADHD, OCD, and schizophrenia, highlighting the need to model brain-behaviour relationships to isolate neural correlates of ASD-specific behavioural dimensions (Sha et al., 2019).

1.2.2. Preprocessing methods

While neuroimaging has yielded clues about ASD's neurobiological roots, robust biomarkers remain elusive (Raznahan et al., 2013; Ecker et al., 2015; Ha et al., 2015; Sha et al., 2019; He et al., 2020; Lee et al., 2021). As outlined above, methodological challenges persist, including small samples, cross-sectional designs, clinical heterogeneity, and developmental factors.

Additional issues arise in MRI data acquisition and analysis. Scan quality variation, especially from head motion (Backhausen et al., 2016; Ecker et al., 2015; Gilmore et al., 2019; Ha et al., 2015; Reuter et al., 2015; White et al., 2018), requires rigorous quality assessment. Large open datasets (Alexander et al., 2017; Bellec et al., 2017; Di Martino et al., 2014, 2017; Sudlow et al., 2015; Thompson et al., 2020; Volkow et al., 2018) demand automated quality control, but standard methods are lacking (Backhausen et al., 2016).

Preprocessing techniques like template registration may introduce confounds by obscuring group differences or reducing reproducibility (Horien et al., 2022). Harmonisation techniques developed in neurotypical participants could have similar effects (Horien et al., 2022).

Overall, poor standardisation of quality control and preprocessing likely contributes to inconsistent findings (Dadi et al., 2019; Ecker et al., 2015; Heinsfeld et al., 2018; Horien et al., 2022). Recent efforts like BIDS, MRIQC, fMRIPrep, QSIprep, and, more globally, the

NiPreps project (https://www.nipreps.org/), aim to establish standards and enhance reproducibility (Cieslak et al., 2021; Esteban et al., 2017, 2019; Gorgolewski et al., 2016, 2017).

Several questions remain to be explored:

- What quality control methods are optimal for large multisite datasets vs. smaller single site data?
- How can preprocessing choices avoid obscuring potential group differences or reducing reproducibility?
- Is it possible to build a better preprocessing pipeline on MRI data to be able to study more accurately the brain characteristics of ASD?

I address these questions in my first empirical study - **Chapter 3**.

1.2.3. Machine Learning approaches

Combined with the emergence of large-scale open science data repositories, recent advances in data analysis techniques are beginning to transform neuroimaging and biological psychiatry. Machine learning in particular offers multivariate analytical advantages over univariate techniques. Autism researchers have capitalised on ML's predictive capacity to build diagnostic classifiers from MRI (Dekhil et al., 2020; Ecker et al., 2015; Jiang et al., 2018; Kunda et al., 2023; Lake et al., 2019; Pagnozzi et al., 2018; Retico et al., 2016; Riddle et al., 2017; Subbaraju et al., 2017; Wolfers et al., 2019; Zabihi et al., 2019).

However, many early studies lacked independent validation due to small samples (Pagnozzi et al., 2018; Traut et al., 2021). When properly validated, ML approaches achieve moderate prediction accuracy of 65-75% for ASD classification (Dekhil et al., 2020; Kunda et al., 2023; Pagnozzi et al., 2018; Retico et al., 2016; Wolfers et al., 2019). A large multisite challenge further demonstrated 70-80% accuracy, though performance declined when the algorithm was tested on novel sites (Traut et al., 2021).

Although promising, ML has limitations. Large samples are required, and available data may still be insufficient given the heterogeneity of ASD and confounds like comorbidities (Traut et al., 2021). Multisite differences and derivative inputs (e.g. volumetrics) can further bias results (Horien et al., 2022). More recently, deep learning has been explored to mitigate these challenges. DL can learn predictive features directly from minimally processed data, reducing confounds caused by preprocessing steps (LeCun et al., 2015). But DL has its own challenges including architecture optimization, generalisability, reproducibility, and computational demands (LeCun et al., 2015).

1.2.4. Deep Learning approaches

DL is a machine learning approach that learns hierarchical, multi-scale representations from raw data (LeCun et al., 2015). By minimising preprocessing, DL can learn predictive features directly (LeCun et al., 2015). Various DL architectures exist including Multilayer Perceptron (Hastie et al., 2009), Convolutional Neural Networks for images (Lecun et al., 1998), and Recurrent networks for sequences (Rumelhart et al., 1988). Since the field is constantly evolving, new types of DL algorithms that may be revolutionary in certain fields may emerge in the future, just like the development of Transformers architecture in 2017 (Vaswani et al., 2017).

DL has shown initial promise for MRI-based ASD prediction, achieving accuracies of 65-75% (Arya et al., 2020; Dekhil et al., 2020; Heinsfeld et al., 2018; Hu et al., 2020; Khosla et al., 2019; Lu et al., 2020; Traut et al., 2021; Wang et al., 2020). However, a recent challenge found that DL models tended to overfit compared to ML approaches (Traut et al., 2021).

DL is similar to ML in terms of its data demands, needing large samples to mitigate factors such as phenotypic and clinical heterogeneity. It also remains sensitive to input quality and preprocessing biases. Architectural complexity introduces challenges like overfitting and intensive computation (Traut et al., 2021). Sharing code and parameters openly for reproducibility is difficult.

This PhD aims to address these limitations by designing DL pipelines leveraging large open datasets. Goals include boosting predictive performance, enhancing model interpretability, and promoting reproducible practices.

1.3. PhD Project

1.3.1. Importance of the proposed research

Developing interpretable DL models that identify ASD neuroimaging biomarkers could advance precision psychiatry. Such tools could aid diagnosis, inform individualised interventions, and elucidate neural-behavioural links. This could benefit clinicians, educators, families, and autistic individuals themselves.

For example, linking neurobiology to autistic traits could improve societal empathy and reduce discrimination in social and vocational settings (Bird & Flint, 2019; Heasman, 2017; Heasman & Gillespie, 2018; Milton et al., 2018). Characterising early neural patterns may enable earlier intervention and improved outcomes (Clark et al., 2018; Dawson & Burner, 2011; Rogers et al., 2014; Van 'T Hof et al., 2021). Models could also track brain changes during care and education process.

More broadly, this work aligns with evolving efforts to integrate neuroscience, AI, and genomics to better characterise mental health conditions based on underlying mechanisms. Advanced analytics hold promise for precision medicine but require continued advances in techniques like interpretability to be clinically applicable.

1.3.2. Research aims and objectives

The overall objective of the PhD project is to advance research on MRI-based biomarkers of Autism by designing new analytical pipelines that include DL algorithms.

To achieve this aim, I used several large open science databases (described in **Chapter 2**), and have openly shared all code, to maximise the value of this work for the community.

Other goals were to:

- Improve existing pipelines for MRI data preprocessing;
- Create predictions of ASD diagnosis from two modalities of MRI data: structural data and resting-state functional data;
- Tailor new DL pipelines to each MRI modality;
- Boost the acceptability of DL applications in medicine by improving the explainability and interpretability of models. This was achieved by finding and implementing methods to explain DL models and to interpret the brain patterns driving to prediction outcomes;
- Participate in and contribute to open science;
- Help raise the current standards of reproducibility for neuroimaging research;
- Share high quality, readable code for better reusability;
- Grow as a young researcher by developing my skills in psychiatry, DL, communication, leadership, management, and, more globally, in topics related to technology and Health.

1.3.3. Thesis plan

The work performed for this thesis is described in five chapters:

- **Chapter 3** describes my first empirical study, which aimed to build a fast, reliable quality control pipeline for brain structural MRI data using DL. The best-performing algorithm was integrated into an open BIDS-app, which was shared with the neuroimaging community.
- **Chapter 4** aimed to build an interpretable pipeline for the prediction of detection of ASD diagnosis from structural MRI data using DL. Key innovations were (1) the model was trained on minimally preprocessed data (no registration to template), (2) the characterisation of regions that contributed to the prediction of ASD (interpretability), and (3) the examination of how age, gender, and Comorbidities influenced the characterisation of such regions.

- My third empirical study is described in **Chapter 5**. This project aimed to build a new DL approach to prediction from resting state fMRI data; we applied it to the detection of ASD, but also to gender, age, and performed an analysis of the brain areas that contributed to prediction outcomes.
- **Chapter 6** describes efforts aimed at promoting reproducibility and fostering better practices in neuroimaging research.
- **Chapter 7** summarises my "extra-curricular" projects - the summer schools I attended during the PhD, and several projects undertaken for these school programmes, which helped me grow as a young researcher in psychiatry and in AI.

Finally, the General Discussion and Conclusion in **Chapter 8** provides a global summary of the thesis and of each study. I give an overall interpretation of all the results obtained during the PhD, as well as the implications this project has for the broader community. Finally, I discuss all the limitations that this project has, and I establish a synthetic list of recommendations for future work related to the main thesis topics.

2. Data and Methods

2.1. Data

This PhD project made use of five large (totalling more than 5000 individuals) publicly available datasets comprising phenotypic and neuroimaging data: Autism Brain Imaging Data Exchange (ABIDE) version 1 (Di Martino et al., 2014) and 2 (Di Martino et al., 2017), Healthy Brain Network (HBN) (Alexander et al., 2017), Adolescent Brain Cognitive Development (Volkow et al., 2018), and Attention Deficit Hyperactivity Disorder 200 (ADHD 200) (Bellec et al., 2017).

Table 2.1 describes the various datasets and indicates the studies in which these datasets were used.

Dataset	MRI type	Number of patients	Related PhD Chapter
ABIDE 1 ABIDE 2	structural and resting-state functional	1112 (539 ASD) 1114 (521 ASD)	3,4,5
ABCD	structural and resting-state functional	2141 (QC)	3
ADHD200	structural and resting-state functional	973	3,4
HBN	multimodal: resting-state and task functional	2505	5

Table 2.1: Description of datasets

- ABIDE 1 (http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html)

ABIDE I (Di Martino et al., 2014) involved 17 international sites, sharing previously collected resting state functional magnetic resonance imaging (R-fMRI), anatomical and phenotypic datasets made available for data sharing with the broader scientific community. This effort yielded 1112 dataset, including 539 from individuals with ASD and 573 from typical controls (ages 7-64 years, median 14.7 years across groups). This aggregate was released in August 2012. Its establishment demonstrated the feasibility of aggregating resting state fMRI and structural MRI data across sites. In accordance with HIPAA guidelines and 1000 Functional Connectomes Project / INDI protocols, all datasets have been anonymized, with no protected health information included.

- ABIDE 2 (http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html)

ABIDE II (Di Martino et al., 2017) was established to further promote discovery science on the brain connectome in ASD. To date, ABIDE II has aggregated over 1000 additional datasets with greater phenotypic characterisation, particularly about measures of core ASD and associated symptoms. In addition, two collections include longitudinal samples of data collected from 38 individuals at two time points (1-4 year interval). To date, ABIDE II involves 19 sites - ten charter institutions and seven new members - overall donating 1114 datasets from 521 individuals with ASD and 593 controls (age range: 5-64 years). These data have been openly released to the scientific community in June 2016. In accordance with HIPAA guidelines and 1000 Functional Connectomes Project / INDI protocols, all datasets are anonymous, with no protected health information included.

- ADHD200 (http://fcon_1000.projects.nitrc.org/indi/adhd200/)

The ADHD-200 Sample (Bellec et al., 2017) is a grassroots initiative, dedicated to accelerating the scientific community's understanding of the neural basis of ADHD through the implementation of open data-sharing and discovery-based science. Towards this goal, we are pleased to announce the unrestricted public release of 776 resting-state fMRI and anatomical datasets aggregated across 8 independent imaging sites, 491 of which were obtained from typically developing individuals and 285 in children and adolescents with ADHD (ages: 7-21 years old). Accompanying phenotypic information includes diagnostic status, dimensional ADHD symptom measures, age, sex, intelligence quotient (IQ) and lifetime medication status. Preliminary quality control assessments (usable vs. questionable) based upon visual timeseries inspection are included for all resting state fMRI scans.

In accordance with HIPAA guidelines and 1000 Functional Connectomes Project protocols, all datasets are anonymous, with no protected health information included.

- ABCD (https://abcdstudy.org/about/)

The Adolescent Brain Cognitive Development (ABCD) Study (Volkow et al., 2018) is the largest long-term study of brain development and child health in the United States. The

National Institutes of Health (NIH) funded leading researchers in the fields of adolescent development and neuroscience to conduct this ambitious project. The ABCD Research Consortium consists of a Coordinating Center, a Data Analysis, Informatics & Resource Center, and 21 research sites across the country (see map on website), which have invited 11,880 children ages 9-10 to join the study. Researchers can track their biological and behavioural development through adolescence into young adulthood.

- HBN (https://healthybrainnetwork.org/about/)

The Healthy Brain Network (Alexander et al., 2017) is the signature scientific initiative of the Child Mind Institute. The goal of this community-centred program is to collect data from and provide diagnostic consultations to 10,000 children and adolescents (ages 5–21) in New York City to further the study of child and adolescent mental illness.

The HBN Biobank houses data about psychiatric, behavioural, cognitive, and lifestyle phenotypes, as well as multimodal brain imaging (resting and naturalistic viewing fMRI, diffusion MRI, morphometric MRI), electroencephalography, eyetracking, voice and video recordings, genetics and actigraphy (Alexander et al., 2017).

2.2. Background on deep learning algorithms

In this section, I summarise key deep learning (DL) concepts and models relevant to the work performed for this PhD.

2.2.1. Deep Learning as a child class of Machine Learning

One important point to keep in mind about deep learning (DL) is that it is a category of machine learning (ML). In other words, as an analogy with object-oriented programming, if ML was a superclass of objects (e.g., algorithms), DL would be a child class of the ML class, with similar attributes and methods, as well as overridden ones. Thus, ML and DL share similar important concepts (e.g., training and testing a model, supervised vs unsupervised approach, regression vs classification) but differ on several points, which I describe below.

The main idea behind ML is to build algorithms that combine features to model a certain phenomenon, and that can return predictions for new input data (Hastie et al., 2009). Features involved in machine learning generally require considerable engineering and knowledge about the given problem.

For a given problem, the optimal ML model is obtained thanks to an iterative learning process of optimization, specific to the type of algorithm. The learning process is called “training”, and the dataset used at this stage is called the “train set” or the "training set".

The “testing” step corresponds to the prediction of new input data - commonly called the test set - that was not used to train the algorithm. Testing is very important to assess the generalisability of a model and the replicability of outcomes (Hastie et al., 2009; He et al., 2020).

For instance, in the case of a Linear Regression, the training step consists of finding the optimal weights $(w_i)_{i\in[1,n]}$, where $i, n \in \mathbb{N}$, that make the best prediction $\hat{y}$ of a target variable y from explicative variables $(x_i)_{i\in[1,n]}$, under the equation $\hat{y} = \sum_i^n \quad w_i . x_i$.

When the target variable is known and used in the optimisation process, the model is called *supervised*. When the target variable is unknown or is not used in the optimisation process, the model is called *unsupervised*. In addition, when the target is continuous, the algorithm performs a *regression* whereas when the target is categorical, the algorithm performs a *classification*. **Figure 2.1** represents these concepts.

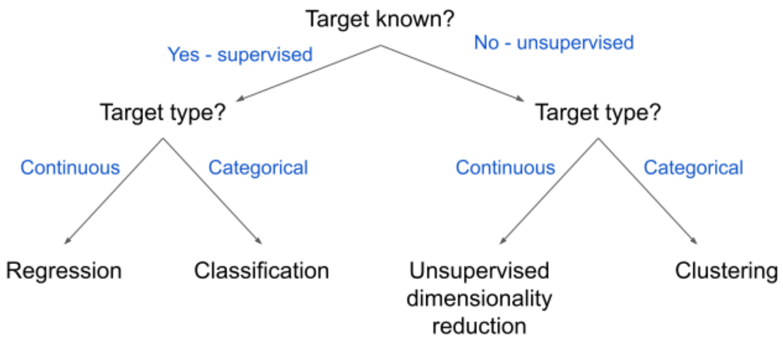


Figure 2.1: Nomenclature of algorithms in machine learning

Depending on the case (supervised or unsupervised, regression or classification), the performance of an algorithm can be evaluated with different metrics, including, for instance, accuracy for supervised classification, or mean absolute error in the case of supervised regression.

Three metrics are particularly important when using supervised classification to develop diagnostic tools: sensitivity, specificity and AUROC.

- **Sensitivity** represents the ability of the algorithm to correctly identify people with the diagnosis. Also called the True Positive Rate, it is the number of people correctly classified as having the diagnosis, divided by the total number of people with the diagnosis.
- **Specificity** represents the ability of the algorithm to correctly identify people without the diagnosis. Also called the True Negative Rate, it is the number of people correctly classified as not having the diagnosis, divided by the total number of people without the diagnosis.
- **AUROC** - area under the Receiver Operating Characteristic curve - is a metric that summarises both sensitivity and specificity. The closer to 1 the AUROC is, the better the model is.

Additional explanations about ML techniques can be found in Hastie et al. (2009).

As stated in the introduction, the idea behind deep learning is that a complex nonlinear function of variables relevant for a given problem can be learned hierarchically, and that the multi-scale relationship between variables can be learned implicitly. Designing the architecture of a DL algorithm in the shape of successive layers of analysis hence appears natural. In other words, as stated by LeCun et al. (2015), DL allows computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction.

One main advantage of DL algorithms compared to more traditional ML algorithms is that the preprocessing applied to input data is minimised (LeCun et al., 2015). The layers of the

DL algorithm are trained to learn implicit relevant features automatically for the given problem, without the need to perform demanding feature engineering.

Training such models requires new techniques because of the hierarchical structure of Artificial Neural Networks (ANN) and because millions of parameters need to be updated at each iteration step.

In particular, the training loop involves two main steps: a *forward step* and a *backward step*.

- During the *forward step*, at each iteration, the inputs are processed by the successive layers of the ANN, the computations are performed with the model parameter values of the current iteration, and outputs are returned. A loss function value for the iteration step can then be computed from the current outputs, as well as performance metrics (e.g., accuracy, AUROC). Estimating the loss function - also called the cost function - serves to guide the optimisation of the model. When the chosen performance metrics values are stable and high, training can stop.
- During the *backward step*, backpropagation is performed across the ANN in order to optimise all the parameters of the model. This optimization process is performed in a hierarchical way thanks to an optimization algorithm that takes into account the loss function value computed at the end of the forward step.

Importantly, the number of epochs is a hyperparameter that defines the number of times that the learning algorithm optimises the parameters on the entire training dataset. For instance, one epoch means that the forward step and backward step have been performed only once on the entire training set. For machine-related and regularisation reasons, it is often impossible to run one epoch at once, and a technique called “batch optimisation” is widely used to perform these steps on smaller data samples. Hence, one iteration corresponds to one forward step and one backward step performed on a batch (i.e. a sub-sample of the training set). As a consequence, batch size is another hyperparameter that plays an important role in the optimisation process.

Nowadays, there exist many optimisers, like the procedures of Stochastic Gradient Descent (Bottou et al., 2018), Adam (Kingma & Ba, 2017) or RMS Prop (Ruder, 2017) to name a few that are widely used. These optimisers all depend on a decisive hyperparameter called the learning rate, which makes the optimisation go faster or slower, and which also makes the process converge towards a local minimum or the global minimum (the one expected) of the estimated loss function. **Figure 2.2** illustrates the optimization of weights during training and the resulting loss value, and the fact that there are multiple local minima where the optimiser could become stuck. The learning rate value can be fixed or changed during the optimization.

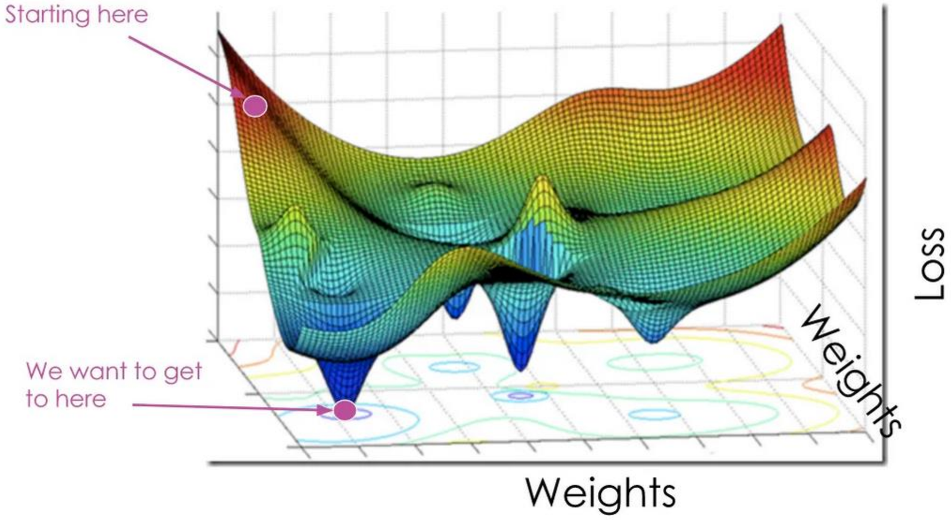


Figure 2.2: An illustration of the optimisation process: the loss function depends on the weight values. At the beginning of the optimisation process (i.e., the training), the loss value is high. At the end, the process should converge to the global minimum of the surface, which is made challenging because of the various local minima.

Image source: firsttimeprogrammer.blogspot.co.uk

This introduction to the main concepts of DL shows that there is a greater number of parameters to consider in DL than for traditional ML models. While challenging, this fact is also why DL models are potentially more powerful tools, particularly in domains where a priori knowledge is limited, such as in psychiatric neuroimaging.

For further information on the fundamentals of DL, I recommend Hastie et al. (2009), LeCun et al. (2015), Goodfellow et al. (2016).

In the following **sub-chapters 2.2.2 to 2.2.4**, I provide an overview of concepts related to three different types of ANN that are relevant to this thesis.

2.2.2. Multilayer Perceptron

A Multilayer Perceptron (MLP) is considered as the most basic deep learning model.

The basic unit of such models is called a neuron. In the neuron, the basic operation of an MLP is performed, leading to the "activation" or not of the neuron. One layer of an MLP is made of several neurons. The global architecture of an MLP consists of successive layers where each layer receives as inputs the outputs of the previous layer and returns outputs that form the inputs of the subsequent layer. **Figure 2.3** represents an MLP.

For instance, let's call an input vector $x = (x_i)_{i \in [1,n]}$ where $n > 1, n \in \mathbb{N}$. Hence, in neuron 1 of layer 1 of an MLP, the output y is : $y = f(\sum_1^n \quad w_{1i}x_i + b_1)$ where the $(w_{1i})_{i \in [1,n]}$ are the weights and $b_1$is the bias computed for neuron 1 of layer 1, and where $f: z \in \mathbb{R} \to \mathbb{R}$ is an activation function. The bias term of each neuron is optional (i.e., can be 0), and is not represented in **Figure 2.3**. An example of activation function is the ReLU function: $f: z \in \mathbb{R} \to \max(0, z)$.

With these notations, the weights of an MLP layer can be represented as a matrix $(w_{ki})_{k \in [1,m], i \in [1,n]}$ where m is the number of neurons in this layer.

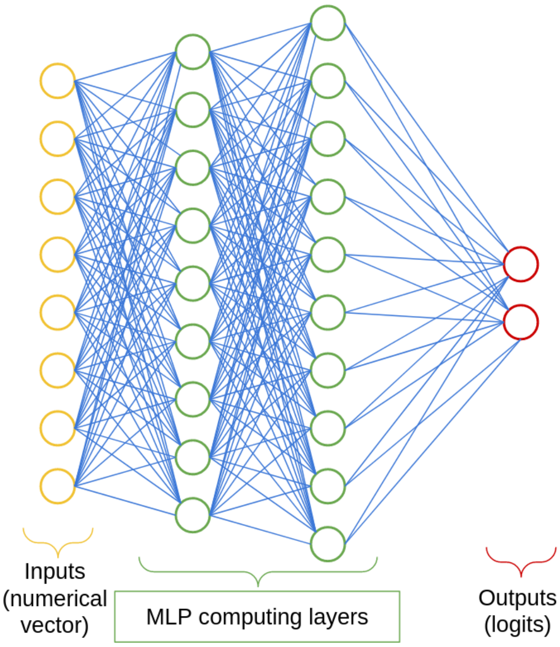


Figure 2.3: Architecture of a Multilayer Perceptron (binary classification)

2.2.3. Convolutional Neural Networks

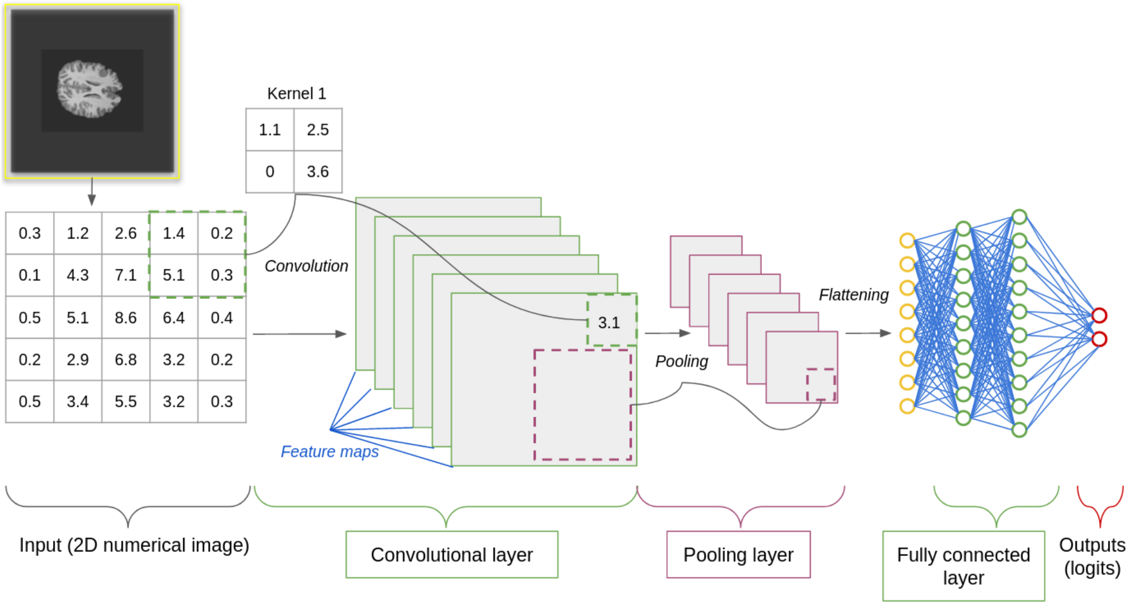
Traditional image processing pipelines include filtering in order to extract patterns. Filtering involves multiple computations by sliding a *kernel* over an input image. Mathematically, each computation consists of performing a convolution (https://en.wikipedia.org/wiki/Convolution) between a patch of the input image – a matrix or a tensor – and a kernel – also a matrix or a tensor of the same size as the patch image matrix/tensor.

Hence, it is logical to think that applying multiple different kernels can extract various features from an image. In addition, more implicit patterns may be learned by leveraging the hierarchical nature of ANN, that is by applying successive layers of multiple kernels.

This intuition led to the development of Convolutional Neural Networks (CNN), a type of DL algorithm that has revolutionised the field of object detection and recognition.

Figure 2.4 illustrates the architecture of a simple CNN applied to the task of binary classification. An input 2D image is represented as a matrix of numbers. As an example, a convolution is performed between a sub-matrix of the input and a kernel matrix of dimension 2x2. More globally, for a kernel matrix of size 2x2 $(w_{ij})_{i,j\in\{1,2\}}$and for a sub-matrix $(x_{ij})_{i,j\in\{1,2\}}$, the result of the convolution is given by: $w_{11}.x_{11} + w_{12}.x_{12} + w_{21}.x_{21} + w_{22}.x_{22}$. This operation is performed on the whole input to generate a feature map corresponding to this kernel. In particular, a stride must be defined that sets the amount of movement the kernel filter has over the image. For instance, a stride of 1 means the kernel filter moves one pixel at a time. For a given kernel size, the smaller the stride is, the larger the output feature map is. Thus, one convolutional layer consists of many feature maps that correspond to many different kernels applied to the input layer. A pooling operator is then used to reduce the dimension of the feature maps. This step plays the role of a regularizer for the network. Next, the feature maps are flattened into a numerical vector that is input to a fully connected layer (FCN). The FCN has the same architecture as an MLP, and it returns two logits in the case of a binary classification.

Generally, a CNN has many convolutional blocks (convolutional+pooling layers). Many variants of CNN architectures exist, including, for instance, VGG (Simonyan & Zisserman, 2015), ResNet (He et al., 2015), DenseNet (Huang et al., 2018).



Kernel 1

1.1	2.5
0	3.6

0.3	1.2	2.6	1.4	0.2
0.1	4.3	7.1	5.1	0.3
0.5	5.1	8.6	6.4	0.4
0.2	2.9	6.8	3.2	0.2
0.5	3.4	5.5	3.2	0.3

Figure 2.4: Architecture of a simple CNN for binary classification.

In neuroimaging, we typically work with 3D image volumes. Training a 3D CNN is possible, but technically challenging because the number of parameters to optimise soars because, relative to the 2D CNN, the input image, the kernels, and the feature maps are 3D tensors instead of 2D matrices.

2.2.4. Transformer

In late 2022, ChatGPT sparked a revolution by democratising the use of Large Language Models (LLMs). ChatGPT is derived from a particular type of Natural Language Processing (NLP) algorithm called a Transformer (Vaswani et al., 2017). This kind of algorithm is of interest because both neuroimaging and the field of NLP share a need to consider spatio-temporal features when building models.

Recently, Transformer algorithms (Vaswani et al., 2011) have risen to dominance in NLP. Transformers have also inspired new approaches in medical image processing (Bedel et al., 2022; Kan et al., 2022; Luo et al., 2021; Malkiel et al., 2022; Nguyen et al., 2020; Thomas et al., 2023; Yu et al., 2022; Zhang et al., 2021; Zhao et al., 2022), demonstrating the relevance of such algorithms in image analysis.

The original architecture of a Transformer is displayed in **Figure 2.5** Globally, it consists of an encoder part and a decoder part. In the pipeline, the inputs (e.g., the words in a text) are embedded as token numerical vectors that are summed with a positional encoding vector (that represents a function on the relative positions between the inputs). Next, there is a succession of multiple blocks that each include a multi-head attention module followed by a feed-forward network (similar to an MLP).

The idea of the attention module is that, in order to best describe the context of a word in a text and understand its meaning, paying attention only to a couple of relevant words in the text is more efficient than considering all the words with equal importance.

In that respect, each attention head computes weights of importance of all the words for a given word, and it performs this computation for all the words in the text.

Each attention module is multi-head, meaning that various sets of weights can be computed for a given word. The multi-head design serves to find more interesting patterns that could be relevant for the task, as CNNs do with multiple kernels in one convolutional layer. Further, the architecture is also hierarchical, and builds implicit patterns and features optimised for a task during training.

The encoder outputs are input to the decoder blocks that estimate a function of the encoded features in order to return consistent probabilities for a given task.

More details on the Transformer algorithm are available in the original paper by Vaswani et al. (2017).

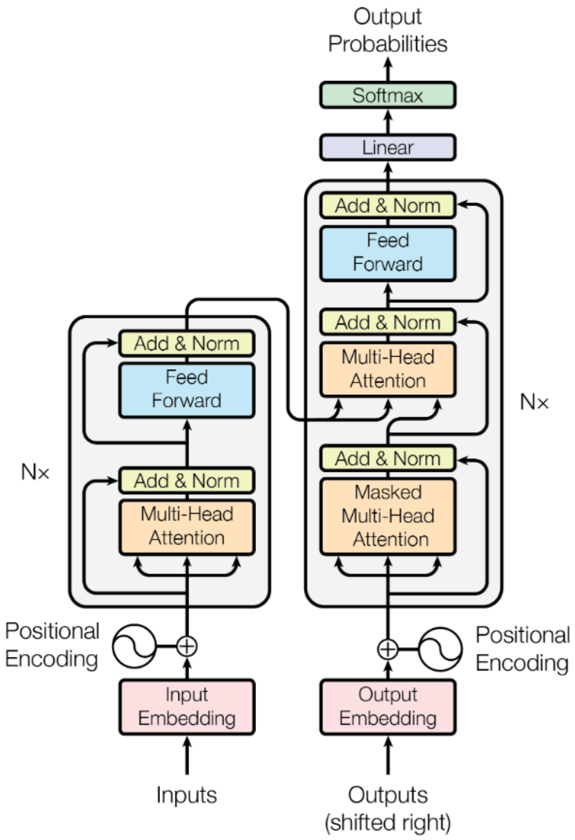


Figure 2.5: Transformer architecture - adapted from Vaswani et al. (2017).

3. Manuscript: “BrainQCNet: a Deep Learning attention-based model for the automated detection of artefacts in brain structural MRI scans”

Chapter 1 introduced the complex landscape and ambitions of this thesis. It demonstrated the value of building models using brain sMRI data to identify ASD, and ultimately, derive reproducible biomarkers for ASD. The creation and sharing of large-scale datasets has fueled great aspirations. However, upon working directly with sMRI data, I realised that standard neuroimaging preprocessing pipelines are ill-equipped to scale up to these goals. In particular, available quality control methods make this critical initial step extremely time-consuming when handling thousands of scans.

Seeking to develop an ASD detection pipeline that is reusable, scalable, and acceptable to the medical community through explainability, I wondered if DL could automate and accelerate quality control. Further, what is the best way to make such a tool reusable and improvable by the community? This study presents a new scalable model I built and shared for detecting sMRI quality that moves toward those goals.

The following corresponds to a manuscript that was revised on the basis of comments received from three reviewers, following submission to the journal *NeuroImage* in June 2022. The manuscript, authored by myself, Dr. Nico Dosenbach (Department of Neurology, Washington University School of Medicine, St. Louis), and my supervisor Dr. Clare Kelly, has now been resubmitted to the journal Imaging Neuroscience. It is re-written in English UK.

3.1. Abstract

Analyses of structural MRI (sMRI) data depend on robust upstream data quality control (QC). It is also crucial that researchers seek to retain maximal amounts of data to ensure reproducible, generalisable models and to avoid wasted effort, including that of participants. The time-consuming and difficult task of manual QC evaluation has prompted the development of tools for the automatic assessment of brain sMRI scans. Existing tools have proved particularly valuable in this age of Big Data; as datasets continue to grow, reducing execution time for QC evaluation will be of considerable benefit. The development of deep learning (DL) models for artefact detection in structural MRI scans offers a promising avenue toward fast, accurate QC evaluation. In this study, we trained an interpretable deep learning model, ProtoPNet, to classify minimally preprocessed 2D slices of scans that had been manually annotated with a refined quality assessment (ABIDE 1; n = 980 scans). To evaluate the best model, we applied it to 2141 ABCD scans for which gold-standard manual QC annotations were available. We obtained excellent accuracy: 82.4% for good quality scans (Pass), 91.4% for medium to low quality scans (Fail). Further validation using 799 scans from ABIDE 2 and 750 scans from ADHD-200 confirmed the reliability of our model. Accuracy was comparable to or exceeded that of existing ML models, with fast processing and prediction time (1 min per scan, GPU machine, CUDA-compatible). Our attention model also performs better than traditional DL (i.e., convolutional neural network models) in detecting poor quality scans. To facilitate faster and more accurate QC prediction for the neuroimaging community, we have shared the model that returned the most reliable global quality scores as a BIDS-app (https://github.com/garciaml/BrainQCNet).

3.2. Introduction

Analyses of structural MRI (sMRI) data depend on robust upstream data quality control. This is particularly true for predictive analyses incorporating machine learning techniques, where artefacts and noise may severely bias results and jeopardise generalisability (Reuter et al., 2015; Backhausen et al., 2016; White et al., 2018; Gilmore et al., 2019). Artefacts related to participant motion are a particular concern when working with very young participants, or those with neurodevelopmental diagnosis, such as Autism Spectrum Disorder and Attention-Deficit/Hyperactivity Disorder (Rauch, 2005; Nordahl et al., 2016). In such settings, data collection is usually a demanding and costly task, and it is crucial that researchers retain the maximum amount of usable data to build realistic models.

In this age of big data, manual QC evaluation of sMRI data through visual inspection is a time-consuming and monotonous task, prompting the development of new tools for automatic (full or partial) quality assessment of brain sMRI scans (Marcus et al., 2013; Shehzad et al., 2015; Glasser et al., 2016; Esteban et al., 2017; Alfaro-Almagro et al., 2018; White et al., 2018; Keshavan et al., 2019; Sujit et al., 2019). Such tools typically compute a number of diagnostic metrics using sMRI data to help researchers sort images prior to any analysis (Marcus et al., 2013; Shehzad et al., 2015; Glasser et al., 2016; Esteban et al., 2017; White et al., 2018; Alfaro-Almagro et al., 2018). For example, MRIQC (Esteban et al., 2017) has revolutionized QC of MRI data by providing a reliable and accurate machine learning-based assessment of scan quality that has been made freely available to the neuroimaging community as an open-source application. The tool generates 64 image quality metrics, including Contrast to Noise Ratio and Entropy Focus Criterion (Esteban et al., 2017), chosen on the basis of the Preprocessed Connectomes Project (PCP) Quality Assessment Protocol (Shehzad et al., 2015). The MRIQC algorithm uses machine learning to find a function that predicts a global quality score for each scan using these metrics. Although highly accessible, automated, and accurate, growth in the size of datasets (e.g., thousands to tens of thousands of sMRI scans for database such as ABCD (Volkow et al., 2018; Karcher and Barch, 2021), ENIGMA (e.g., Whelan et al., 2018) and UK Biobank (Sudlow et al., 2015), prompts a search for developments that can further reduce execution time for QC evaluation. In this study, we evaluate whether deep learning models can help advance this goal.

Deep learning models may prove particularly useful for the task of automated QC. While training a deep learning model, such as a convolutional neural network (CNN), may initially take longer than training a traditional machine learning (ML) algorithm (because there are more parameters to train), the subsequent processing and inference time is reduced compared to ML (which requires more data preprocessing before inference). This rapid inference makes DL models more scalable for Big Data applications. Studies have already successfully applied DL models to the task of sMRI QC. For example, (Sujit et al.; 2019) built a CNN model for each axis (sagittal, coronal, axial), and used a fully connected network to return a final prediction based on the intermediary predictions generated by each CNN. Although the model performed well on an multi-site test dataset, it showed poor sensitivity (0.41) when applied to an independent sample. Keshavan et al. (2019) trained a CNN model on slices of scans from a database comprising 200 scans for which expert/gold-standard manual QC was available and 722 scans judged by "citizen scientists." The AUROC for predicted labels (pass/fail) on a left-out (but non-independent) dataset was 0.99. The authors explained that this high score was due to the fact that the left-out dataset contained scans from similar sites as the training set and the fact that these scans were either very high quality or very low quality, with no intermediate quality scans included in the evaluation. These studies suggest that DL can usefully be applied to predict sMRI scan quality, but highlight the need to ensure that models are generalisable to unseen and independent data that is representative of the range of quality typically observed.

Beyond generalisability, DL models suffer from a lack of interpretability. Visual attention models offer a means to address this. These models mimic human visual attention by identifying the parts of the input image most relevant to the task. For example, when recognising a bird species from a single image, a person might rely on specific details, such as the size, colour, or shape of the beak or feathers. Attention-based DL algorithms mimic this process such that the parts of an input that contribute most to prediction (i.e., the most strongly predictive features) can be identified, leading to improved interpretability.

Here, we built on the successes of existing ML and DL approaches and leveraged the advantages of DL attention models to perform automated QC of sMRI data. Specifically, we trained the attention CNN ProtoPNet (Chen et al., 2019), as well as three standard CNNs (VGG19 - (Simonyan and Zisserman, 2015); ResNet152 - (He et al., 2015); DenseNet161 - (Huang et al., 2018)) on 2D slices of sMRI data which had been manually annotated as either good or poor quality. The process used by the ProtoPNet algorithm is similar to the one humans use when we perform manual classification of MRI scans. First, we visually search for the presence of artefacts, slice by slice, in 2D. To judge the quality of a given scan, we focus on specific features in a slice (e.g., the presence of rings or blurring) and compare these features to prototypically corrupted scans. ProtoPNet imitates this human attention process artificially, and returns interpretable output: information about the areas of the input slice identified as being poor quality or defect-free (good). The model also provides another level of interpretability: it points to prototypical cases containing the predictive features.

To train a deep learning model, it is crucial that the inputs are correctly labelled. We manually rated 980 structural MRI scans from the ABIDE 1 dataset (Di Martino et al., 2014) guided by (Backhausen et al., 2016), who described four types of artefacts. To train our algorithms, we developed an augmented training set of 270000 2D image slices, derived from 60 scans and a validation set of 1800 2D image slices from 12 scans, perfectly balanced for good quality and very poor quality slices. To identify the best-performing model, we tested the models on the remaining 908 scans from the ABIDE 1 dataset, which had been manually QCed. Finally, we evaluated the best-performing model on independent, multisite datasets: using 2141 scans from ABCD (Volkow et al., 2018; Karcher and Barch, 2021), 799 scans from ABIDE 2 (Di Martino et al., 2017) and 751 scans from ADHD-200 (Bellec et al., 2017).

A key advantage of our algorithm over existing approaches is that it requires only minimal preprocessing, which dramatically reduces the total processing time for every scan. For instance, on a machine with a processor Intel I9-10850K, RAM 64Gb DDR4, GPU Nvidia GeForce RTX 3090 24 Gb, running the CPU-version of the model took 10 minutes while the GPU version took 50 seconds. On a laptop with a processor Intel i7-7700HQ, RAM 16Gb DDR4, GPU Nvidia GeForce GTX 1060, running the CPU-version of the model took 30 minutes while the GPU version took 90 seconds. Across our independent testing datasets, we observed excellent accuracy that matched or surpassed existing automated QC algorithms. In the context of the growth of open science datasets to tens of thousands of participants, our method could offer substantial savings in terms of time and computational resources.

To facilitate fast and accurate QC prediction for the neuroimaging community, we have shared the model that returned the most reliable global quality scores, local predictions of quality, and maps and prototypes of local artefacts as a BIDS-app (https://github.com/garciaml/BrainQCNet). For the fastest performance, we recommend using the GPU version of our app.

3.3. Materials and Methods

3.3.1. Datasets

In our study, we used structural MRI data from ABIDE 1 (Di Martino et al., 2014), ABIDE 2 (Di Martino et al., 2017), ADHD-200 (Bellec et al., 2017) and ABCD (Volkow et al., 2018; Karcher and Barch, 2021). Details of each of the datasets used are provided in **Figure 3.1**.

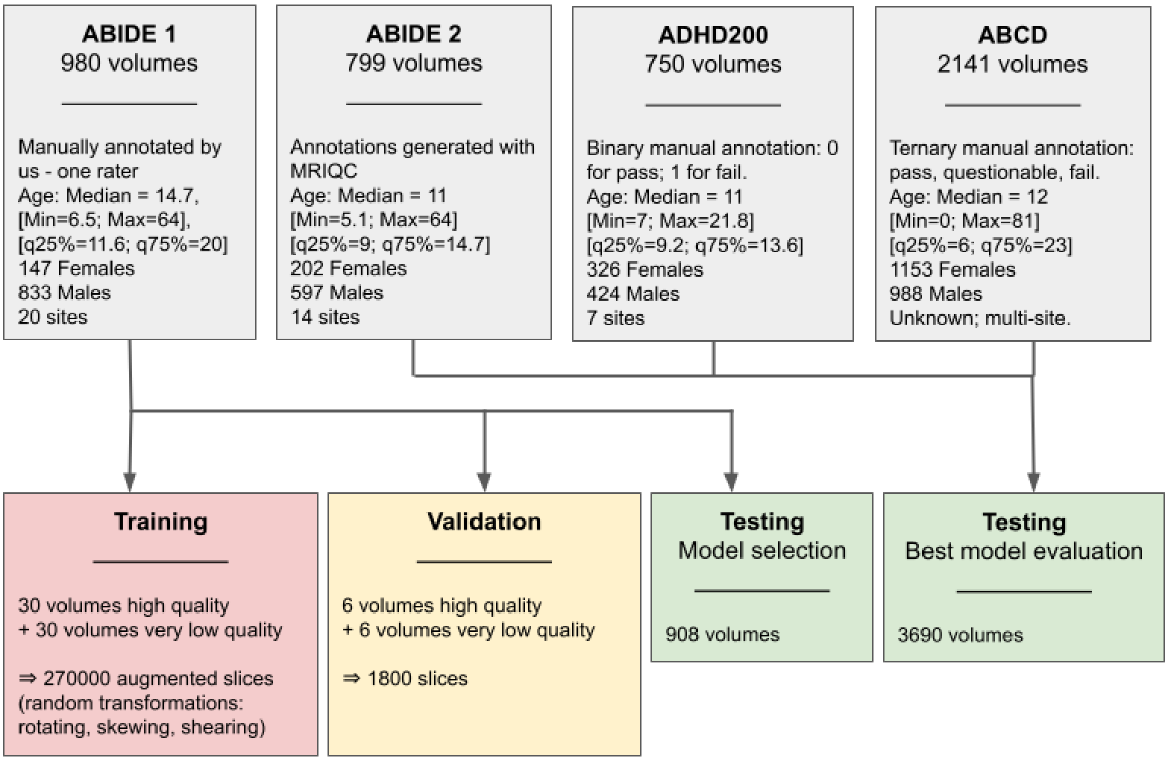


Figure 3.1: Dataset descriptions and division into training, validation, and testing sets.

3.3.2. Ethics statement

The three databases used in the project - ABIDE 1, ABIDE 2, ADHD200 - are shared by the International Neuroimaging Data-sharing Initiative (http://fcon_1000.projects.nitrc.org/).

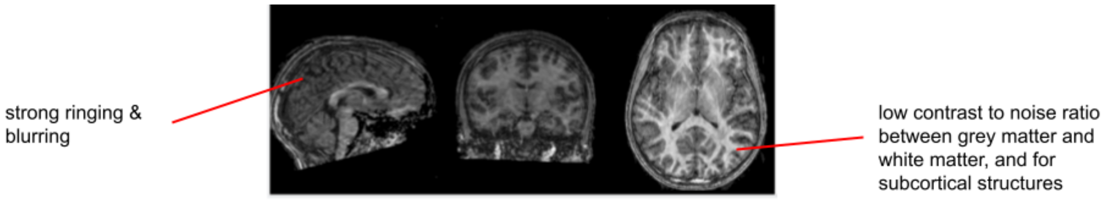
Each dataset was fully de-identified and anonymized in accordance with the US Health Insurance Portability and Accountability Act (HIPAA). All the datasets were collected and shared in accordance with the local regulations on ethics and data protection. Data usage is unrestricted for non-commercial research purposes; it is openly shared with the scientific community under the licence Creative Commons BY-NC-SA. Our work with these open data is approved by the Research Ethics Committee of the School of Psychology at Trinity College Dublin.

Data from the ABCD study were fully de-identified and anonymized, and each data-collecting site obtained informed consent from participants and their parents/guardians. The ABCD study developed guidelines for ethical considerations to be applied by each data-collecting site, and organised a hierarchy of workgroups who assessed whether each step of the collection process conformed to the ABCD guidelines (Clark et al., 2018). Data from the ABCD study were used under a Data Agreement between Trinity College Dublin and Washington University.

3.3.3. Manual Quality Control

One rater (MG) manually annotated 980 MRI scans from ABIDE 1. The annotation was guided by the work of Backhausen et al., (2016), which specified four different types of artefacts: (1) blurring (global or local), (2) ringing, (3) low contrast noise ratio between grey matter and white matter, and (4) low contrast noise ratio (CNR) of subcortical structures. For further details of the artefacts, please see the Supplementary Materials of Backhausen et al., (2016). For each scan and each artefact type, a score between 1 and 4 was given, such that a score of 1 indicates absence of that artefact while scores of 2, 3, and 4 indicate the presence of that artefact at worsening degrees of severity (where 4 is the worst).

For each 3D scan, we also noted whether each of the four artefacts was evident either locally or globally. When no artefact was observed (score = 1,1,1,1), we labelled the 3D scan as good quality (Class 0). Otherwise, we labelled the 3D scan as poor quality (Class 1; see **Figure 3.2**). Class 1 is a wide spectrum that includes scans with localised artefacts (e.g., score = 1,2,2,1) as well as very low quality, globally disrupted scans (score = 4,4,4,4 and artefacts present on all the slices of the volume). These labels - Class 0 and Class 1 - were used as the true values on which our models were trained and tested.


strong ringing & blurring
low contrast to noise ratio between grey matter and white matter, and for subcortical structures

Blurring	Ringing	CNR WM / GM	CNR subcortical structures
1 → 4 Local/Global	1 → 4 Local/Global	1 → 4 Local/Global	1 → 4 Local/Global

For Training/Validation steps:
- 1,1,1,1 Global : Highest quality
- 4,4,4,4 Global : Lowest quality

Binary classification labels:
- if 1,1,1,1: class 0 (no artifact)
- else: class 1 (presence of artifact)

Figure 3.2: Description of our system for manual sMRI scan quality annotation

3.3.4. *Training and Validation Datasets*

To create a set of images on which to train our deep learning algorithm, we identified 30 high quality scans (randomly selected from those labelled Class 0) and 30 highly corrupted/poor quality scans (randomly selected from all the scans labelled Class 1 and scored 4,4,4,4) from the 980 ABIDE 1 scans we had manually annotated. We also created a within-training validation set comprising 6 further high quality Class 0 scans and 6 very low quality Class 1 (i.e. score=4,4,4,4 and artefact present on all the slices) scans. Importantly, these training and validation sets included all the highly corrupted scans (i.e., score=4,4,4,4). We did this to provide a balanced training (same number of Class 1 and Class 0 scans) and to maximise the chances of obtaining meaningful prototypes representative of scan artefacts and corruption.

Chen et al. (2019) found that the ProtoPNet algorithm worked better on cropped images, so each 3D scan was tightly cropped to remove empty space, then converted from Nifti format to 2D PNG images (using Med2Image https://github.com/FNNDSC/med2image). For each scan there were between 150-200 2D slices for each of the 3 orientations (sagittal, coronal, axial); resulting in approximately 450-600 images per scan. The first and last 20 slices of each image stack were discarded since they contained little brain tissue. Taking a random sample of 50 slices per axis, per scan, we created a training set comprising 4500 high quality and 4500 poor quality 2D slices from all the 60 scans in the training set. A validation set of 1800 slices, also balanced for quality, was created in the same way.

Next, the training set was augmented with a set of transformations chosen randomly from a uniform distribution (using the library Augmentor https://github.com/mdbloice/Augmentor) which rotated (probability of 1, maximum left rotation of 15 degrees, maximum right rotation of 15 degrees), skewed (probability of 1, random skewing, magnitude of 0,2), left-right flipped (probability of 0,5) and sheared (probability of 1, maximum left shearing of 10 degrees, maximum right shearing of 10 degrees) the images. This yielded an augmented training set of 270000 images. Data augmentation is used to prevent overfitting in deep learning, thus improving generalisability of the algorithms.

All 2D images from good quality scans (Class 0) were defined as Label 0 and all 2D images from poor quality scans (Class 1) were defined as Label 1. The algorithm was trained to perform a binary classification between Label 0 and Label 1 2D slices using the augmented training set (*n* = 270000 slices), and validation accuracy was computed every 2 epochs (n = 1800 slices). An epoch is a hyperparameter that defines the number of times that the learning algorithm has optimised the parameters on the entire training dataset. This process of data preparation, training, and validation is summarised in **Figure 3.1**.

Since predictions were performed at the level of slices, to generate a global prediction for each scan, we computed the proportion of slices with a prediction of Label 1 (poor quality) and applied a threshold of 0.5. If greater than 50% of slices for a given scan were predicted Label 1, the entire scan was classified as Class 1 (poor quality). Below this threshold, the entire scan was classified Class 0 (good quality). We note that this is an arbitrary threshold and that different thresholds may be preferable, depending on the particular goal of subsequent analyses. Our BIDS-app (https://github.com/garciaml/BrainQCNet) returns a CSV file containing scan identifiers and probability scores, allowing for the specification of a new threshold for tailored scan classification.

3.3.5. Testing set for Model Selection

To identify the best-performing model (see Section 3.3), we generated predictions for the remaining 908 scans from ABIDE 1 (Di Martino et al., 2014), which we had manually annotated. For each scan, 450-600 2D slice images were created using the process described above (**Section 3.3.4**).

3.3.6. Independent Testing sets for Evaluation

After identifying the best-performing model, we performed an evaluation using independent testing sets comprising 2D slice images created using the process described above, for 3690 sMRI scans obtained from the following sources (see **Figure 3.1**):

- 2141 scans from ABCD (Volkow et al., 2018; Karcher and Barch, 2021). These scans had been manually QC'ed by two or more reviewers (Hagler et al., 2019), following the recommendation from the ABCD Data Analytics and Informatics Core (DAIC) (Saragosa-Harris et al., 2022), with ternary classification: pass, questionable, fail;
- 799 scans from ABIDE 2 (Di Martino et al., 2017) with QC classification generated by the MRIQC algorithm (see **Section 3.3.8**, below);
- 750 scans from ADHD-200 (Bellec et al., 2017). These scans had been manually QC'ed by 1 or 2 human raters (Bellec et al., 2017) with binary classification: pass, fail.

3.3.7. Deep Learning Algorithm

The algorithm we used, ProtoPNet (Chen et al., 2019), is a deep learning attention model that reproduces the human manual process for classifying images. The network consists of a regular convolutional neural network, followed by a prototype layer and a fully connected layer with weight matrix and no bias. Here, we compared three different architectures for the regular convolutional network: VGG19 (Simonyan and Zisserman, 2015), ResNet152 (He et al., 2015) and DenseNet161 (Huang et al., 2018). These three models are well known deep learning algorithms for image classification, and have shown good performance for 2D images (Simonyan and Zisserman, 2015; He et al., 2015; Huang et al., 2018). In machine learning, it is common to compare different types of algorithm for a given problem, to detect overfitting and to identify the best-performing algorithm (Hastie et al., 2009).

In their approach, Chen et al. (2019) constrained each convolutional filter to be identical to a latent training patch, to make every convolutional filter interpretable as visualisable prototypical image parts. In our study, the "prototypes" or "prototypical images" corresponded to the Class 0 (good quality) and Class 1 (poor quality) images of the augmented training set. The algorithm works, in part, by comparing images in the validation and test sets to parts of the prototypes. The number of images selected randomly as prototypes during each epoch of training was set to 2000.

In the ProtoPNet global architecture, the prototype layer computes similarity scores between the convolutional filters of the input image and the ones from the 2000 prototypes at a fixed epoch. The similarity scores are computed with an inverted L2 norm distance.

Chen et al. (2019) explained that given a convolutional output $z = f(x)$, the j-th prototype unit g_{p_j}in the prototype layer g_p computes the squared L^2 distances between the j-th prototype p_j and all patches of z that have the same shape as p_j, and inverts the distances into similarity scores. The result is an activation map of similarity scores whose value indicates the strength of similarity between the input image and a prototype.

Mathematically, the prototype unit g_{p_j} computes $g_{p_j}(z) = max_{\tilde{z} \in patches(z)} log((||\tilde{z} - p_j||_2{}^2 + 1)/(||\tilde{z} - p_j||_2{}^2 + \epsilon))$. The function g_{p_j} is monotonically decreasing with respect to $||\tilde{z} - p_j||_2$ (if $\tilde{z}$ is the closest latent patch to p_j). If the output of the j-th prototype unit g_{p_j} is large, then there is a patch in the convolutional output that is (in 2-norm) very close to the j-th prototype in the latent space, and this in turn means that there is a patch in the input image that has a similar concept to what the j-th prototype represents.

Next, the fully connected layer predicts the label of the input image from the 2000 similarity scores. We obtained probability scores by applying the softmax function to the output logits of the fully connected layer. In theory, this method of regularisation and comparison should improve the generalisability of the algorithm. More mathematical details of the ProtoPNet model are given in (Chen et al., 2019); **Figure 3.3(b)** illustrates its architecture in our context.

We initiated training using ImageNet (Deng et al., 2009), drawn from the model zoo of Pytorch (https://pytorch.org/serve/model_zoo.html). We used the same initialisation parameters as previous experiments (Chen et al., 2019), including 5 “warming” epochs for which no accuracy was computed (where each epoch is a step during which the algorithm is optimised by all the images of the training set). Because of the GPU memory demands of this process, optimization is achieved iteratively using small batches of data. Here, we used the same batch sizes as (Chen et al., 2019): 80 for the training and 100 for the testing phase. During training time, we validated every 2 epochs by assessing the prediction accuracy of the model for slices from the scans in the validation set.

We trained our models in a distributed way on AWS cloud instances of type p3.8xlarge and p3.16xlarge initialised with the AMI Deep Learning. The instances correspond to 4 or 8 NVIDIA V100 GPUs. We trained ResNet152 on 20 epochs and VGG19 and DenseNet161 on 30 epochs. We saved models and associated prototypes every 10 epochs.

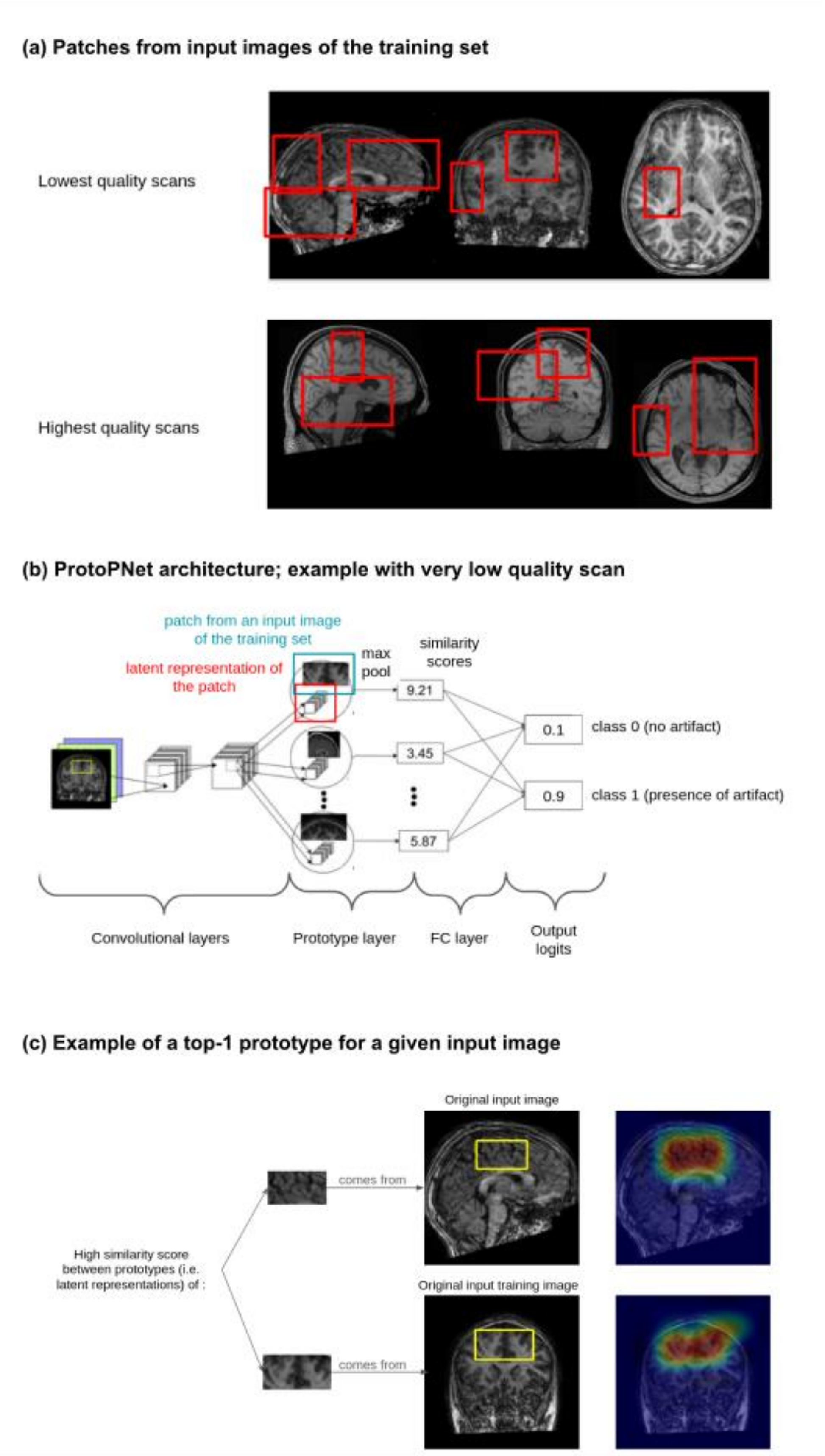

(a) Patches from input images of the training set
Lowest quality scans
Highest quality scans
(b) ProtoPNet architecture; example with very low quality scan
patch from an input image of the training set
latent representation of the patch
max pool
similarity scores
9.21
3.45
5.87
0.1
0.9
class 0 (no artifact)
class 1 (presence of artifact)
Convolutional layers
Prototype layer
FC layer
Output logits
(c) Example of a top-1 prototype for a given input image
High similarity score between prototypes (i.e. latent representations) of :
comes from
comes from
Original input image
Original input training image

Figure 3.3: The ProtoPNet approach for automatic QC of brain sMRI scans. (a) Patches taken from input 2D slices of the training set; (b) Architecture of the ProtoPNet model; (c)

Example of a top-1 prototype (i.e., the prototype from the training set with the highest score for similarity with the input patch) for a given input 2D slice.

3.3.8. MRIQC

MRIQC (Esteban et al., 2017) was conceived as a tool to permit more reliable and efficient QA/QC of MRI data through visual reports. It integrates a classifier to provide an automatic assessment of the quality of brain structural and functional MRI scans. The MRIQC classifier is based on a machine learning algorithm that was trained on a large number of metrics of quality previously extracted and computed from raw scans. As outlined in the introduction, these metrics were chosen as part of the Preprocessed Connectomes Project (PCP) Quality Assessment Protocol (Shehzad et al., 2015) to harmonise the assessment of the quality of brain MRI scans (Shehzad et al., 2015), like the signal-to-noise ratio. The output of MRIQC is a score and a binary prediction (pass/fail) for each scan.

This method is reliable (accuracy estimated to 76%±13% on new sites, using leave-one-site-out cross-validation, accuracy of 76% on a held-out dataset of 265 scans; Esteban et al., 2017), and widely employed.

Here, we used the MRIQC classifier to generate predictions of the quality of each scan on ABIDE 2 (Di Martino et al., 2017; 799 scans). We used the default MRIQC threshold for classification. In particular, we used the BIDS-app poldracklab/mriqc:0.9.6 (on DockerHub) to run the MRIQC classifier as is. We treated these MRIQC-based predictions as the “ground truth” against which we compared the results of our algorithm.

We also compared the distribution of the scores returned by MRIQC for ABIDE 1 (n = 980 scans; Di Martino et al., 2014) with the distribution of scores returned by our models. In particular, we examined the discrimination between good quality scans (score=1,1,1,1) and medium quality (artefacts present only locally on the volume and/or medium intensity artefacts) and low quality ones (score=4,4,4,4 and artefacts present on all the slices of all the volume).

3.3.9. Comparison with traditional CNN models

To provide a comprehensive evaluation of the attention model (ProtoPNet) approach, we also built three traditional CNN models for comparison. To do this, we used the pre-trained CNN models, VGG19, ResNet152, DenseNet161, drawn from the model zoo of Pytorch (https://pytorch.org/serve/model_zoo.html). We used the same training and validation sets, learning parameters, and methods described above.

3.3.10. Data and Code availability

Three of the datasets used in the project - ABIDE 1, ABIDE 2, ADHD200 - are openly shared by the International Neuroimaging Data-sharing Initiative (http://fcon_1000.projects.nitrc.org/). Access to ABCD data is available upon request (https://nda.nih.gov/abcd/request-access).

All global predictions of quality for the 4670 scans we used from the ABIDE 1 & 2, ADHD200 and ABCD databases are available through the GitHub repository: https://github.com/garciaml/BrainQCNet_paper_results.

To maximise the reproducibility of our analyses and usability of our model, all the code to build the BIDS-apps is available on two other GitHub repositories (https://github.com/garciaml/BrainQCNet_CPU for users of CPU machines and https://github.com/garciaml/BrainQCNet_GPU for users of GPU machines compatible with CUDA technology). Non-containerized version for CPU is also available (https://github.com/garciaml/BrainQCNet_CPU_non_containerized).

We have integrated the best-performing QC model into an open-source BIDS-app (Gorgolewski et al., 2017), to share it with the neuroimaging community in a ready-to-use format. Documentation for our BIDS-app for CPU or GPU is available here: https://github.com/garciaml/BrainQCNet. We have also shared our trained CNN baseline models for reuse: https://github.com/garciaml/BrainQCNet_CNN_GPU .

The following BIDS-apps are available on DockerHub:

- garciaml/brainqcnet-cnn: the best CNN model (which provides a control/comparison for the model based on ProtoPNet architecture);
- garciaml/bids-pytorch-cuda: a template for deep learning BIDS-app running on GPU/CUDA machines using the Pytorch framework;
- garciaml/brainqcnet: the best-performing model identified in this study, for use on GPU/CUDA machines;
- garciaml/brainqcnetcpu: the best-performing model of this study, for us on CPU machines.

Our apps and code are available under the Apache License, Version 2.0, January 2004.

We have also created and shared two demo videos explaining how to run our app on CPU and on GPU machines compatible with CUDA technology (links available on https://github.com/garciaml/BrainQCNet).

3.4. Results

3.4.1. Annotations

Manual QC inspection of 980 scans from ABIDE 1 (Di Martino et al., 2014) identified 564 high quality scans (Class 0), 36 very low quality scans (i.e. globally corrupted and score=4,4,4,4; which we used in the training and validation sets), and 380 scans with either local artefacts or with mild-moderate global corruption. Local ringing (likely reflecting motion) was the most commonly occurring local artefact, and was often combined with other artefact types.

3.4.2. Training performance

In the results and figures below, we use the following naming convention: the prefix "proto-" corresponds to the ProtoPNet algorithm, while the suffix indicates the CNN architecture: V19 for VGG19, R152 for ResNet152, or D161 for DenseNet161 (see Section 2.7).

We obtained excellent accuracy for the detection of good (Class 0) and bad (Class 1) quality slices during training. From epoch 10, accuracy for the three attention models - proto-V19, proto-R152, proto-D161, was above 99% on the Training set and above 95% on the Validation set. This means that more than 99% of the 270000 training images were accurately classified from epoch 10. Likewise, more than 95% of the 1800 validation slices were accurately classified from epoch 10. Looking at performance on the validation set, the model proto-D161 out-performed proto-V19 and proto-R152 (see **Figure 3.4, left**).

The traditional CNN comparator models also converged quickly (see **Figure 3.4, right**). The CNN models (VGG19, ResNet152, DenseNet161) trained on 15 epochs were used as comparators for the main attention models (proto-V19, proto-R152, proto-D161) in all further analyses.

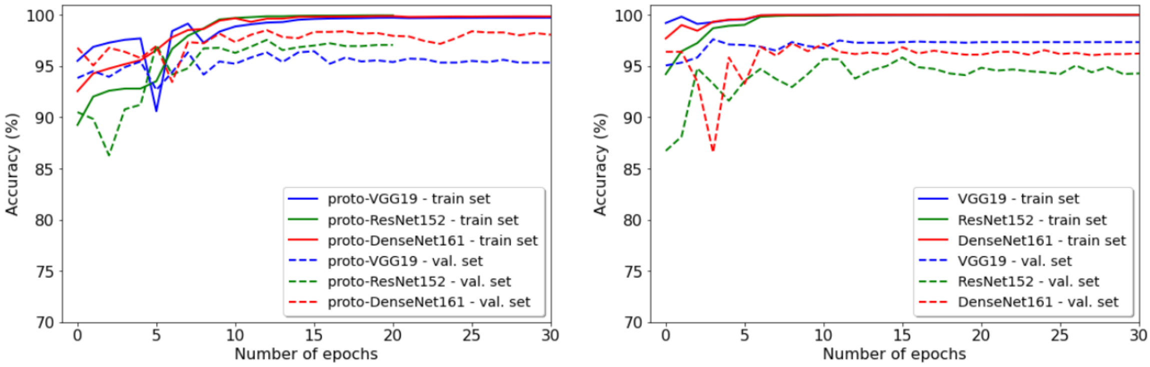


Figure 3.4: Evolution of accuracy across epochs for the Training and Validation sets; (left) training performance of the ProtoPNet models; (right) training performance of the traditional CNN models.

3.4.3. Selecting the best model using ABIDE 1

As described above (**Section 3.3.4**), predictions (Class 0/1) were performed at the level of 2D slices from a given scan. To generate a global prediction for each scan, we applied a threshold such that if >50% of slices for a given scan were predicted Label 1, the entire scan was classified as Class 1 (poor quality). Below this threshold, the entire scan was classified Class 0 (good quality). Producing a binary scan-level class prediction is useful in the QC context, because it provides a pass (Class 0) or fail (Class 1) outcome. However, there are likely to be applications for which an examination of the value of the proportion itself might be warranted, since this value gives more information about the quality of the scan. In analyses and comparisons performed below, we have operationalised this proportion as a probability - specifically, it is the frequentist probability that a given scan is corrupted by an artefact. Similarly, there will be applications where a different threshold (e.g., >0.4 = Class 1) may be preferable, depending on the particular goal of subsequent analyses. Our BIDS-app (https://github.com/garciaml/BrainQCNet) allows for the specification of a threshold for scan classification.

Table 3.2 compares the specificity and sensitivity scores for each model. While specificity is very high (>95%) for all the models (with the exception of MRIQC = 91.1%), sensitivity is relatively low. The highest sensitivity is achieved by the model proto-R152 trained on 10 epochs (47.89%) followed by the MRIQC classifier (41.58%). This may be explained by the fact that since the most severely corrupted scans were used for training, the Test set contains scans that are generally of lower and more variable severity of artefact and poor quality. Scans of moderate quality (less severe global artefact, or very localised artefact) likely yield probabilities between 0.4 and 0.5. This means that the Class predicted is 0 (good quality), the scan is of moderate rather than high quality. Supplemental **Figure A1.2** in **Appendix 1** shows the distribution of probabilities for each model and each dataset.

Table 3.2 compares the classification accuracies for global quality of the Training, Validation, and Test sets, obtained for each of the models, including MRIQC and the CNN models. These results show that the best model for the prediction of sMRI scan global quality is proto-R152 trained on 10 epochs. This model is at least as accurate as MRIQC and the CNN models. Supplemental **Figures A1.1** and **A1.2** in **Appendix 1** provide further illustrations of the distribution of probability scores across models.

Model	Training (60 scans)	Validation (12 scans)	Test (908 scans)		
			All Scans	artefact-free Class 0 (528 scans)	With artefact Class 1 (380 scans)
proto-D161 10 epochs	Acc. = 100% AUC = 1	Acc. = 100% AUC = 1	Acc. = 69.8% AUC = 0.775	Sp. = 99.4%	Sens. = 28.7%
proto-D161 20 epochs	Acc. = 100% AUC = 1	Acc. = 100% AUC = 1	Acc. = 64.7% AUC = 0.774	Sp. = 100%	Sens. = 15.5%
proto-D161 30 epochs	Acc. = 100% AUC = 1	Acc. = 100% AUC = 1	Acc. = 62% AUC = 0.758	Sp. = 100%	Sens. = 9.2%
proto-R152 10 epochs	Acc. = 100% AUC = 1	Acc. = 100% AUC = 1	Acc. = 75.4% AUC = 0.825	Sp. = 95.3%	Sens. = 47.9%
proto-R152 20 epochs	Acc. = 100% AUC = 1	Acc. = 100% AUC = 1	Acc. = 68.7% AUC = 0.811	Sp. = 99.6%	Sens. = 25.8%
proto-V19 10 epochs	Acc. = 100% AUC = 1	Acc. = 100% AUC = 1	Acc. = 67.2% AUC = 0.823	Sp. = 99.6%	Sens. = 22.1%
proto-V19	Acc. = 100%	Acc. = 100%	Acc. = 70.0%	Sp. = 99.1%	Sens. = 29.7%

20 epochs	AUC = 1	AUC = 1	AUC = 0.849		
proto-V19 30 epochs	Acc. = 100% AUC = 1	Acc. = 100% AUC = 1	Acc. = 71.8% AUC = 0.847	Sp. = 98.5%	Sens. = 34.7%
MRIQC_CLF	Acc. = 96.7% AUC = 0.767	Acc. = 100% AUC = 1	Acc. = 70.4% AUC = 0.724	Sp. = 91.1%	Sens. = 41.6%
CNN-DenseNet161 15epochs	Acc. = 100% AUC = 1	Acc. = 100% AUC = 1	Acc. = 68.1% AUC = 0.787	Sp. = 99.6%	Sens. = 24.2%
CNN-ResNet152 15 epochs	Acc. = 100% AUC = 1	Acc. = 100% AUC = 1	Acc. = 69.3% AUC = 0.792	Sp. = 99.4%	Sens. = 27.4%
CNN-VGG19 15 epochs	Acc. = 100% AUC = 1	Acc. = 100% AUC = 1	Acc. = 68.6% AUC = 0.781	Sp. = 99.6%	Sens. = 25.5%

Table 3.2: Accuracy (Acc.) and ROC AUC (AUC) scores for Training, Validation, and Test sets. Specificity (“Sp.”) and Sensitivity (“Sens.”) scores on the testing set. For each of the attention models, performance after 10, 20, and 30 training epochs (parameter optimization steps) is shown.

We identified proto-R152 (after 10 epochs) as the best model among those compared. Supplemental **Figure A1.3** in **Appendix 1** shows the distributions of probability scores for the proto-R152 model for ABIDE 1 scans with different types/levels of severity of artefact.

As described above, each algorithm selected 2000 prototype images from the augmented training set of 270000 images during each training epoch. **Figure 3.3** and Supplemental **Figures A1.5** and **A1.6** in **Appendix 1** provide examples of the prototypes. Examination of the prototypes for proto-R152 after 10 epochs suggested a set of diverse prototypes that were highly relevant for the type of artefacts detected in the ABIDE I dataset.

Further, the distribution of accuracies across categories and sites does not appear to suggest a site effect (see Supplemental **Table A1.5** in **Appendix 1**), and there was no difference in the global distribution of probabilities between the three axes (sagittal, coronal, axial).

3.4.4. Evaluation using ABCD (2141 scans)

The ABCD dataset was annotated with gold-standard manual QC judgments thanks to the workgroups performing data collection and quality control (Karcher and Barch, 2021). We tested our algorithm on 2141 of these manually QCed scans. **Figure 3.5** compares the distribution of probabilities between QC categories (pass, questionable, fail) for these 2141 ABCD scans, computed by the best-performing model (proto-R152 trained on 10 epochs). It shows that, although there is some overlap, the central tendency and distribution of probability scores differ between pass and fail categories. There is greater overlap between scores of the questionable and pass categories, which is to be expected. We confirmed this observation by performing Mann-Whitney U-tests (because the normality assumption for a T-test was not verified for any of the samples; see Supplemental **Table A1.2** in **Appendix 1**).

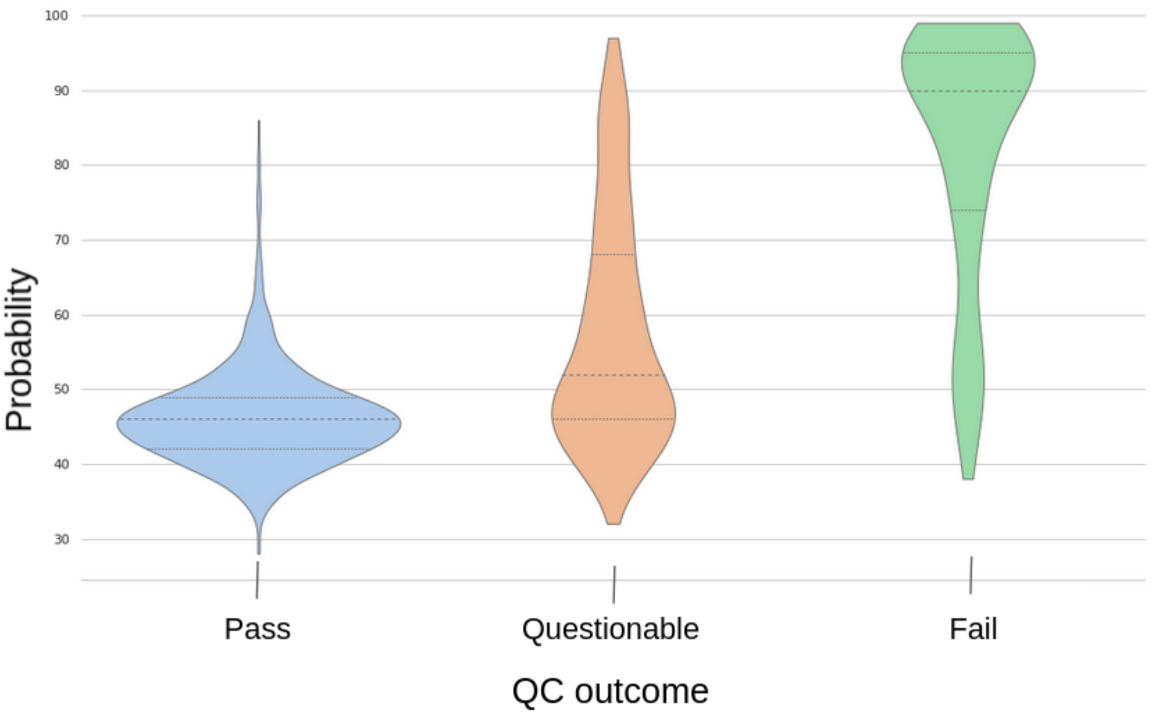


Figure 3.5: The distribution of probabilities between the true QC categories (pass, questionable, fail) for ABCD data (2141 scans), computed by proto-R152 trained on 10 epochs.

Table 3.3 shows that our algorithm showed better accuracy for the category "fail" than the comparison models. Conversely, the three CNN baseline models and MRIQC (tested on 410 of the 2141 scans, due to the time required for processing) initially performed better than proto-R152 when predicting the category "pass". Upon closer inspection, we found that 311 "pass" scans had probabilities between 0.5 and 0.6. When these scans are removed and only scans with probabilities lower than 0.5 or greater than 0.6 are retained, accuracy was 96.4% for the pass category. It is possible that our algorithm detected mild artefacts that were not considered significant by human raters. Accordingly, depending on the application, we suggest a second verification - either manual checking or a second model - for scans with "borderline" probabilities (0.5-0.6).

ABCD (2141 scans)	pass	questionable	fail

proto-R152 10 epochs	Accuracy = 82.4%	class 0: 255
		class 1: 304
			Accuracy = 91.4%

MRIQC (on 410 scans only)	Accuracy = 90.4%	class 0: 43
		class 1: 7
			Accuracy = 76.1%

DenseNet161 - 15 epochs	Accuracy = 99.9%	class 0: 484
		class 1: 75
			Accuracy = 70.7%

ResNet152 - 15 epochs	Accuracy = 99.9%	class 0: 498
		class 1: 61
			Accuracy = 67.2%

VGG19 - 15 epochs	Accuracy = 99.2%	class 0: 445
		class 1: 114
			Accuracy = 81.8%

Table 3.3: Accuracy of predictions for each of the manually determined QC categories (pass, questionable, fail) for ABCD data (2141 scans).

3.4.5. Evaluation using ABIDE 2 (799 scans) and ADHD-200 (750)

To further evaluate our tool using independent data , we ran the MRIQC classifier on 799 scans from the ABIDE 2 dataset and treated its predictions as ground truth. The MRIQC classifier predicted 588 Class 0 (pass) scans and 211 Class 1 (fail). Accuracy for our proto-R152 was 75.5%. The ROC AUC score was 0.72.

We also evaluate our model using the ADHD200 dataset, which includes manual QC (pass, fail) annotations for 750 scans. Our proto-R152 model attained an accuracy score of 79.2%

and a ROC AUC score of 0.76. Sensitivity was greater than for the CNN baseline models but specificity was lower. These results are summarised in **Table 3.4**.

	ABIDE 2 - QC prediction by MRIQC			ADHD200		
	All	588 uncorrupted scans - class 0	211 corrupted scans - class 1	All	711 uncorrupted scans - class 0	39 corrupted scans - class 1
proto-R152 10 epochs	Acc. = 75.5% AUC = 0.718	Sp. = 83.5%	Sens. = 53.1%	Acc. = 79.2% AUC = 0.76	Sp. = 80.2%	Sens. = 61.5%
DenseNet161 15 epochs	Acc. = 80.1% AUC = 0.726	Sp. = 94.6%	Sens. = 39.8%	Acc. = 90.0% AUC = 0.747	Sp. = 92.4%	Sens. = 46.2%
ResNet152 15 epochs	Acc. = 79.8% AUC = 0.742	Sp. = 93.7%	Sens. = 41.2%	Acc. = 88.4% AUC = 0.674	Sp. = 90.9%	Sens. = 43.6%
VGG19 15 epochs	Acc. = 79.5% AUC = 0.679	Sp. = 94.7%	Sens. = 37.0%	Acc. = 89.3% AUC = 0.696	Sp. = 91.6%	Sens. = 48.7%

Table 3.4: Accuracy (“Acc.”), ROC AUC (“AUC”), Specificity (“Sp.”) and Sensitivity (“Sens.”) scores for the proto-R152 and CNN comparison models for ABIDE 2 (true quality annotations obtained by the predictions of the MRIQC classifier) and ADHD200.

3.4.6. *Model interpretability*

What features of the input data does our model rely on for prediction? This question relates to the interpretability of the model, which is often challenging for deep learning models, relatively to conventional machine learning methods. Interpretability is important, not only for revealing the input features that contribute most to classification, but also for pointing to opportunities for model improvement.

First, we considered the prototypes (the 2000 images from the augmented training set of 270000 images selected during each training epoch) used by the attention models (proto-V19, proto-R152, proto-D161) and assessed whether these were well balanced in terms of the types of artefacts represented. We identified the top 5 prototypes (i.e. the 5 prototypes with the highest similarity scores with patches of 2D input slices) for each of the three axes (axial, sagittal, coronal) and observed that two prototypes (ringing and blurring) were highly prevalent among the top 5 (**Figure A1.4** in **Appendix 1**). We observed that the prototypes used by the best-performing model, proto-R152 exhibited greater diversity and less redundancy than the ones used by proto-D161 and proto-V19.

Second, to evaluate artefact localization, we examined whether the areas that the proto-R152 algorithm compares (the focus of "attention") between an input slice and associated top-prototypes (prototypes with the highest similarity scores to the input slices) appeared relevant. We selected 100 2D slices at random from the original training set of 62 Class 1 scans from ABIDE 1, and examined the top 5 prototypes and the associated attention maps. One rater - Melanie Garcia - estimated that 52.4% of the attention maps were visually meaningful, in that artefacts were visible on the 2D image. For the remaining maps, either the artefact appeared elsewhere in the slice, or no obvious artefact could be detected by eye. Two examples of such attention maps are provided in **Figures A1.5** and **A1.6** in **Appendix 1**. This outcome suggests that while there is some congruence between human-identified and automatically identified artefacts, the algorithm may detect and rely on information that is not visible to the human eye. Future work will evaluate the attention maps and performance at the local scale in greater detail.

3.4.7. BIDS Docker app

We developed a BIDS-app (Gorgolewski et al., 2017) to share our model with the neuroimaging community. It is available on the open-source platforms GitHub and DockerHub. The model and instructions are available at: https://github.com/garciaml/BrainQCNet. The GPU/CUDA version is optimal. The average time to process a 3D sMRI scan using was about 1 minute 30 seconds on a laptop with one GPU Nvidia GEFORCE GTX 1060 (6GB memory) and 50 seconds on a machine with one GPU Nvidia RTX 3090 (24GB memory). While we strongly recommend the GPU version, there is also a CPU version available. Runtime will depend on the architecture available; in our experience, the average time to process a scan was about 30 minutes on a laptop with Intel Core I7-7700HQ processor (16GB memory), while it took about 10 minutes on an Intel Core i9-10850K (64GB memory).

3.5. Discussion

In this age of “big data'', manual quality control of T1-weighted MRI scans is a time-consuming task requiring substantial experience and training. Our goal was to further advance the automatic detection of artefacts in sMRI scans by increasing the efficiency of the process. We trained an attention deep learning algorithm, ProtoPNet, paired with several different CNN architectures for the convolutional layer, to classify minimally preprocessed sMRI scans as pass/good quality and fail/poor quality. Specifically, the algorithms yielded class (0/1) predictions at the level of 2D image slices. These were converted to a probability value for each scan by computing the proportion of slices classified as fail/poor quality. Binary pass/fail global scan-level predictions were then generated by applying a threshold of 50% to the probability values. We evaluated our models’ performance by comparison to a reference tool in neuroscience (MRIQC) and to three traditional (non-attention) CNN models. Training, validation, and test sets comprised 4598, largely openly available sMRI scans from a large number of data collection sites, enabling the validation of the best-performed model using fully independent data.

Across convolutional layer architectures, the attention model ProtoPNet combined with a ResNet152 CNN architecture and trained on 10 epochs showed the best performance. On the first, non-independent, testing set (908 scans from ABIDE 1; Di Martino et al., 2014), this model performed equally as well as the reference tool, MRIQC (accuracy for high quality scans: 95.27% vs 91.1% for MRIQC; accuracy for medium and low quality scans: 47.89% vs 41.58% for MRIQC). Proto-R152 was also more sensitive than traditional CNNs, although less specific. On the second, independent, testing set (2141 scans from ABCD; Volkow et al., 2018; Karcher and Barch, 2021), the model showed excellent (91.4%) accuracy for low quality scans (i.e. high sensitivity). For high-quality scans, our model showed good prediction accuracy (82.4%), but this was lower than that of comparison models, including MRIQC (90.4%) and the CNN baseline models (from 99.2% to 99.9%). When we examined this more closely, we found that scans with a prediction falling in the mid-range of probabilities [0.5; 0.6] contained a mixture of good quality scans and moderately corrupted scans with more localised artefacts. If this "borderline" range was excluded, our model exhibited excellent accuracy for both pass and fail classes (accuracy for pass scans: 96.4%; accuracy for fail scans: 92.2%).

These data illustrate an advantage of our model - the ability to adjust global classification thresholds, or to isolate scans with probabilities falling within a specific range for further quality assessment. These parameters can be adjusted to make the classification categories more or less inclusive according to study needs. For applications where large samples are available and very high quality (artefact-free) data are required (e.g., computation of cortical thickness), the conservative 0.5 threshold could be retained. In other words, all the scans with a returned probability higher than 0.5 could be ruled out. This would have the disadvantage of removing some relatively good quality scans but the advantage of ruling out a greater proportion of lower quality scans than any other automatic method. If, on the other hand, a researcher had a smaller sample and less stringent quality requirements, a more liberal threshold of 0.6 could be set. This would mean that some scans with low severity or localised artefacts would be included in the study, but would offer the advantage that no good quality scans would be unduly eliminated. A third possibility is for researchers to retain all scans that have a global probability lower than 0.5, and to run one of our CNN models (or to manually evaluate or run MRIQC) on scans that have a global probability between 0.5 and 0.6 to separate the good from moderately corrupted scans. To facilitate these possibilities, our BIDS-app (https://github.com/garciaml/BrainQCNet) outputs a CSV file containing probability scores for each scan.

Our study demonstrates that deep learning is a promising method for increasing the speed of scan quality evaluation by reducing the computational time required, without compromising classification accuracy. Importantly, preprocessing was minimal - avoiding even the need for data reorienting, since our model was trained to process transformed (rotated, skewed, sheared) 2D image slices from the three axes (sagittal, coronal, axial). This simplifies the process compared to approaches where knowing the data orientation is necessary (Sujit et al., 2019). To generate a global prediction for a single 3D scan on a GPU machine, our model currently takes 1 minute to process one scan (50 seconds on a machine with one GPU Nvidia RTX 3090, 24GB memory; 1 minute 30 seconds on a laptop with one GPU Nvidia GEFORCE GTX 1060, 6GB memory). On a CPU machine, our model is slower but still relatively fast (10 minutes on an Intel Core i9-10850K; 64GB memory; 30 minutes on an Intel Core I7-7700HQ processor, 16GB memory). We have openly shared our code so it can be further adapted to other architectures.

In order to save resources and encourage sustainable practices, we have also shared the global scores predicted by our best model for the scans we used from ABIDE 1 and 2 (Di Martino et al., 2014; Di Martino et al., 2017), ADHD200 (Bellec et al., 2017) and ABCD (Volkow et al., 2018; Karcher and Barch, 2021). The scores are available through our GitHub repository: https://github.com/garciaml/BrainQCNet_paper_results. In addition, we have shared a version of the app containing the traditional (non-attention) CNN models. Even though our data showed that these algorithms are less sensitive (have a greater number of false negatives), they nonetheless show excellent accuracy (true negatives) for good quality (pass) scans. These characteristics may be of use for certain applications or may offer possibilities for further refinement.

Deep learning models often lack interpretability - attention models reflect an attempt to address this. As implemented here, the attention ProtoPNet model enables the localisation of regions in the input images that contribute significantly to classification. This might help to identify specific brain regions that are more vulnerable to artefacts, such as motion, or highlight a scanner quality issue that can be addressed to avoid future data loss. We have made it easy to inspect regions exhibiting local artefacts using our BIDS-app, using the parameter “n_area.” Details on how to do this can be found in the documentation.

Future work will focus on improving our algorithm by running further experiments with other CNN-bases, such as ResNet34 or DenseNet121, and examining the effects of prototype selection. In addition, we plan to increase the training set, as well as the variety of artefacts in the set of prototypes, since our approach was not exhaustive. It is likely that signals in the background are leveraged by the current attention algorithm and this behaviour should be studied more precisely. To take a wider view, it is clear that MRI scan quality is a continuous spectrum; pass/fail (good/bad) thresholds can seem arbitrary and simplistically binary. Scan quality would be better captured by a more sophisticated label, but this is very difficult to implement concretely. Moreover, reliance on human observers as the ground truth for scan quality assessment introduces its own limitations, including subjectivity and inter-observer variability. Human judgment is inherently subjective and can vary significantly between observers, leading to potential inconsistencies in the ground truth data. Even experienced clinicians or researchers may have biases or blind spots that affect their evaluations. Furthermore, human observers may not be able to detect subtle or complex patterns in data that advanced computational methods can uncover, potentially leading to underestimation of a model capabilities. The ground truth established by human observers is often limited to the knowledge and assessment criteria available at the time, which may evolve with further research. Future work should explore these limitations to understand the potential and constraints of human-validated benchmarks in MRI quality control, and should investigate whether such a quality prediction can be refined by incorporating additional information about the location/extent of artefact.

In addition, we chose to use native space for classification in this work instead of performing registration to template space on scans. This choice impacts the accessibility and interpretability of the data; certain anatomical details may be more pronounced and easier to analyse in native space, while the variability across subjects can make other information more challenging to interpret for the algorithm. Although processing scans in native space accelerates each run, it would be interesting to train a model with registered brains to determine which approach better captures inter-individual variability.

Investigating whether our approach could be applied to other MRI modalities is another important future direction. Quality Control of functional MRI is a considerable challenge that is exacerbated by the advent of Big Data. Future work will examine whether our approach can be adapted for data with a temporal dimension so that it could be applied to fMRI data in a framewise manner to enable faster and automated data quality control.

There is further scope for improvement of our algorithm and app - particularly in terms of processing speed. While the model already exhibits fast performance on GPU, we have not yet attempted to optimise the implementation by better distributing the computations or better use of infrastructure types. These possibilities will be investigated for future versions of the app, to further foster reusability.

Finally, to our knowledge, our BIDS-app is the first app that applies deep learning to neuroimaging and is built to be used on CUDA GPU machines. By sharing our code, we are providing the community with a new BIDS-app template for deep learning applications, facilitating the sharing of deep learning models in the community and helping to maximise reproducibility and collaboration.

3.6. Conclusions

In this work, we introduced a novel deep learning approach for the automatic evaluation of the quality of minimally preprocessed structural MRI scans. Our method is scalable to big datasets by taking advantage of new technologies like GPU machines with high-computing capacity. Paths to improve our model include incorporating additional CNN architectures and manually selecting the prototypes used by the model to increase the diversity of artefacts represented during training. Our approach could be further adapted to functional MRI, as well as to other types of MRI scans and organs. Our model is already freely available for use and development by the community via the app BrainQCNet (https://github.com/garciaml/BrainQCNet). Since all our code is open-source, the app can be used as a template for future applications of deep learning in neuroimaging.

3.7. Acknowledgements and Funding

This research was supported by the Hypercube Institute (Paris, France), a Trinity College Dublin Postgraduate Research Studentship (1252) awarded by the School of Medicine, and an Irish Research Council Government of Ireland Postgraduate Scholarship award (GOIPG/2021/1508).

3.8. Disclosure of competing interests

None.

4. Manuscript: "Towards 3D Deep Learning for neuropsychiatry: predicting Autism diagnosis using an interpretable Deep Learning pipeline applied to minimally processed structural MRI data"

Chapter 3 provided an example of an interpretable DL model that performed well when applied to MRI data. However, the task of detecting artefacts on a scan is solvable by a human whereas the task of detecting Autism on a scan is not (to my knowledge). Does a DL model perform as effectively on a task not achievable by humans, like predicting Autism on the basis of a structural MRI scan?

In addition, **Chapter 1** provided an overview of studies that have leveraged sMRI data in the quest to better understand the biological underpinnings of autism. Nevertheless, **Chapter 1** also outlined the precarious framework of neuroimaging, and warned about biases that may be introduced as a result of preprocessing steps that imply or rely on a "neurotypical" baseline(e.g., normalisation to template space). What novel approaches can be developed to identify ASD based on structural MRI data that move away from such questionable preprocessing steps, yet maintain interpretability? This study proposes one such approach, and outlines the methodology, results, interpretations, and limitations.

The following corresponds to a manuscript, authored by myself and my supervisor Dr. Clare Kelly, submitted to the journal Plos ONE in October 2022 that is currently undergoing revisions and due to be resubmitted by the end of September 2023. It is rewritten in English UK below.

4.1. Abstract

By capitalising on the power of multivariate analyses of large datasets, predictive modelling approaches are enabling progress toward robust and reproducible brain-based markers of neuropsychiatric conditions. While deep learning offers a particularly promising avenue to further advance progress, there are challenges related to implementation in 3D (best for MRI) and interpretability. Here, we address these challenges and describe an interpretable predictive pipeline for inferring Autism diagnosis using 3D deep learning applied to minimally processed structural MRI scans. We trained 3D deep learning models to predict Autism diagnosis using the openly available ABIDE I and II datasets (n = 1329, split into training, validation, and test sets). Importantly, we did not perform transformation to template space, to reduce bias and maximise sensitivity to structural alterations associated with Autism. Our models attained predictive accuracies equivalent to those of previous machine learning studies, while side-stepping the time- and resource-demanding requirement to first normalise data to a template, thus minimising the time required to generate predictions. Further, our interpretation step, which identified brain regions that contributed most to accurate inference, revealed regional Autism-related alterations that were highly consistent with the literature, such as in a left-lateralized network of regions supporting language processing. We have openly shared our code and models to enable further progress towards remaining challenges, such as the clinical heterogeneity of Autism, and to enable the extension of our method to other neuropsychiatric conditions.

4.2. Introduction

Autism Spectrum Disorder (Autism) is a complex and heterogeneous neurodevelopmental condition characterised by divergence from typical development on a number of behavioural dimensions, including communication, social interaction, and repetitive or restricted behaviours or areas of interest (*APA, Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*, 2013). These manifest behaviours likely reflect developmental neurological alterations over the lifespan (Baker et al., 2019; Fishman et al., 2018; Jiang et al., 2018; Lake et al., 2019; McKinnon et al., 2019; Walbrin et al., 2018), a suggestion supported by structural MRI studies (Bedford et al., 2020; Dickie et al., 2018; Ecker et al., 2015; Emerson et al., 2017; Fishman et al., 2018; Ha et al., 2015; Haar et al., 2016; Heinsfeld et al., 2018; Hong et al., 2018, 2020; Kishida et al., 2019; Lake et al., 2019; Lord et al., 1989, 1994; McKinnon et al., 2019; Pagnozzi et al., 2018; Pereira et al., 2018; Sha et al., 2019; Subbaraju et al., 2017; Yang et al., 2016; Zheng et al., 2020). Despite substantial research effort, however, no compelling brain-based biomarkers have yet emerged. Autism Spectrum Disorder is diagnosed through clinician judgment and gold standard observational tests, such as the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 1989) and the Autism Diagnostic Interview-Revised (ADI-R) (Lord et al., 1994), typically around age 43 months (Van 'T Hof et al., 2021). Given the considerable heterogeneity inherent to the diagnosis, and the wide range of long-term outcomes, the availability of robust and reproducible brain biomarkers for Autism could help refine diagnoses and potential treatment plans, thus promoting better outcomes. The availability of predictive models could also help clinicians build personalized care paths (Horien et al., 2022).

One challenge in the search for biomarkers and in the development of predictive models is the attainment of sample sizes that afford adequate statistical power. This challenge is exacerbated by clinical heterogeneity (Horien et al., 2022). Multi-site collaborative studies yielding well-powered samples, such as ABIDE I and II (Di Martino et al., 2014, 2017), have gone some way to addressing this challenge, and analyses of these samples suggest a distributed pattern of Autism-related structural alterations (Bedford et al., 2020; Ecker et al., 2015; Ha et al., 2015; Nakagawa et al., 2019; Pagnozzi et al., 2018; Pereira et al., 2018; Yang et al., 2016; Zhang et al., 2020). The application of multivariate approaches, such as machine and deep learning, offer another promising avenue for the search for brain-based biomarkers and the construction of predictive models.

These methods enable the simultaneous exploration of a very large set of features, offering much more powerful analytical capacity than univariate approaches. To date, such approaches have had moderate success, with recently reported prediction accuracies (for Autism diagnosis) in the range of 65-70% for models built using both functional and structural MRI data (Arya et al., 2020; Dekhil et al., 2020; Lu et al., 2020; Wang et al., 2020). In an effort to boost accuracy through competition, (Traut et al., 2021) held an international challenge in which competing teams predicted Autism diagnosis using a large multisite dataset comprising preprocessed anatomical and functional MRI data from > 2,000 individuals. Of the 589 models submitted, the 10 best were combined and evaluated using a subset of unseen data (from one of the sites included in the main dataset), as well as data from an additional, independent acquisition site. The blended model achieved an ROC AUC of ~0.66 using features extracted from anatomical data only. One observation from this effort was the fact that prediction accuracy increased with increasing sample size. Another was that while prediction accuracy for the subset of unseen data was similar to validation accuracy, accuracy for the novel site was poorer, illustrating the challenge of generalisation, particularly to new data collection sites.

Although recent gains in prediction accuracy are promising, machine learning studies conducted to date have two main limitations. The first is that preprocessing pipelines often have many steps, each of which can introduce biases to prediction models. In particular, preprocessing typically includes transformation to a template space, such as MNI152, which was created using anatomical scans acquired from neurotypical adults. Template normalisation may therefore negatively impact the ability to detect Autism-related alterations in brain structure, introduce biases, and lead to poorer reproducibility (Horien et al., 2022). A second limitation is that datasets used for prediction tend to be clinically heterogeneous, but this heterogeneity is not explicitly accounted for in the models, leading to inconsistent results between separate datasets (Benkarim et al., 2022). Many Autistic participants have a secondary diagnosis, which is often another psychological condition such as ADHD or anxiety, or a neurological condition such as epilepsy or Fragile X syndrome (Ecker et al., 2015; Pagnozzi et al., 2018; Sha et al., 2019). Ignoring these comorbidities may introduce biases or lead to non-specific biomarkers (Ecker et al., 2015), since in such analyses, the label “autism” is not well delimited.

In the current study, we sought to develop a prediction pipeline that could overcome these challenges. To do this, we trained 3-dimensional deep learning models to predict Autism diagnosis from minimally preprocessed structural MRI data, to avoid biases introduced by template normalisation. To address the influence of clinical heterogeneity, we built our models using a large sample of 1329 patients (521 with autism) without comorbidities, following the classical framework of train-validate-test. To test if the patterns identified by the best models were robust to comorbidity, we tested the three best models on a second dataset comprising 270 patients (155 with autism) with comorbid diagnoses.

Deep learning models can extract meaningful implicit features during the optimization process, which minimises the preprocessing required and ultimately reduces prediction time. While 2D deep learning models are increasingly popular, 3D deep learning is not widely used in medical imaging applications, in part because of the large number of parameters to optimise (greater than in 2D) and concerns related to interpretability. To address the challenge of extracting information about predictive features (i.e., interpretability), we leveraged recently developed methods to build an interpretation pipeline that identifies predictive brain areas while avoiding the requirement for template normalisation.

In this paper, we described our novel pipeline for interpretable 3D deep learning prediction of Autism diagnosis from structural MRI data. In our proof-of-concept analyses, our models achieved the same prediction accuracy as is typical for machine learning models, while avoiding the potential biases introduced by template normalisation. Our interpretation pipeline identified a set of regions that replicated well across datasets (including participants with comorbidities), and models, and which converged with previous structural imaging studies on Autism. To facilitate further development of our pipeline, we have openly shared all our code through GitHub (https://github.com/garciaml/Autism-3D-CNN-brain-sMRI).

4.3. Materials and Methods

4.3.1. Data and Quality Control

We used T1-weighted structural MRI data from the ABIDE I (980 scans) and II (857 scans) datasets (Di Martino et al., 2014, 2017) and 140 scans from ADHD200 (Bellec et al., 2017). We performed quality control using BrainQCNet (Garcia et al., 2022), retaining scans with a probability score below 60% as advised in (Garcia et al., 2022); 797 scans from ABIDE I, 704 from ABIDE II and 98 from ADHD200 remained after this step.

Our primary analysis focused on participants with a diagnosis of Autism but no reported comorbidity and comparison participants with no psychiatric diagnosis. Excluding participants with comorbidities resulted in a dataset of 1329 participants which were used for training, validating and testing the models.

All participants in the testing set (n = 65, 26 with Autism) were obtained from different (independent) data collection sites than participants in the training (n = 1074, 421 with Autism) and validation (n = 190, 74 with Autism) sets.

To examine the impact of comorbidities on prediction accuracy, we created a second evaluation set of participants who had at least other diagnoses in addition to Autism, such as ADHD, phobias, depression, and anxiety. This dataset (testing set 2) contained scans from 270 participants (155 with Autism diagnosis).

Further details on the datasets are provided in **Appendix 2, A2.1 - Detailed Data Description**.

4.3.2. Preprocessing

We employed a minimal preprocessing pipeline that did not apply transformation to template space, to avoid any impact of brain normalisation on the detecting of Autism-related alterations in brain structure. Instead, we applied FSL's Brain Extraction Tool (BET; https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET) to remove non-brain tissue, followed by a number of minor non-deforming transformations, to prepare our data to be processed by the deep learning algorithm:

- *Resolution homogenization:* the ABIDE datasets comprise data from different data collection sites, each of which has different scanners and acquisition protocols, Accordingly, the T1-weighted volumes have heterogeneous voxel spacing that could bias the analysis. We used Linear Interpolation to perform resampling, with the Resample function from the Python library TorchIO (https://torchio.readthedocs.io/_modules/torchio/transforms/preprocessing/spat ial/resample.html#Resample), built from the Insight Toolkit (https://itk.org/Doxygen/html/index.html) to resample all volumes to a fixed resolution of 1.5mm*1.5mm*1.5mm. We also reordered the data to RAS+ orientation.

- *Intensity normalisation:* We removed the noise generated by voxel value outliers in every image by truncating the intensities to the range of 0.5 to 99.5 percentiles using the RescaleIntensity function from TorchIO. We also normalised each volume by z-scoring, i.e. by subtracting the mean intensity value v_m to each voxel value v_i , and then dividing by the standard deviation v_{sd} , obtaining a new voxel value v'_i .
$$v'_i = (v_i - v_m)/v_{sd}$$
- *Cropping or Padding:* We cropped or padded each volume to obtain a uniform shape for all the volumes of 256*256*256. This shape was sufficiently large to fit the full brains and was also appropriate as an input shape to our deep learning models, in view of the filters applied all along each network (described in detail below).

4.3.3. Classification models in 3D

Comparing different types of algorithm enables the detection of overfitting and retention of the best type of algorithm for the given problem (Hastie et al., 2009). We compared two models: (1) DenseNet121 (Huang et al., 2018) and (2) Med3D-ResNet50 (Chen et al., 2019), based on well-known CNN architectures with good 2D performance (Huang et al., 2018; Chen et al., 2019). DenseNet121 is more compact and has fewer parameters than ResNet50 making it possible to train on 3D data, while Med3D-ResNet50 (Chen et al., 2019) is a version of ResNet50 that has been pre-trained on medical images, including brain sMRI scans. Logically, pre-trained models enable better convergence and performance on new data and tasks of the same context. We fine-tuned Med3d-ResNet50 to adapt it to our task by training the last convolutional layers (corresponding to the 4th convolutional block). We also appended the last classifier block, consisting of a global average pooling layer and a fully connected layer (see **Appendix 2, A2.2 - Model architectures**).

Like in (Huang et al., 2018) and in (Chen et al., 2019), we used the ReLU function as the activation function, the cross-entropy loss, and the Adam optimiser with a fixed learning rate of 0.001.

4.4. Interpreting outcomes of deep learning algorithms

4.4.1. Guided Grad-CAM

In order to interpret and evaluate the reliability and relevance of our 3D deep learning models, we used Guided Grad-CAM (Selvaraju et al., 2019), which combines guided backpropagation (Springenberg et al., 2015) and Grad-CAM (Selvaraju et al., 2019). This represents a good trade-off between the precision offered by feature maps produced by interpretability algorithms and the processing time required. Mathematically, guided Grad-CAM (Selvaraju et al., 2019) is an element-wise product of the results of the two algorithms. It returns a high resolution map of the fine-grained features that is also class-discriminative.

In the context of our study, for a given trained CNN model (either DenseNet121 or Med3DNet-ResNet50), we used guided Grad-CAM to generate one “attention map” for each participant at the inference step (i.e. the first layer of the CNN). This attention map matched the input scan resolution and voxel dimensions, and its voxel values corresponded to scores of “importance” for the prediction of Autism/non-Autism by the trained CNN model. Mathematically, for a given input participant’s scan, we computed $q_{50\%}$ - the median of the voxel values of the attention map obtained with guided Grad-CAM. We then built a binary mask by returning all the voxel values lower than $q_{50\%}$ to 0 and all voxels greater than $q_{50\%}$ to 1. We used this mask M to identify the brain regions that are the most important for the prediction of Autism across the sample and across algorithms.

4.4.2. HighRes3DNet

As noted above, a key feature of our preprocessing pipeline was our avoidance of normalisation to a group template. This creates a significant challenge for the identification of the brain areas that were most predictive of diagnosis across participants. We solved this challenge by segmenting individual scans into anatomical units and combining this information with the mask M created in the preceding step.

HighRes3DNet (Li et al., 2017) is a deep learning algorithm that segments brain MRI scans following the GIF brain parcellation (V3, http://niftyweb.cs.ucl.ac.uk/program.php?p=GIF ; (Cardoso et al., 2015)). The GIF algorithm was especially built to be robust to brain morphological differences, especially those encountered in populations with atypical brain development like Autism (Cardoso et al., 2015).

We segmented each participant's brain with the HighRes3dNet algorithm (first homogenising scans to voxel size 1mm*1mm*1mm using linear Interpolation). The resulting segmented images were resampled to 256*256*256 images of voxel size 1.5mm*1.5mm*1.5mm to match the resolution of the attention maps obtained from the guided Grad-CAM algorithm, while retaining the segmented voxel values.

Specifically, we know that the information on the transformations applied to the segmented image is contained into the affine matrix of the resulting transformed segmented image.

Mathematically, we note $X = [x, y, z, 1]$, the column vector of the coordinates x, y, z of a voxel in a segmented image obtained with HighRes3DNet (voxel size: 1mm*1mm*1mm), $Y = [x', y', z', 1]$ the column vector of the coordinates x', y', z' of a voxel in the corresponding transformed segmented image (size: 256*256*256; voxel size: 1.5mm*1.5mm*1.5mm), and $A \in |R^4$ its affine matrix. We note B, the inverse matrix of A, such that $BA = A^{-1}A = I$, where I is the identity matrix in R^4.

Thus, we have the relationship:

$$AX = Y$$

$\Leftrightarrow X = BY, \forall (x', y', z') \in [1, 256]^3$.

Thus, if we take x', y', z' the coordinates of a voxel in the mask M obtained from guided Grad-CAM, we can obtain the corresponding x, y, z voxel coordinates in the segmented image, and thus get the voxel value and the name of the area at (x, y, z).

Applying this procedure for every scan, we obtained a table containing, for every area of the HighRes3DNet atlas, a relative frequency corresponding to the number of voxels in the area with value = 1, divided by the total number of voxels in this area in the segmented image. This relative frequency corresponds to the proportion of the area that is considered important for the prediction by a CNN model, for that participant. These proportions were then used to compare different brain areas and to draw up a ranking of brain areas for each model, dataset (training, validation, testing sets), and type of prediction (True Positives, True Negatives, False Positives, False Negatives), to improve interpretability for our CNN models.

4.4.3. Machine and Code availability

We trained our model on a GPU Nvidia RTX 3090 (24 GB memory) with a batch size of 2.

We openly shared the code of this project on GitHub, in the repository: https://github.com/garciaml/Autism-3D-CNN-brain-sMRI. The models are also shared so that they can be reused as pre-trained models for similar applications.

4.5. Results

4.5.1. Training Performance

For all the probability scores of all the models, we chose a threshold of 0.5 for the class “Autism diagnosis” to define the prediction and compute the accuracy and ROC AUC scores.

We trained each model up to 100 epochs and computed model accuracy using the validation set (190 scans) every two epochs. Details on the validation set accuracy during training for the two models DenseNet161 and Med3d-ResNet50 are provided in **Appendix 2**, **Figure A2.1** in **A2.3 - Performance of the models**.

For ResNet50, the best validation set accuracy was 62.6%, achieved at 42 epochs. For DenseNet121, 66.3% accuracy was achieved at 32 epochs and 67.4% was achieved at 70 epochs. Next, we compared the performance of these three best models (one ResNet50 model and two DenseNet121 models) for the prediction of diagnosis in the training, validation, and testing sets.

4.5.2. Prediction Performance: Autism diagnosis

For the prediction of Autism diagnosis, the three best models behaved differently, as shown by the Receiver Operating Characteristic curves in **Figure 4.1**. Med3d-ResNet50-42ep overfitted the data - the accuracy and ROC AUC scores were very high on the training set (94.2% and 99.9% respectively) but much lower on the validation (acc = 62.6% and AUC = 62.1%) and testing sets (acc = 53.8% and AUC=57.3%). DenseNet121-32ep appeared to be more stable in terms of its overall performance on the training (acc = 65.5% and AUC = 69.1%), validation (acc =66.3% and AUC = 68.8%) and testing (acc =55.4% and AUC = 60.7%) sets. DenseNet121-70ep had better performance on the training (acc = 69.7% and AUC = 77.1%) and validation (acc = 67.4% and AUC = 68.1%) sets than DenseNet121-32ep, but poorer performance on the testing set (acc = 40% and AUC = 38.1%).

Table 4.1 displays the sensitivity and specificity of each model for each dataset. DenseNet121-32ep exhibited high specificity on the training and validation sets, but low sensitivity. Paradoxically, it had high sensitivity but low specificity on the testing set. DenseNet121-70ep behaved similarly on the testing set while on the training and validation sets, sensitivity and specificity were balanced and fairly high. Finally, for Med3d-ResNet50-42ep, sensitivity and specificity were very high on the training set, unbalanced on the validation set with low sensitivity and very high specificity, and balanced on the testing set, but with moderate values.

Sensitivity on the second testing set, which included participants with comorbidities, was low for all models. This demonstrates that when the training and testing sets include only participants without known comorbidities, predicting Autism diagnosis for participants with comorbidities is particularly challenging. Here, we found that this produces a large increase in False Negatives in particular. One potential explanation is that neuroimaging markers of Autism are less salient when individuals have another diagnosis involving similar or other neuroimaging markers. Another explanation is that more data is needed to adequately train DL algorithms on the whole spectrum of Autism in the context of comorbidities.

Further details and comments on the performance of the models are given in **Appendix 2, A2.3 - Performance of the models**, and a comparison of the predicted scores with the scores of diagnosis are given in **Appendix 2, A2.4 - Analysis of ADI-R and ADOS scores, age, gender and full IQ**.

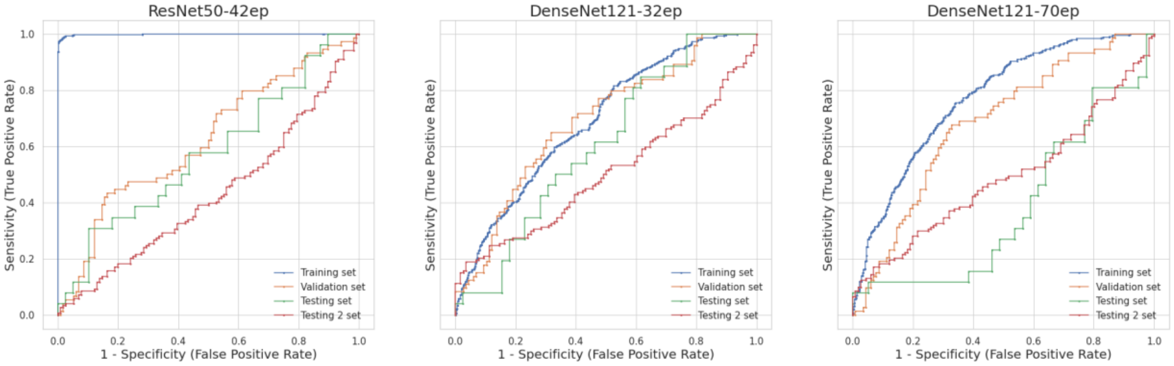


Figure 4.1: Receiver Operating Characteristic curves for all the three models and all the four datasets

	Med3dNet - Resnet50; trained on 42 epochs	DenseNet121; trained on 32 epochs	DenseNet121; trained on 70 epochs
Sensitivity	Train: 85,3%	Train: 32,8%	Train: 68,2%
	Validation: 17,6%	Validation: 36,5%	Validation: 66,2%
	Test: 50%	Test: 84,6%	Test: 69,2%
	Test 2: 8,4%	Test 2: 7,6%	Test 2: 31%
Specificity	Train: 100%	Train: 86,7%	Train: 70,8%
	Validation: 91,4%	Validation: 85,3%	Validation: 68,1%
	Test: 56,4%	Test: 35,9%	Test: 20,5%
	Test 2: 87,8%	Test 2: 100%	Test 2: 73%

Table 4.1: Sensitivity and Specificity of each model on each dataset (training, validation, testing sets with no comorbidity and testing set 2, which included patients with comorbidities).

4.5.3. Interpretability: True Positive discriminative ROIs

We segmented each participant's scan using HighRes3DNet (GIF parcellation), to extract a measure of "prediction importance" (the output of the guided Grad-CAM algorithm) for each of the three best models. We identified the regions that best contributed to True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), across the whole dataset (training + validation + testing 1 & 2 sets).

For every pair of model and dataset, we defined the "most predictive" regions as those with relative frequency values (see **Section 4.4.2,** above) greater than the 90% percentile.

This yielded 16 regions for each model and dataset pair. To compare the most predictive regions across models and datasets (training, validation, testing Set 1 - no comorbidities, testing set 2 - with comorbidities), we summed the presence (1) or absence (0) of the most predictive regions over all the datasets, separately for *True Positives* and *True Negatives*. Across all three models, 79 areas were found to be most predictive for *True Positives*, including 26 areas spanning both left and right hemisphere, 23 areas in the left hemisphere only, 3 areas in the right hemisphere only, and the Corpus Callosum. Retaining only areas that replicated across all four datasets (training, validation, and test 1/2), we found that areas in the left hemisphere were more replicable than those in the right, and that the majority of areas were in the prefrontal cortex. In **Appendix 2**, the section **A2.5 - Most important regions for the prediction of True Positives** provides **Table A2.6** that summarises the most replicable regions across models and datasets that are important to predict *True Positives*, and a detailed analysis of these most replicable regions.

Overall, 17 regions were found to best predict *True Positives* across models and replicate across datasets (training, validation, and testing 1/2). These regions are shown in **Figure 4.2** and include regions in the left frontal lobe (medial frontal cortex, inferior and middle frontal gyrus, lateral and medial precentral gyrus, anterior and subcallosal cingulate gyrus, and posterior orbital gyrus), left temporal lobe (temporal pole, planum temporale, parahippocampal gyrus), parietal lobe (parietal operculum, supramarginal gyrus, and superior parietal lobe), as well as left parietal white matter and the right ventral thalamus.,

Looking at these data another way, and taking the regions that were most predictive across datasets and which replicated across the three models, we again obtained left hemisphere regions that are located in the frontal lobe - middle and inferior frontal gyrus (pars triangularis) and medial precentral gyrus - and in the limbic system and its associated structures - anterior cingulate gyrus, subgenual cingulate gyrus, parahippocampal gyrus (**Figure 4.2(b)**).

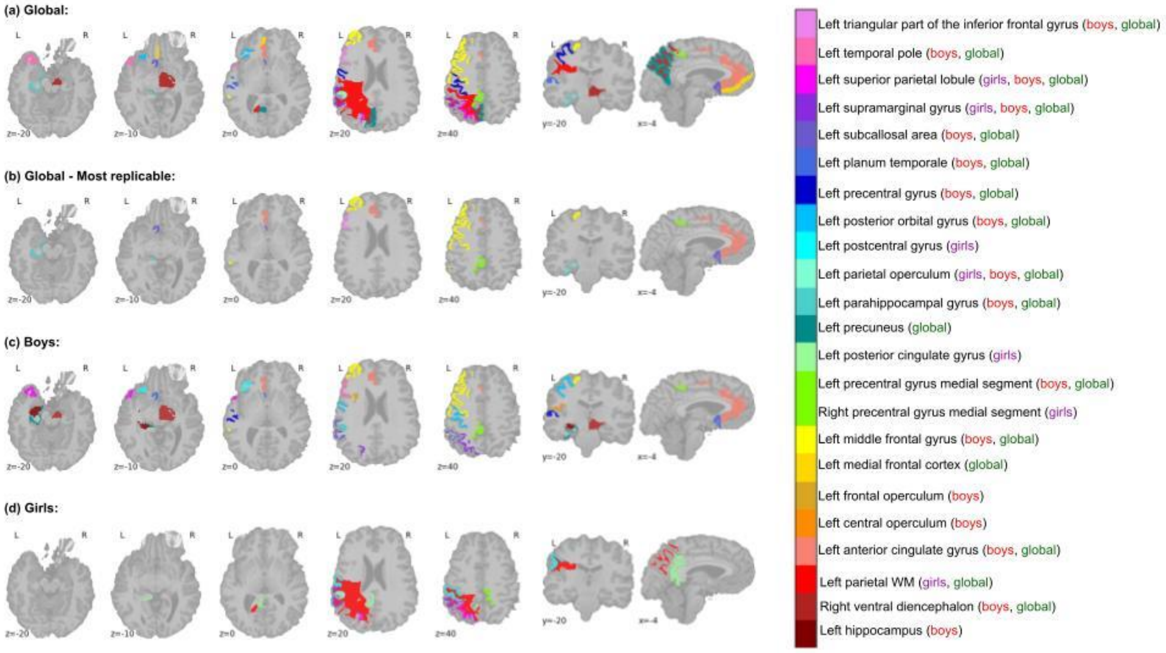

(a) Global:
(b) Global - Most replicable:
(c) Boys:
(d) Girls:
L
R
z=-20
z=-10
z=0
z=20
z=40
y=-20
x=-4
Left triangular part of the inferior frontal gyrus (boys, global)
Left temporal pole (boys, global)
Left superior parietal lobule (girls, boys, global)
Left supramarginal gyrus (girls, boys, global)
Left subcallosal area (boys, global)
Left planum temporale (boys, global)
Left precentral gyrus (boys, global)
Left posterior orbital gyrus (boys, global)
Left postcentral gyrus (girls)
Left parietal operculum (girls, boys, global)
Left parahippocampal gyrus (boys, global)
Left precuneus (global)
Left posterior cingulate gyrus (girls)
Left precentral gyrus medial segment (boys, global)
Right precentral gyrus medial segment (girls)
Left middle frontal gyrus (boys, global)
Left medial frontal cortex (global)
Left frontal operculum (boys)
Left central operculum (boys)
Left anterior cingulate gyrus (boys, global)
Left parietal WM (girls, global)
Right ventral diencephalon (boys, global)
Left hippocampus (boys)

Figure 4.2: (a) Regions most predictive of Autism diagnosis; (b) Most predictive regions that replicate across datasets, (c) Most predictive regions for boys; (d) Most predictive regions for girls.

- *Effect of gender*

Regions important for predicting *True Positives* for boys were different from those for girls. Regions common to both genders were located in the left parietal lobe: parietal operculum, supramarginal gyrus, and superior parietal lobule (**Figure 4.2(c), (d)**). Globally, regions found important to predict *True Positives* for boys were more replicable across the datasets (training, validation, testing 1/2) than for girls. For boys, several left prefrontal regions were replicably predictive of Autism diagnosis: left anterior cingulate gyrus, middle frontal gyrus, inferior frontal gyrus (pars triangularis; ResNet50-42ep only), medial precentral gyrus (DenseNet121-32ep) and precentral and parahippocampal gyrus (DenseNet121-70ep).

In Appendix 2, section A2.8 - True Positives by Gender shows these results in Tables A2.10 and A2.11.

- *Relationship with age*

Autism has been associated with disrupted brain development across the lifespan. To assess whether there were any developmental trends in the most predictive areas, we created four age categories (5-10yrs, 10-15yrs, 15-20yrs, >20yrs) and identified the most predictive (True Positives) regions for each category, separately for boys and girls. In **Appendix 2**, section **A2.9 - True Positives by Gender and Age** shows these results.

Our results showed that the most discriminative regions varied with the age. In particular, left precentral gyrus, central operculum, and posterior orbital gyrus replicably predicted *True Positives* in boys aged 5-10yrs, while left inferior frontal gyrus (pars triangularis), subcallosal/subgenual cingulate cortex, and supramarginal gyrus, were most predictive for boys aged 10-15 years old.

In addition, we found that the replicability of each region decreased as age increased. Indeed, we found that the left and inferior frontal (pars triangularis) gyrus, posterior orbital gyrus and putamen were most predictive for 15-20 years old, but only for participants without comorbidities. Left temporal areas - parahippocampal gyrus, superior temporal gyrus and temporal pole - were most predictive for males aged 20-64yrs without comorbidity.

Examining global prediction performance for these different age groups reveals other interesting trends, such as a decrease in the number of *False Negatives* and *True Negatives* with increasing age, for both boys and girls. This suggests that our prediction of Autism diagnosis tended to be more sensitive but less specific as age increased.

- *True Negatives*

We adopted the same approach described above to identify regions most predictive of *True Negatives* (i.e., absence of an Autism diagnosis). The results (see **A2.6 - Most important regions for the prediction of True Negatives** in **Appendix 2**) showed that the most replicable regions for predicting *True Negatives* were in the left hemisphere and included the frontal operculum, the precuneus, the planum polare, the inferior occipital gyrus, the occipital fusiform gyrus, the superior occipital gyrus and the thalamus proper. It also included the cerebellar vermal lobules VI and VII.

Another result is that the regions left precuneus, parietal operculum, and superior parietal lobe, and right thalamus were important (at various degrees of replicability and for different models) for the prediction of both *True Negatives* and *True Positives*. The 23 other regions important for the prediction of True Negatives are different from those that were important to the prediction of True Positives.

- *Bad predictions - False Positives and False Negatives*

We adopted the same approach described above to identify regions most predictive of *False Positives* (i.e., incorrectly predicted Autism diagnosis) and *False Negatives* (i.e., incorrectly failed to predict Autism diagnosis). The results (see **A2.7- Most replicable regions for *False Positives* and *False Negatives*** in **Appendix 2**) showed that no highly replicable regions (replicable over all datasets) were found for *False Positives.* However, regions with a high level of replicability for *False Positives* for DenseNet121-70 overlapped with replicable regions for the prediction of *True Positives* for the two other models and included the middle frontal gyrus, precentral gyrus medial segment, and triangular part of inferior frontal gyrus. This illustrates differences in the calibration of each algorithm and demonstrates the importance of comparing different models. For *False Negatives*, the most replicable regions were again found in the left hemisphere and included the left frontal operculum, left precuneus, left superior temporal gyrus, left planum polare, left inferior occipital gyrus and left occipital fusiform gyrus.

4.5.4. Does image background contribute to model predictions?

As a final test, we examined whether image background (i.e., information outside the brain) contributed to predictions. For Med3d-ResNet50-42ep the relative frequency of the Background (RF) is the smallest (RF=0.97%) and the second smallest for DenseNet121-70ep (RF=0.28%), meaning that this area is not considered predictive for the models. For

DenseNet121-32ep, it is among the last 4% informative areas of the model (RF=0.74%). These results confirm that the models use information from inside rather than outside the brain to make a prediction, supporting their validity.

4.5.5. Multi-site effect

We observed an inhomogeneous consistency of the distributions of probability scores between the different sites (see **Appendix 2**, **A2.10 - Multi-site effect**). We displayed the accuracy scores for every site in the whole dataset (training+validation+testing sets) in **Appendix 2**, **Table A2.20**, and it also confirmed the multi-site effect.

4.6. Discussion

This study outlines and demonstrates a novel approach for inferring Autism diagnosis from structural brain imaging data using 3D deep learning algorithms. To maximise the interpretability of the model outputs, we also used a second type of algorithm - guided Grad-CAM (Selvaraju et al., 2019) - to extract patterns important for the predictions. This step revealed a set of regions predominantly located in the left hemisphere, including lateral and medial prefrontal cortex, anterior cingulate, the superior temporal gyrus, lateral parietal regions including supramarginal gyrus, parahippocampal gyrus. The only right hemisphere region highlighted in our analyses was the right thalamus. The regions highlighted by this interpretability analysis, the brain structural features of which were most important for accurate inference of Autism diagnosis (i.e., *True Positives*), are highly consistent with the literature. Our predictive modelling framework has considerable potential to be extended to further datasets to identify and refine sensitive and specific brain biomarkers of Autism using MRI data.

4.6.1. 3D deep learning applied to minimally processed data

To our knowledge, this is the first time that 3D-DL CNNs have been used to predict Autism diagnosis from 3D structural MRI scans. Our findings show that these algorithms are capable of inferring Autism diagnosis on the basis of structural MRIs with at least the same level of accuracy as traditional machine learning algorithms, while requiring a smaller number of training epochs. The average accuracy score (64.1%) and ROC AUC score (0.67) obtained for participants without comorbidities is consistent with previous machine learning models trained on sMRI data (e.g., (Traut et al., 2021)). The comparable accuracy we achieved should be viewed in the context of the speed of inference of deep learning models over machine learning approaches. While machine learning algorithms require inputs derived following extensive preprocessing of structural MRI data, including normalisation to template space, our deep learning models used minimally preprocessed data. In particular, we avoided transformation to template space, a near-universal requirement of neuroimaging analyses that may negatively impact the ability to detect structural alterations associated with the diagnosis of interest. Although our pipeline included some minimal preprocessing steps to address the fact that a diversity of scanners and acquisition protocols was used across data collection sites, resulting in heterogeneous voxel spacing and signal intensities. Resolution homogenization and intensity normalisation were applied to address these variations, and it is possible that these steps could bias the algorithm. Further, despite these steps, a clear effect of the data collection site was observable. Future studies will incorporate specific preprocessing steps like the ComBat algorithm (Radua et al., 2020) to integrate scan parameters during training and minimise site effects.

4.6.2. Interpretability

The outputs of deep learning models are not straightforwardly understandable, giving rise to the challenge of poor interpretability. This challenge arises because mathematically, deep learning models are composed of multiple functions. Each of these functions is nonlinear and is itself the sum of multiple functions. Further, models such as the 3D CNNs used in the current study have a large number of parameters that must be optimised. One of the goals of our study was to address this drawback by devising a pipeline that would allow for the extraction of predictive brain regions, providing interpretability. Guided Grad-CAM (Selvaraju et al., 2019) was chosen for this purpose, due to its reasonable computation time and its ability to return fine-grained class-specific segmentations of important (predictive) voxels in the input images.

A challenge for our novel interpretability process was to identify brain areas that were predictive of Autism diagnosis across participants while avoiding the requirement for template normalisation. To address this issue, we used a segmentation algorithm to partition individual volumes into established anatomical regions. We used HighRes3DNet (Li et al., 2017) for this task because it was built to be pathology-agnostic, robust to brain morphology differences, and has reduced computation time compared to other algorithms (e.g., the GIF algorithm (Cardoso et al., 2015)). We performed a detailed analysis of the regions that were most relevant for inferring an Autism diagnosis, by examining true and false positives and negatives separately for each dataset and algorithm. We also identified regions that were reproducibly identified across algorithms and datasets. This detailed analysis is important because each model has biases, likely resulting in a differential weighting of anatomical features and brain areas. This analysis showed that regions of left prefrontal cortex (inferior and middle frontal gyrus, medial prefrontal gyrus, anterior and subgenual cingulate cortex), along with the parahippocampal gyrus were brain regions whose morphological features contributed most to the accurate inference of Autism across models and datasets without and with comorbidities. The areas highlighted are consistent with previous studies reporting Autism-related disruptions to cortical development (Chien et al., 2021; Nordahl et al., 2007; Pagnozzi et al., 2018; Zielinski et al., 2014) and gyrification processes (Pagnozzi et al., 2018; Kohli et al., 2019) in these regions. Further, also consistent with the literature, we found that the most predictive regions varied according to both gender and age, as well as the presence of comorbidities (Ecker et al., 2015; Pagnozzi et al., 2018; Retico et al., 2016). This is consistent with observations that Autism is a complex condition, with patterns of neurological divergence that vary with age (Chien et al., 2021; Ecker et al., 2015; Pagnozzi et al., 2018; Zielinski et al., 2014) and sex (Ecker et al., 2015; Pagnozzi et al., 2018; Retico et al., 2016). Interestingly, the left parietal white matter was found to be important for accurately predicting ASD in boys. This region contains tracts that may connect the parietal lobe with visual regions in the occipital lobe, among other parts of the brain. Girault et al. (2022) identified a significant association between sibling brain connectivity and proband behaviour – in relation to the SCQ questionnaire (Rutter et al., 2003) – specifically for functional connections between the visual and posterior frontoparietal networks. Therefore, examining sMRI data may provide valuable insights and complement findings from fMRI analyses. This underscores the value of conducting multi-modal experiments in future research.

Reproducibly predictive regions in the limbic system (left parahippocampal gyrus, anterior cingulate gyrus, and subcallosal area), dorsal medial frontal cortex, and precentral gyrus fit well with previous work on the role of atypical socio-emotional and motor circuitry in Autism (Ameis & Catani, 2015; Carper et al., 2015; Mundy, 2003; Nebel et al., 2014; Patriquin et al., 2016). Many of the left-hemisphere regions identified as contributing to accurate inference of Autism diagnosis fall within the canonical left-lateralized language network, including inferior prefrontal and inferior parietal regions, and the planum temporale in superior temporal gyrus (Kelly et al., 2010; Malik-Moraleda et al., 2022; McAvoy et al., 2016). Divergent structure and function in the language network is a robust and reproducible finding in Autism (Floris et al., 2016; Lindell & Hudry, 2013; Sharda et al., 2016; van Rooij et al., 2018). Since early language processing appears to be an important predictor of long-term outcomes in Autism (Lombardo et al., 2015; Szatmari et al., 2015; Tager-Flusberg & Kasari, 2013) identification of early-emerging structural alterations in the underlying language network has the potential to yield a powerful marker of Autism or Autism subtypes, which could, in turn, direct individualised interventions and improve prognosis.

An important caveat is that while our novel interpretation step identified which regions of the brain had morphological features relevant to the model-based inference of Autism, it did not provide information on what these morphological features were. For example, features such as cortical thickness, the location of the grey-white boundary, surface area, and gyral/sulcal morphometry could all play a role in prediction of Autism (Andrews et al., 2017; Hong et al., 2018, 2020; Zielinski et al., 2014); and different morphological features may be relevant in different brain areas. While the precise nature of the Autism-related morphological features are not discernable from our analyses, our predictive modelling analyses can be followed up with in-depth, targeted, and hypothesis-driven examinations of the areas highlighted in independent samples to uncover the nature of these features.

4.6.3. Limitations and Ethics

Our pipeline for prediction of neuropsychiatric diagnosis (Autism) on the basis of minimally preprocessed T1 MRI scans advances progress toward interpretable 3D deep learning applications in biological psychiatry and toward the identification of reproducible brain biomarkers that will help refine diagnoses and potential intervention plans across conditions. Our study had several limitations, however, which may be addressed in further refinements of our pipeline.

First, we trained our models on 100 epochs, which is an acceptable number relative to other studies using 3D MRI scans (Lam et al., 2020), but which may have limited the convergence and optimization of the algorithms. Future work may train on a larger number of epochs or may employ earlystopping (Yao et al., 2007) to optimise training. Using the entire structural MRI scans (to explore prediction across the whole brain) may also have posed a challenge for convergence towards the "True" solution. Further, although we used a large dataset (1074 participants to train the models, 525 to validate and test the models), the amount of data available is still rather limited when we consider the clinical heterogeneity of Autism. This idea is supported by the poor prediction performance we observed for test set 2, which included participants with comorbid diagnoses (average accuracy = 46.3%, ROC AUC = 0.47 and average sensitivity = 15.7%). There are still questions in the literature about whether predicting a binary label, "Autism vs non-Autism" is a useful or appropriate endeavour, since Autism is a wide spectrum of behaviours and abilities which may encompass as many as four subtypes (Hong et al., 2020), and there is also considerable overlap of symptoms and neuromarkers across psychological conditions (Ecker et al., 2015). Future analyses will need to leverage even larger datasets to better address the clinical heterogeneity of Autism and to explore the prediction of categories beyond Autism and non-Autism.

Another limitation is related to the segmentation algorithm we used in the interpretation step. We used HighRes3DNet (Li et al., 2017) to obtain rapid segmentation for each brain using the GIF algorithm (Cardoso et al., 2015), which was built to be robust on atypically developing brains. The segmentation produced is rather coarse, however - the algorithm outputs relatively large parcels, encompassing anatomically heterogeneous regions such as the anterior cingulate gyrus or superior parietal lobule. Further, as noted above, while our interpretation process localised regions that were important for prediction of Autism, it did not provide information on what the predictive morphological features of those regions were.

A novel aspect of this study is the decision to perform classification in native space rather than template space. This approach has the potential to make certain types of information more readily accessible, particularly those that are sensitive to individual anatomical variability. However, it can also make other information more challenging to learn from, as it may introduce variability that is irrelevant to the classification task. This trade-off warrants a thorough discussion. For instance, information that is spatially normalised in template space might become obscured or distorted in native space, while unique anatomical features may become more pronounced. Additional studies are necessary to study such effects.

In considering the ethical landscape of our research, it is paramount to reflect on the implications of our findings, especially in scenarios where models exhibit high sensitivity but low specificity. Such outcomes, while adept at identifying true positives, also raise the likelihood of false positives. In the context of ASD, the consequences of false positives—such as unwarranted difference, unnecessary intervention or treatments, and the social stigma associated with misdiagnosis—can be profound and, thus, must be weighed carefully.

Conversely, the costs of false negatives, where true conditions go undetected, can be equally grave, potentially resulting in delayed or missed opportunities for early intervention and support. The ethical calculus of these outcomes is complex and varies across ASD people and contexts. Therefore, developing and producing an identification tool of ASD must be accompanied by a robust ethical framework that carefully considers these trade-offs and strives to minimise harm.

4.6.4. Future Directions

There is considerable scope to extend our interpretable deep learning pipeline to the prediction of other neurological or neuropsychiatric conditions or to other MRI modalities. (Traut et al., 2021) reported that prediction of Autism was considerably improved (from AUC=0.66 using only anatomical MRI to AUC=0.79 using both anatomical and functional data) for a blended model that incorporated both functional and structural MRI data. Future work will examine whether functional MRI data can also improve our models. Other efforts to improve our model will include training the models on more epochs, exploring other architectures, integrating scanning parameters and other confounds such as gender and age, and using different and extended class labelling.

In addition, to evaluate the costs and benefits of native space analysis more systematically, future work could involve a comparative study where the same classification tasks are performed in both native and template spaces. Metrics such as classification accuracy, generalisability across diverse datasets, and the interpretability of learned features could provide a quantitative basis for assessing the relative merits of each approach. Additionally, the impact of native space analysis on computational efficiency and the requirement for more complex data augmentation strategies should be considered. Ultimately, such investigations could lead to a set of guidelines or criteria for determining when native space analysis is most advantageous for neuroimaging studies.

As for the paradigm chosen, our study has primarily focused on direct classification approaches, where the goal is to categorise MRI data into distinct classes, such as the presence or absence of a neuropsychiatric condition (e.g. detecting ASD). However, an alternative and complementary perspective is offered by normative modelling. This method involves constructing models of typical brain development or structure and then assessing individual deviations from this norm, which may be indicative of pathology. Normative models can be particularly informative for understanding complex neuropsychiatric conditions that exhibit a high degree of inter-individual variability like for ASD. Using unsupervised DL algorithms such as autoencoders or Generative Adversarial

Networks as dimensionality reduction techniques could further help to build such normative modelling.

We have shared all our code (https://github.com/garciaml/Autism-3D-CNN-brain-sMRI) to enable other researchers to apply, reuse, and further develop our models and approach.

4.7. Conclusion

In this paper, we described a novel methodology to build a predictive model to infer Autism diagnosis using 3D deep learning applied to structural MRI scans, coupled with an interpretation step in the form of a descriptive method that identified the brain regions that were most important for accurate inference. Importantly, we applied our models to minimally preprocessed data - completely avoiding the template normalisation step, which may obscure diagnosis-related alterations in brain structure. We found that the predictive performance of our models was equivalent to that of machine learning models reported in the literature, while requiring less time to generate predictions (due to minimal preprocessing). There is considerable scope to refine our method or to incorporate other modalities (e.g., fMRI) to further boost predictive performance.

Our method for interpreting the output of deep learning models revealed highly predictive brain regions that were consistent with the literature, demonstrating that 3D deep learning models produce biologically plausible results without a priori knowledge or the requirement for pre-computation of morphological derivatives (e.g., volumes, cortical thickness, surface area). Although challenges related to the clinical heterogeneity of Autism remain to be addressed, we have openly shared our code and models for others to build on and extend, and to further progress the field towards the identification of robust and reproducible brain biomarkers for neuropsychiatric conditions.

4.8. Acknowledgements

We have much gratitude and appreciation for the Irish Research Council and the Hypercube Institute (Paris) for their funding.

We also thank Pr. Louise Gallagher as well as Dr. Robert Whelan for their feedback and guidance.

4.9. Disclosure of competing interests

None.

5. Project 3: Transformer and multi-tasking to detect ASD using rs-fMRI

Chapter 4 highlighted the challenges of developing a DL algorithm that is both robust and interpretable for detecting ASD using structural MRI data.

Previous work summarised in **Chapter 1** highlighted an interest in utilising *functional* MRI data for ASD detection. **Chapter 1** also emphasised the potential confounding effects of factors like age and gender on the diagnosis of ASD.

This prompts the question: What novel approaches can be developed to identify ASD based on resting-state functional MRI data, which also integrate gender and age into the optimization of the model? This study proposes one such approach, and describes the methodology, results, interpretations, and limitations.

5.1. Abstract

This study pioneered applications of Transformer neural networks, a leading DL architecture, for decoding predictive patterns in resting-state fMRI data related to ASD. A methodological framework encompassing data preprocessing, cross-validation strategies, multi-task learning, and interpretability analyses was developed.

While accuracy gains over single-task models were modest, multi-task approaches significantly altered model behaviours in nuanced ways, demonstrating the value in joint training. This establishes a strong basis for refinements like loss weighting and learning rate optimization. With hyperparameter tuning and expanded datasets, both approaches show promise for distilling insights about spatiotemporal brain dynamics.

For model interpretation, reasonable initial techniques were implemented, including representation visualisation and LIME relevance mapping. Analysing intermediate layers, aggregating local explanations, and integrating alternative interpretable modules offer

exciting future directions. Enhanced interpretation can uncover how predictive fMRI patterns are encoded.

5.2. Introduction

The advent of Functional MRI (fMRI) sparked a revolution in neuroimaging in the 1990s, enabling visualisation of human brain dynamics with MR for the first time. This technology opened new avenues for psychology and neurology research, but also posed new analysis challenges, given its lower spatial resolution, yet high dimensionality. Key questions arose around modelling relationships between brain regions across the newly captured time dimension. While early studies focused on characterising the activity evoked by task performance, more recent "resting state" fMRI (rs-fMRI) studies focus on characterising the functional architecture revealed by correlated intrinsic (task-independent) brain activity (Canario et al., 2021). Low participant demands have dramatically increased researchers' ability to gather fMRI data from participant groups that were often inaccessible to conventional task-based fMRI, including Autistic children. Together with the ease of pooling data across data collection sites, this advantage of rs-fMRI has enabled the creation of large-scale open science data repositories that include both structural and rs-fMRI data, such as ABIDE (Di Martino et al., 2014, 2017), ABCD (Volkow et al., 2018), UK Biobank (Sudlow et al., 2015), ENIGMA (Thompson et al., 2020) and Healthy Brain Network (Alexander et al., 2017).

In capturing activity dynamics, fMRI data may offer an important alternative avenue toward brain markers for ASD. It is interesting to note that some ML models built on fMRI have shown better ASD classification performance than those built using structural MRI data (Traut et al., 2021).

There are a number of caveats, however. Traditional rs-fMRI analysis entails extensive preprocessing, compressing information into derived neuroimaging features like regional homogeneity (Jiang et al., 2018) or functional connectivity (Laird et al., 2011; Smith et al., 2009). While neuroscientifically meaningful, the signal processing involved may discard predictive signals.

Further, unintended replicability issues and non-reproducible fMRI findings present major concerns (He et al., 2019; Churchill et al. 2012; Dadi et al, 2019; Traut et al., 2021) that are compounded by small sample sizes; large-scale well phenotyped data gathered from clinical groups, such as Autistic individuals, remains scarce. Standard fMRI preprocessing may also introduce biases that exacerbate this issue (Traut et al., 2021; Dadi et al, 2019). For example, (Churchill et al.; 2012) found that optimising pipelines individually revealed activation patterns that were absent under fixed preprocessing, demonstrating a significant impact of pipeline choices.

Can deep learning help overcome these challenges? Deep learning (DL) approaches show promise for ASD classification, at times outperforming traditional ML techniques (Arya et al., 2020; Bengs et al., 2020; Dvornek et al., 2017; Eslami et al., 2019, 2021; Eslami & Saeed, 2019; Heinsfeld et al., 2017; Hu et al., 2020; Khosla et al., 2018a, 2018b; M. Leming et al., 2020; M. J. Leming et al., 2021; M. Leming & Suckling, 2019; J. Li et al., 2018, 2021; X. Li, Dvornek, Papademetris, et al., 2018; X. Li, Dvornek, Zhuang, et al., 2018; Rakić et al., 2020; Sherkatghanad et al., 2020; Subah et al., 2021; Traut et al., 2021; Tzourio-Mazoyer et al., 2002; Yang et al., 2021). Unlike ML, the DL philosophy entails using data that has been less extensively preprocessed (LeCun et al., 2015). To date, however, many DL studies using fMRI data have trained their algorithms on highly processed derived functional connectivity matrices. Common workflows extensively process the functional data, then parcellate 4D scans into regions of interest (ROIs) using atlases like AAL (Tzourio-Mazoyer et al., 2002), extract mean time series within ROIs, and compute Pearson correlation matrices (Biswal et al., 1995; Dadi et al., 2019), commonly referred to as a functional connectome “fingerprint” (Finn et al., 2015). Viewed as graphs or concatenated 3D images, connectivity matrices have been classified with graphical networks or 3D CNNs for ASD (Arya et al., 2020; Khosla et al., 2018; Leming et al., 2020; Li et al., 2021; Yang et al., 2021). For instance, Khosla et al. (2018) trained 3D CNNs to predict ASD diagnosis on connectivity "fingerprint" images. Similarly, Li et al. (2021) constructed ASD and neurotypical connectivity graphs as spectral convolution network templates. Despite some promising findings, Pearson correlation matrices - functional connectome fingerprints - may discard important temporal patterns like phase shifts in comparing time series, where asynchronous responses could be meaningful. Are there alternative methods?

Several studies have incorporated temporal dynamics for ASD prediction using DL approaches. For example, Dvornek et al. (2017) generated rs-fMRI time series embeddings with LSTMs for classification. (Bengs et al., 2020; Li, Dvornek, Papademetris, et al., 2018; Li, Dvornek, Zhuang, et al., 2018) applied high-dimensional 3D/4D convolutional networks.

Capturing spatiotemporal representations is also critical in Natural Language Processing (NLP), where Transformers - described in **Chapter 2** - currently reign supreme (Vaswani et al., 2017). Inspired by their success, Transformers have been applied in medical imaging (Luo et al., 2021; Nguyen et al., 2020; Zhang et al., 2021) including fMRI analysis. I give here an overview of novel methodological approaches developed in recent works on brain task and rs-fMRI processing in order to better understand how my experiments differed from those.

Notably, Nguyen et al. (2020) compressed 4D task-fMRI into 3D embeddings using a 3D CNN, and fed these to a Transformer encoder to determine important frames for each task. In a similar study, Zhao et al. (2022) predicted task states from fMRI time series sequences using a Transformer applied to compressed spatial data. A second model took the latent representation for state prediction. While effective for task fMRI, the learned embeddings may lack interpretability and spatial relationships important for resting state modelling.

Bedel et al. (2022) incorporated spatial and temporal dynamics via a cross-window Transformer with a learned CLS token summarising latent features for classification. Malkiel et al. (2022) used a 3D CNN autoencoder to compress volumes into input representations for a downstream Transformer. However, this may discard informative spatial interactions across timeframes, better suited to task fMRI.

Yu et al. (2022) also applied self-supervised Transformers to infer functional networks in space and time. Thomas et al. (2023) compared various Transformer architectures on fMRI, finding pre-training on broad neuroimaging data improved generalisation for mental state decoding over training from scratch. Causal modelling outperformed other approaches. Kan et al. (2022) fed connection profiles of mean time series from known ROIs into a Transformer encoder. An orthonormal clustering projection enhanced discriminability for downstream prediction.

Overall, studies have demonstrated the interest of using Transformers to process fMRI data. However, the focus was on predicting a diagnosis or a defined brain state. No phenotypical or demographical data was incorporated in the optimisation of the models whereas gender, age, comorbidities are known confounds of MRI studies as explained in **Chapter 1**. Furthermore, several approaches are not relevant to study brain region interactions between various time frames (e.g. (Nguyen et al., 2020) where spatial information is compressed prior to analysing time series).

In this project, I aimed to model interactions between brain regions that may underlie autistic functioning at rest. This requires analysing spatial relationships across time series. To do this, I applied Transformers to rs-fMRI data, extracting time series using the Craddock parcellation (Craddock et al., 2012) to represent meaningful brain regions, akin to words in a sentence.

Just as Transformers find linguistic relationships, I hypothesized they could decode relationships between brain region activities related to Autism. In my investigation, I explored whether auxiliary prediction tasks—specifically, predicting demographic variables like gender and age—could improve autism classification, drawing parallels to multi-task learning. The rationale was that optimising for these additional, yet relevant, demographic variables may help the model learn more useful representations. Indeed, Werling & Geschwind (2013) contributed significantly to the discussion of sex differences in ASD (e.g. genetic factors, response to environment and physiological differences, biased diagnosis), and underscored the necessity to tailor experiments to better understand this gender influence. In the multi-task learning framework where sex and ASD diagnosis are concurrently predicted, it is also pertinent to reference Simon Baron-Cohen's influential theory (Baron-Cohen, 2002). This theory suggests that ASD might be conceptualised as an extreme manifestation of certain male-typical traits—a form of hyper-masculinity. Known as the "*Extreme Male Brain*" theory of autism, it posits that individuals on the autism spectrum tend to exhibit male-associated traits, such as a heightened systemising mechanism and a reduced empathising mechanism, but to an extreme degree. While the theory has stirred debate, its alignment with observed sex differences in ASD diagnosis rates and the behavioural characteristics of the condition emphasises its relevance. The concurrent examination of sex and ASD diagnosis in this multi-task learning model could provide insights that deepen the understanding of the interplay between sex differences and ASD, potentially shedding light on the empirical foundations of Baron-Cohen's theory (Baron-Cohen, 2002).

However, contrary to my expectations, my experiments comparing single-task (autism classification alone) and multi-task (autism classification in conjunction with gender and age prediction) Transformer architectures showed no clear performance differences. It is important to note that these auxiliary tasks are distinct from fMRI tasks and involve predicting demographic information. In the discussion, I consider refinements like loss weighting, hyperparameter optimization, and augmented data that may better promote the potential of each modelling approach.

5.3. Methods

5.3.1. Data preparation

For this project, I used rs-MRI data available from the ABIDE 1 (Di Martino et al., 2014) and HBN (Alexander et al., 2017) datasets, that are described in **Chapter 2**.

rs-fMRI data was preprocessed with the C-PAC pipeline (version 0.4.0 for Healthy Brain Network (HBN) and version 0.3.9 for ABIDE 1), with global signal correction and band-pass filtering (0.01-0.1Hz). A functional parcellation - Craddock 200 (Craddock et al., 2012) - was applied, and mean time series were extracted for each of 200 regions. ABIDE 1 preprocessed data is open source (http://preprocessed-connectomes-project.org/abide/). HBN data are available to researchers authorised to use the database. In total, time-series files for 1102 participants were available in the ABIDE 1 dataset, and time-series files for 1096 participants were available in the HBN dataset. The total number of time series files used in the experiments is lower because of the preprocessing pipeline described below.

For ABIDE 1, manual quality control annotations were provided. I retained only the scans where at least one rater assessed the scan to be of good quality. 1022 scans remained after this step. No quality control annotations were provided with the preprocessed HBN data .

For each participant, I first checked that every time-series had at least one non-zero value, that the time-series lengths were sufficiently long (the minimum length of 100 frames was chosen arbitrarily to retain as many participants as possible), and that time-series were present for each of the 200 ROIs (following the Craddock 200 parcellation). I excluded participant data that did not meet these conditions. This process generated time-series harmonised in length (100 frames) across the whole dataset. Next, each time-series was normalised separately by removing the mean and dividing by the standard deviation of the time-series.

For ABIDE 1, the full Sample 1 from the University of Michigan data collection site was selected to be the independent test set. For the HBN dataset, I left out 10% of the dataset as the independent test set, where each site was represented in proportion to its representation in the full dataset.

The remaining data were used to train the model in a 100-fold cross-validation (CV) fashion. The CV was stratified on the ASD/non-ASD labels. I used the StratifiedKFold class from scikit-learn's model_selection module in Python to generate the folds, with the random state set to 42.

Table 5.1 summarises the datasets used for each model.

	Training - Validation sets		Testing set	
	ABIDE 1	HBN	ABIDE 1	HBN
Model 1	773 (353 ASD)	/	94 (42 ASD)	/
Model 2	773 (353 ASD, 659 males)	/	94 (42 ASD, 70 males)	/
Model 3	773 (353 ASD, 277 aged between 10-15)	/	94 (42 ASD, 50 aged between 10-15)	/
Model 4	847 (413 ASD)	975 (67 ASD)	105 (51 ASD)	108 (6 ASD)
Model 5	/	975 (67 ASD)	/	108 (6 ASD)

Table 5.1: Description of data used in training-validation sets (100-folds CV) and in testing set for each model

5.3.2. Models

I first trained a binary classification model for ASD diagnosis using a Transformer encoder followed by a fully connected layer block (see **Figure 5.1**). For this initial experiment, I included data from participants with no diagnosis and those diagnosed with ASD only.

Next, I designed a simple multitask model that utilized a shared Transformer encoder as the common component for feature extraction, with separate fully connected blocks dedicated to predicting different targets—specifically, autism classification and demographic variables such as gender or age (see **Figure 5.2**).

In this study, the simple classification models were implemented with a cross-entropy loss function:

$$H(P^* \mid P) = \sum_i P^*(y|x_i) log(P(y|x_i;\theta)$$

Where:

- H is the cross-entropy between the true class distribution P^* and the predicted class distribution P
- y is the class
- x_i is an input instance
- θ are the parameters of the model

For the multitask models, a weighted sum of cross-entropy computations was used as the total loss function:

$$H_1(P_1{}^* \mid P_1) = \sum_i P_1{}^*(y|x_i) log(P_1(y|x_i;\theta_1)$$

$$H_2(P_2{}^* \mid P_2) = \sum_i P_2{}^*(y|x_i) log(P_2(y|x_i;\theta_2)$$

$$H_{sum} = \alpha H_1(P_1{}^*|P_1) + (1-\alpha)H_2(P_2{}^* \mid P_2)$$

Where:

- H_1 is the cross-entropy between the true class distribution ${P_1}^*$ and the predicted class distribution P_1
- H_2 is the cross-entropy between the true class distribution ${P_2}^*$ and the predicted class distribution P_2
- H_{sum} is the loss criterion of the model
- θ_1 are the parameters of the encoder + the FC block 1 (see **Figure X**)
- θ_2 are the parameters of the encoder + the FC block 2 (see **Figure X**)
- y is the class
- x_i is an input instance
- α = 0,5 (arbitrary choice)

I primarily monitored accuracy and AUROC as model performance metrics, and computed the mean of these two scores to evaluate overall model balance.

The Adam optimiser (Kingma et al., 2017) was used for training with a learning rate of 10^{-3} and weight decay of 10^{-7}.

I arbitrarily set the time-series representation dimension to 16.

The input embedder consists of one linear layer to project the input data into a 16-dimensional embedding space, plus a positional encoding layer similar to Vaswani et al. (2017).

The encoder block (see **Figures 5.1** and **5.2**) comprises 3 encoder layers, each containing 4 multi-head attention modules.

Post-encoding, the 200x16 representation of each input is flattened and passed through a fully connected block with 3 layers. Finally, a softmax function is applied to produce output probabilities.

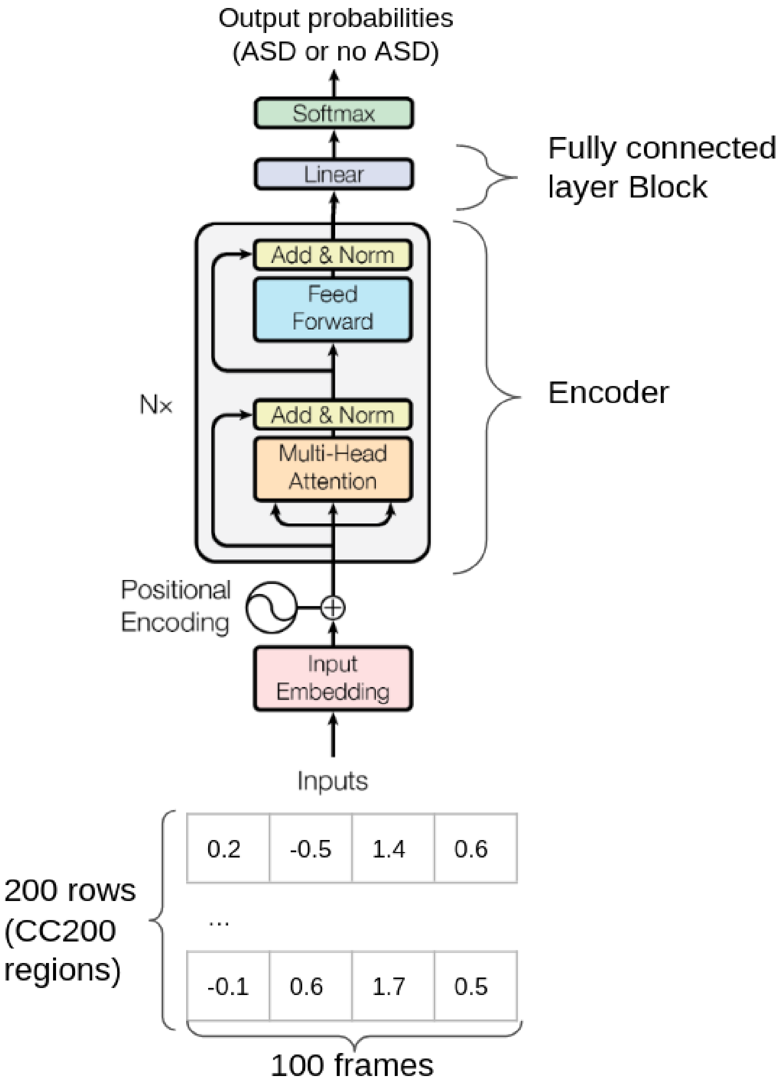
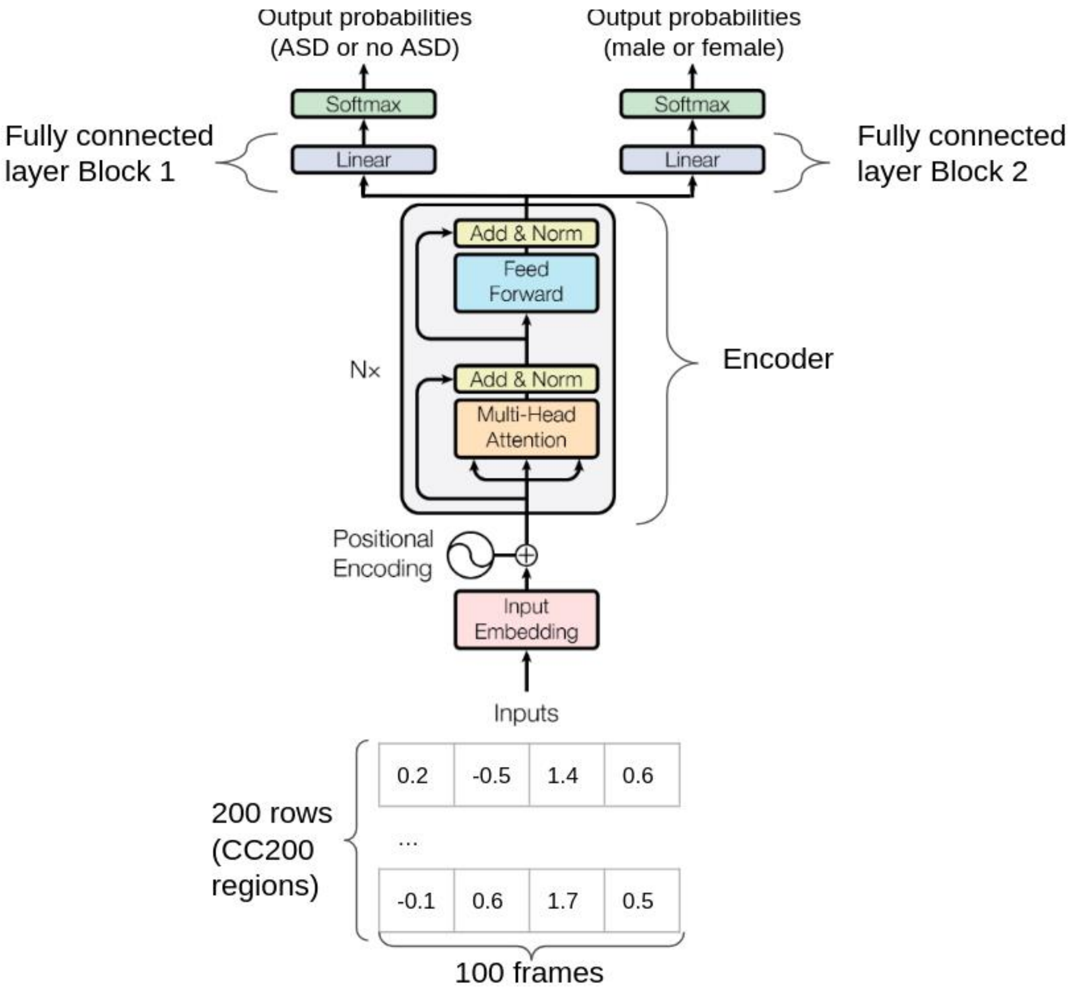


Figure 5.1: Architecture of the models 1, 4 and 5. Inputs are the 200 extracted mean time series (CC200 atlas - Craddock et al., 2012) cropped to 100 non-null frames. The Encoder part is similar to a classical Transformer encoder (Vaswani et al., 2017) and returns an intermediate representation of the inputs. The Fully Connected layer block processes a flattened representation and returns the probability of ASD diagnosis.



0.2	-0.5	1.4	0.6
...			
-0.1	0.6	1.7	0.5

Figure 5.2: Architecture of the multitask models 2 and 3. Inputs are the 200 extracted mean time series (CC200 atlas - Craddock et al., 2012) cropped to 100 non-null frames. The Encoder part is similar to a classical Transformer encoder (Vaswani et al., 2017) and returns an intermediate representation of the inputs. The Fully Connected layer block 1 processes a flattened representation and returns the probability of ASD diagnosis. The Fully connected layer block 2 processes the same flattened representation and returns the probability of being male (for model 2, or for being aged between 10-15 for model 3). The two FC blocks are optimised separately, while the Encoder is optimised taking into account the two tasks.

I led experiments on 5 models:

- Model 1 comprised a Transformer encoder followed by a fully connected block to predict ASD status (ASD or non-ASD) – see **Figure 5.1**. Model 1 was trained only on a subset of ABIDE 1 data, excluding participants with comorbid diagnoses.
- Models 4 and 5 had identical architecture to Model 1 (see **Figure 5.1**), but were trained on ABIDE 1 and HBN, and HBN only, respectively. These datasets included participants with diagnoses other than ASD.
- Models 2 and 3 were intended as multitask models, with a shared Transformer encoder and separate fully connected blocks to predict ASD status plus another binary target (gender for Model 2, age 10-15 years or not for Model 3) – see **Figure 5.2**. The training data for Models 2 and 3 matched Model 1, excluding comorbidities.

5.3.3. Interpretation methods

To evaluate model fit, I plotted training versus validation accuracy and AUROC score curves over epochs to check for underfitting or overfitting.

I defined a metric comprising the mean of accuracy and AUROC scores on the validation set per epoch per fold. The best metric over all epochs for each fold corresponds to the optimal model. Comparing best metrics across folds assessed multitask improvements. Paired t-tests or Mann-Whitney U tests were used after verifying assumptions.

I evaluated the best models (optimal per fold) on independent test sets to assess generalisation.

To visualise representations, I computed the mean post-encoder representation across subjects (see models on **Figures 5.1** and **5.2**). Pearson correlations between all 200 regions were calculated, with correlations above 0.6 visualised in a chord diagram.

Additionally, I implemented LIME (Ribeiro et al., 2016) to locally interpret model predictions. For a given input, LIME approximates the decision boundary and feature importance for that observation. These explanations may provide insights into model behaviour.

5.4. Results

5.4.1. Training performance

This section presents outcomes for the five differently designed models shown in **Figures 5.1** and **5.2**.

Models were trained for 50 epochs with validation performed at each epoch to obtain accuracy and AUROC scores on the validation fold. As a reminder, 100-fold stratified cross-validation was used, so each model was trained 100 times.

Figure 5.3 shows the evolution of accuracy and AUROC over epochs, aggregating repeated fold values to plot mean and 95% confidence intervals for training and validation sets. The models exhibit overfitting - fast convergence on training with stagnating, near-random validation performance. Models 4 and 5 have higher validation accuracy scores, likely due to imbalanced classes in the HBN dataset (~6% ASD). However, the utility of these scores is limited, as the models may be biased towards predicting the majority class. To mitigate this issue and ensure a more robust evaluation, future work could involve testing on a balanced dataset or employing techniques such as oversampling the minority class, undersampling the majority class.

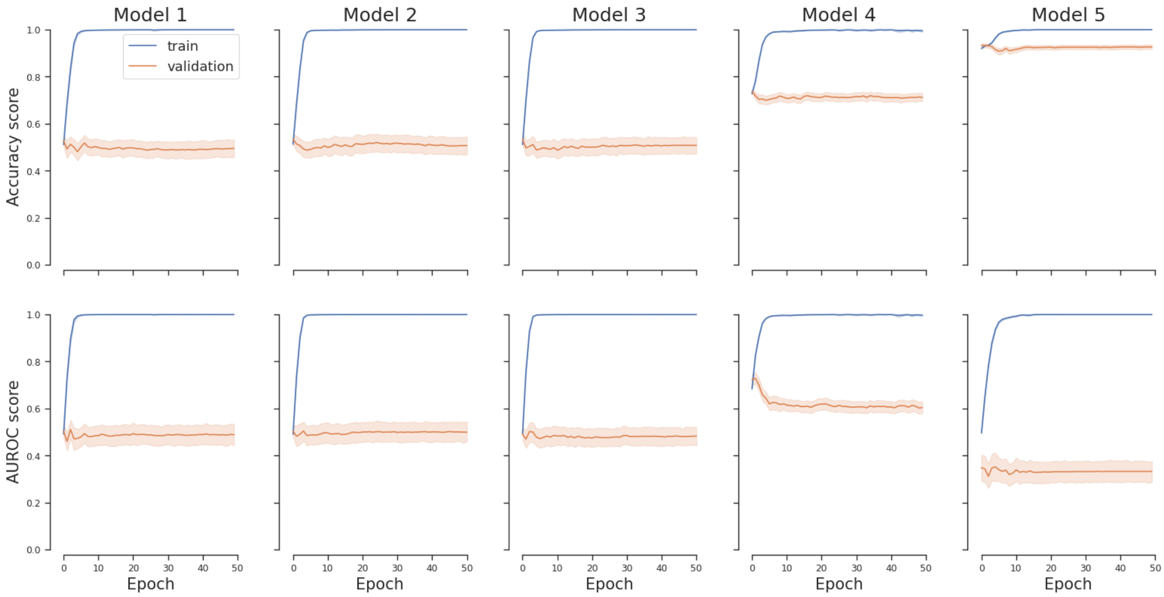


Figure 5.3: Evolution of accuracy and AUROC over epochs; mean and 95% confidence intervals are represented for training and validation sets.

Figure 5.4 shows a comparison of the best metrics computed on validation sets as (accuracy+AUROC)/2) on all the folds between the five models. The boxplots represent the distribution of the best metrics of all the folds for each model. From **Figure 5.4**, it appears that there is no clear difference between Model 1, 2 and 3 best metrics, that have respectively a mean of m_1=0.644, m_2=0.660, m_3=0.644, and a standard-deviation of s_1=0.151, s_2=0.148, s_3=0.144. However, Model 4 and 5 appear different from the others, and have respectively a mean of m_4=0.794, m_5=0.711, and a standard-deviation of s_4=0.078, s_5=0.180.

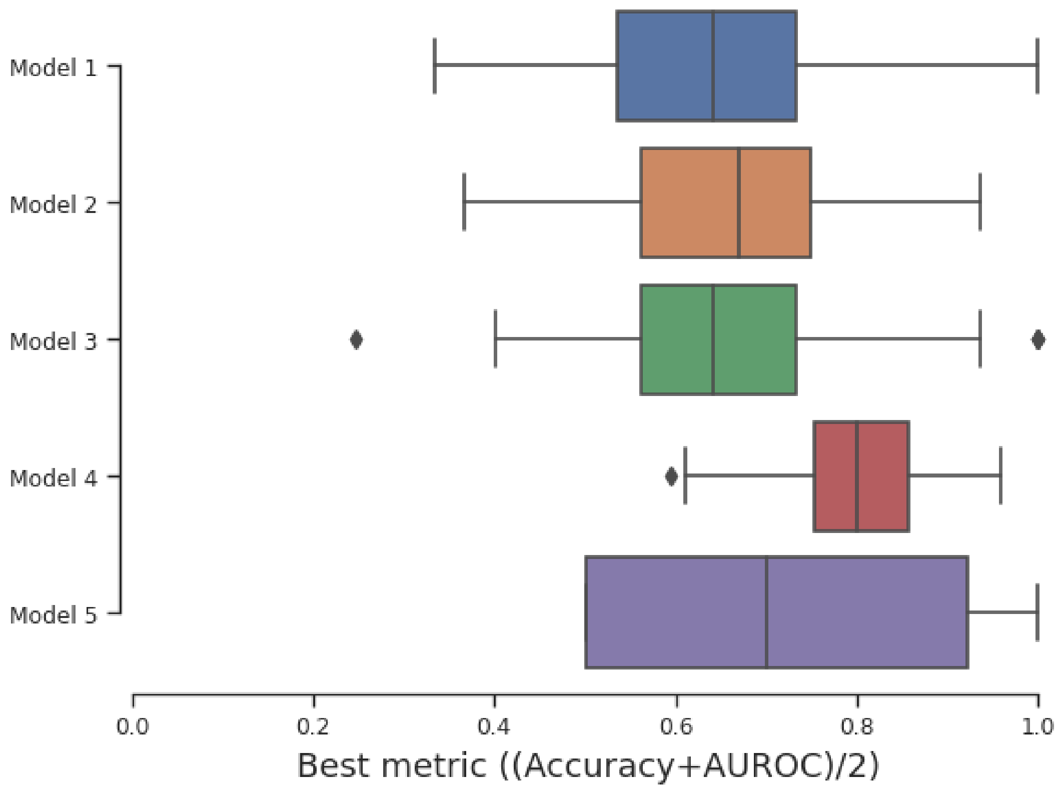


Figure 5.4: Boxplots of the best metrics computed on validation sets as (accuracy+AUROC)/2) on all the folds between the five models.

To confirm observations, I performed statistical tests comparing models under two conditions (simple vs multitask, ASD-only vs ASD+comorbidities, multitask gender vs age). Paired t-tests were suitable for comparing metric differences across folds.

Verifying normality assumptions, all differences passed Shapiro-Wilk except Models 4 and 5 vs 1 (see **Appendix 3**, **Table A3.1**). Thus, I used Mann-Whitney U tests for those pairs.

Results (**Table 5.2**) show no significant difference between Models 1, 2, and 3 (p_{m1-m2}=0.435, p_{m1-m3}=0.984, p_{m2-m3}=0.199). However, there were significant differences between Model 1 and Model 4 (p_{m1_m4}=0.0158 < 5%), and Model 1 and Model 5 ($p_{m1\_m5} < 10^{-13}$).

Paired T-test	T	dof	Alterna-tive	p_val	CI95%	cohen-d	BF10	power
Model 1 - Model 2	0.783	99	2-sided	0.435	[-0.02, 0.05]	0.103	0.149	0.176
Model 1 - Model 3	-0.0206	99	2-sided	0.984	[-0.04, 0.04]	0.003	0.111	0.050
Model 2 - Model 3	1.293	99	2-sided	0.199	[-0.01, 0.04]	0.109	0.248	0.189
MW U-test	U-val		Alternative	p_val	RBC	CLES		
Model 1 - Model 4	8081.0		2-sided	5,17.10^{-14}	-0.616	0.808		
Model 1 - Model 5	5981.5		2-sided	0.0158	-0.193	0.598		

Table 5.2: Statistical tests on the best metrics of the validation fold results between the models.

5.4.2. Inference on test set

	Model 1	Model 2	Model 3	Model 4	Model 5
Accuracy	50%	52.1%	47.9%	67.1%	85.2%
AUROC	0.581	0.551	0.599	0.599	0.368
(Acc. + AUROC)/2	0,541	0,536	0,539	0,635	0,61
Specificity	0.442	0.808	0.327	0.776	0.902
Sensitivity	0.571	0.167	0.667	0.386	0.0

Table 5.3: Summary of test set results for each model: accuracy, AUROC, specificity, sensitivity.

Table 5.3 presents test set results for each model, including accuracy, AUROC, specificity, and sensitivity. Model 5 achieved the highest accuracy (85.2%), while Models 3 and 4 showed the top AUROC scores (0.599). Model 5 had the highest specificity (0.902) and Model 3 the highest sensitivity (0.667). A drop in the mean metric (Accuracy+AUROC)/2 is observed, with Model 4 having the best value (0.635).

The data description in **Table 5.1** shows highly imbalanced classes for Models 4 and 5 (~26% and ~6% ASD prevalence) compared to Models 1-3 (~45% ASD). Despite higher accuracy, Model 4's performance cannot be directly compared to Models 1-3 due to this imbalance. For example, Model 5 has 0 sensitivity but 85.2% accuracy, correctly predicting only non-ASD participants.

In summary, class imbalance introduces biases making accuracy metrics non-comparable between models. Future work should incorporate calibration strategies for balanced benchmarking. No model emerges as singularly optimal, but refinements to both approaches show promise in advancing ASD prediction.

5.4.3. Visualisation and interpretation

The 200 Craddock atlas (Craddock et al., 2012) time series undergo transformations through the Transformer encoder, resulting in a 16-feature representation per region before the fully connected block. I computed Pearson correlations between regions under this implicit 16-dimensional encoding.

To simplify visualisation, I translated the 200 Craddock regions into the 7-network Yeo atlas (Yeo et al., 2011). Chord diagrams in **Figure 5.5** show correlation results in this Yeo space. "Glass brain" visualisations in **Figure 5.6** depict these correlations mapped onto brain networks.

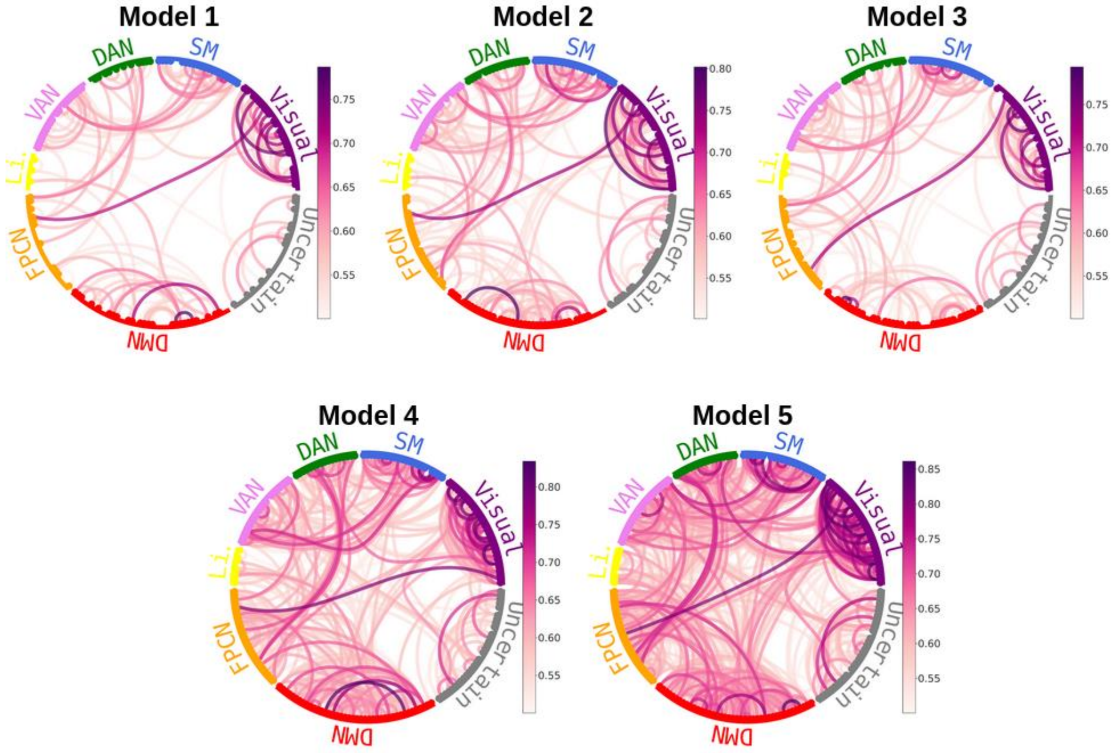
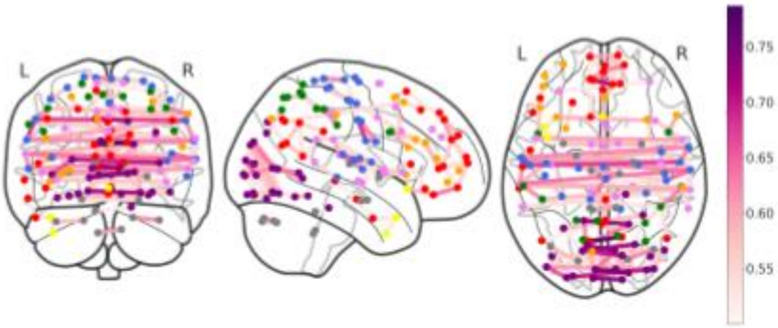
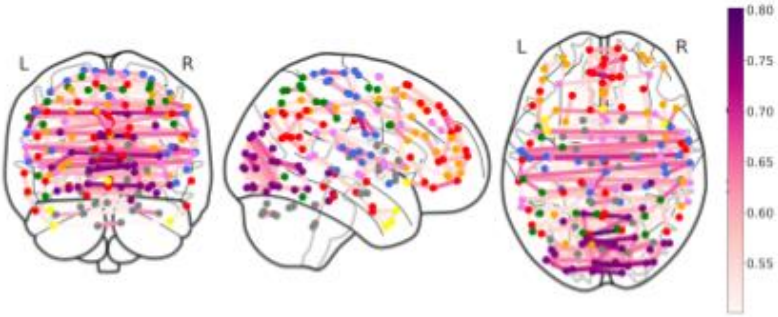


Figure 5.5: Chord representations of the correlations between the implicit representations of 200 regions by 16 features after the last layer of the Transformer encoder part of each model. The regions were translated into the Yeo Network (Thomas Yeo et al., 2011) to simplify the plot: "Visual" (purple) is for the Visual Network; "SM" (blue) is for the Somatomotor Network; "DAN" (green) is for the Dorsal Attention Network; "VAN" (violet) is for the Ventral Attention Network; "Li." (yellow) is for the Limbic Network; "FPCN" (orange) is for the Frontoparietal Control Network; "DMN" (red) is for the Default Mode Network; "Uncertain" (grey) is for regions in the CC200 atlas that did not match any region in the Yeo atlas. Inside the circles, I represented as pink lines with varying intensity levels the correlations between 0.5 and 1 (no negative correlation was found to have an absolute value greater than 0.5). The Python library Nichord (https://github.com/paulcbogdan/NiChord) was used.

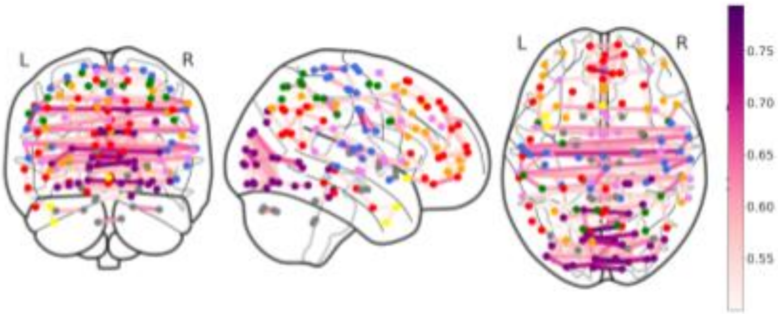
Model 1



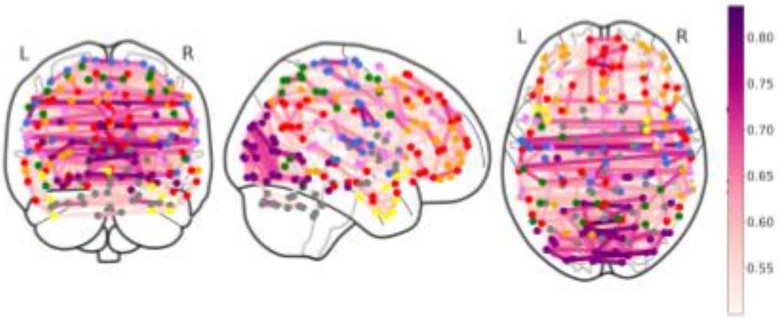
Model 2



Model 3



Model 4



Model 5

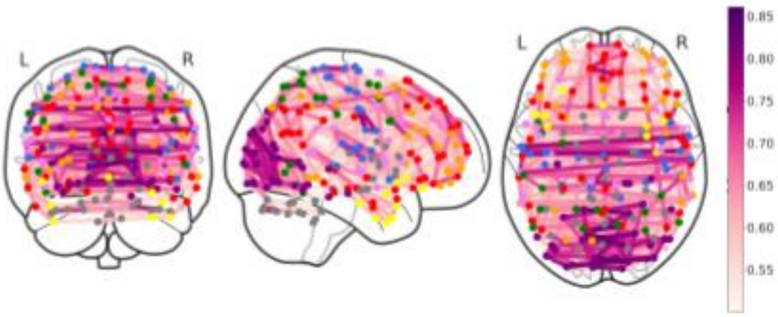
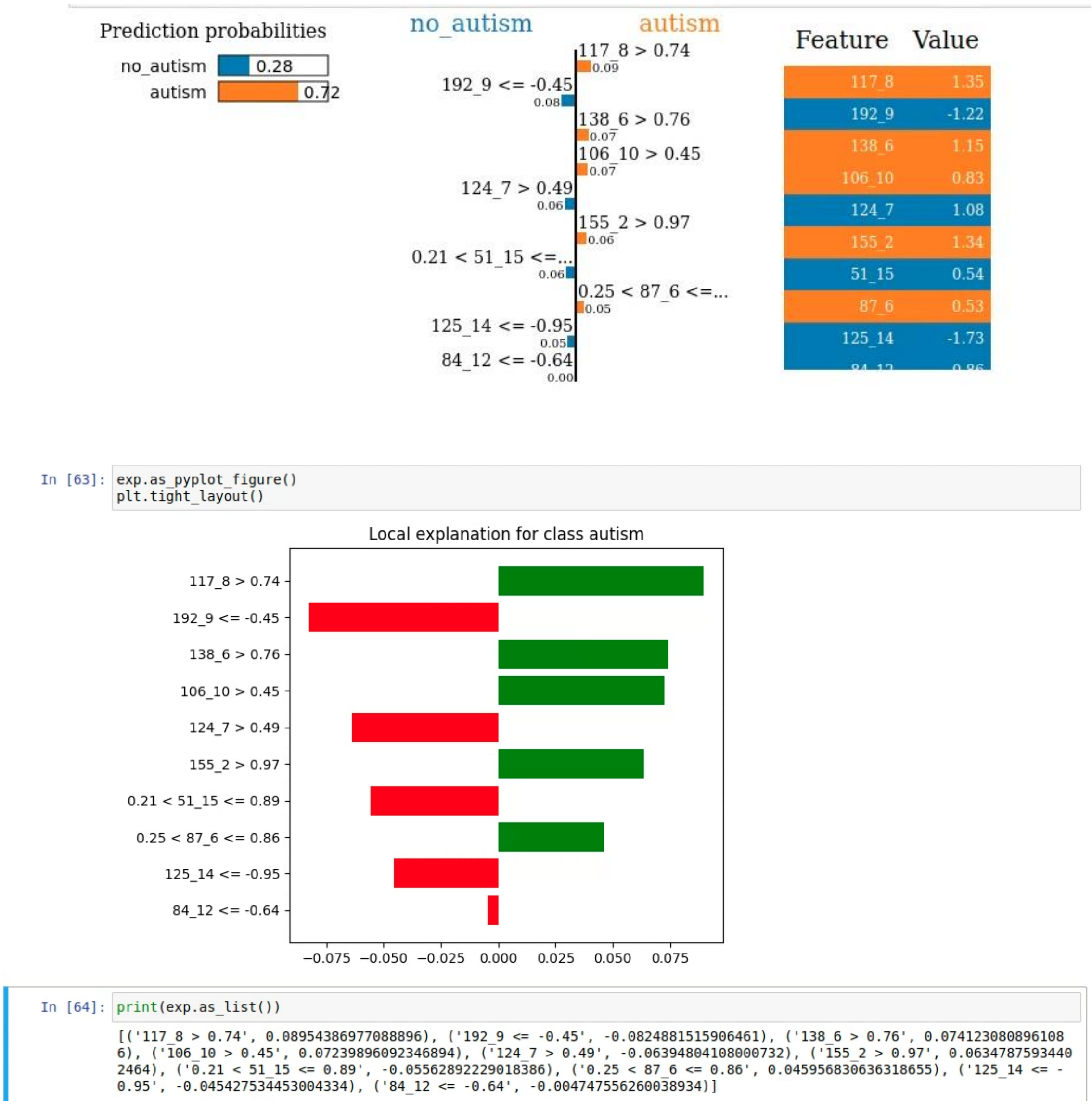


Figure 5.6: Brain representations of the correlations between the implicit representations of 200 regions by 16 features after the last layer of the Transformer encoder part of each model. The regions were translated into the Yeo Network (Yeo et al., 2011) to simplify the plot. Each CC200 region is represented by dot points and coloured in function of their correspondence with the Yeo atlas: Purple is for the Visual Network; Blue is for the Somatomotor Network; Green is for the Dorsal Attention Network; Violet is for the Ventral Attention Network; Yellow is for the Limbic Network; Orange is for the Frontoparietal Control Network; Red is for the Default Mode Network; Grey is for regions in the CC200 atlas that did not match any region in the Yeo atlas. I represented as pink lines with varying intensity levels the correlations between 0.5 and 1 (no negative correlation was found to have an absolute value greater than 0.5). The Python library Nichord (https://github.com/paulcbogdan/NiChord) was used.

In addition to visualisations, I implemented the model-agnostic algorithm LIME (Ribeiro et al., 2016) to locally interpret model predictions. For a given input, LIME approximates the decision boundary and weights feature importance for that sample. These local explanations can provide insights into model behaviour.

I did not conduct a full LIME analysis across all participants, such as separating by diagnosis or segmenting by age, gender, and comorbidities. Proper LIME implementation requires optimising many parameters including data normalisation, model settings, and LIME hyperparameters (e.g. number of random projections). This extensive tuning enables robust feature importance mapping and represents an exciting area for future work to elucidate how models predict autism (and auxiliary targets like age and gender for multitask models).

As an initial example, **Figure 5.7** shows the top 10 features driving autism classification for one participant with only an ASD diagnosis, explaining the importance via decision thresholds on continuous variables.



Feature	Value
117_8	1.35
192_9	-1.22
138_6	1.15
106_10	0.83
124_7	1.08
155_2	1.34
51_15	0.54
87_6	0.53
125_14	-1.73

In [63]:
```
exp.as_pyplot_figure()
plt.tight_layout()
```

In [64]:
```
print(exp.as_list())
```

```
[('117_8 > 0.74', 0.08954386977088896), ('192_9 <= -0.45', -0.0824881515906461), ('138_6 > 0.76', 0.074123080896108
6), ('106_10 > 0.45', 0.07239896092346894), ('124_7 > 0.49', -0.06394804108000732), ('155_2 > 0.97', 0.0634787593440
2464), ('0.21 < 51_15 <= 0.89', -0.05562892229018386), ('0.25 < 87_6 <= 0.86', 0.04595683063631865 5), ('125_14 <= -
0.95', -0.0454275344530043 34), ('84_12 <= -0.64', -0.004747556260038934)]
```

Figure 5.7: LIME algorithm executed on one autistic participant data: it explains the decisions of the Fully Connected Layer of the Model considered (e.g. Model 1). The feature names are provided with numbers. For instance, “117_8” is for region 117 of CC200 atlas and encoding feature number 8 (among the 16 ones) after the Transformer encoder part of the model. For this feature, the value is strictly greater than 0.74 that is the threshold found by the LIME algorithm, meaning that it is consistent with a prediction of Autism for LIME. The two types of graphs are displayed in a Python 3 Jupyter Notebook.

5.5. Discussion

This study explored innovative applications of the Transformer algorithm (Vaswani et al., 2017) to analyse brain activity patterns in resting-state functional MRI data for autism classification. I developed several modelling approaches: Model 1 was a basic binary classifier trained on data from individuals with autism and neurotypical comparisons. Models 2 and 3 took a multi-task learning approach, jointly predicting autism diagnosis along with gender or age group on the same dataset. This was motivated by known gender differences (Beggiato et al., 2017; Van Wijngaarden-Cremers et al., 2014; Zeidan et al., 2022) and age-related changes (Sanders, 2015; Van Wijngaarden-Cremers et al., 2014; Wolfers et al., 2019) in autism phenotypes. As discussed in **Chapter 1**, gender and age are known important variables for the characterisation and identification of ASD. In **Chapter 4**, it was also shown that depending on age and gender, replicable brain regions that were the most important to predict ASD differed. Models 4 and 5 expanded the binary autism classifier to include individuals with comorbid diagnoses like ADHD, anxiety, and depression. Such comorbidities are common in autism, and many psychiatric disorders show overlapping neural correlates and symptoms like discussed in **Chapter 1**. Overall, this work aimed to explore the potential for Transformer architectures to capture informative patterns in brain activity time series and improve autism classification. The multi-task and comorbidity-inclusive approaches were creative ways to incorporate additional, relevant phenotypic information to potentially enhance model performance. This study provides promising initial results and suggests future directions for further developing neuroimaging-based classifiers using state-of-the-art DL methods.

The results demonstrate that no single model emerged as an unambiguous top performer for autism spectrum disorder (ASD) prediction. Imbalances between datasets in terms of the representation of ASD diagnosis introduced biases that made direct accuracy comparisons between certain models unfair. Future work could incorporate strategies to calibrate models trained on imbalanced data to enable fairer benchmarking.

Models 1, 2, and 3 were reasonably comparable overall. Contrary to expectations, the multi-task approach did not boost global performance, though it did impact specificities and sensitivities considerably (sp1 = 0.442, sp2 = 0.808, sp3 = 0.327; se1 = 0.571, se2 =

0.167, se3 = 0.667). This aligns with expectations that auxiliary tasks would modulate model learning. However, multi-task models did not converge to more stable performance like Model 1.

Several factors may explain these observations. Multitask models summed task losses with arbitrary balancing (α=0.5). Optimising α more systematically could help. Additionally, complex inter-task loss relationships beyond a linear sum likely exist. Learning rates were fixed across tasks; optimising these independently may improve outcomes. Furthermore, no quality control was performed on HBN data while this step may be crucial here to avoid distorted information.

In a nutshell, while failing to improve overall accuracy, multi-task learning impacted model performance in nuanced ways. With refined loss weighting, learning rates, and other enhancements, multi-task and single-task approaches show promise for distilling insights about brain function from neuroimaging data to advance ASD prediction.

The use of 100-fold cross-validation for model training enabled robust performance estimation, although at the cost of greater computational demands compared to a standard train-test split. The large number of folds likely improved the reliability of the evaluated metrics. However, this extensive cross-validation constrained the extent of parameter optimization that could be completed within project timelines. Many architectural and training hyperparameters warrant deeper investigation in future work, including encoder layer count, attention heads, positional encoding, learning rates, regularisation, and loss weighting.

In particular, the dimensionality of the time series representations may significantly impact model performance. The extracted 100-frame time series were compressed to 16 feature vectors, aiming to maximise information density. However, this compressed size could overlook important signals or relationships in the resting state data. Optimising representation dimensionality could better capture the complexity of whole-brain dynamics. Overall, this study developed a solid computational framework and baseline modelling results. With expanded hyperparameter tuning, the Transformer-based architectures show strong promise for decoding meaningful spatiotemporal patterns from rs-fMRI in relation to autism diagnosis (Kan et al., 2022).

This study highlights several interesting areas for future investigation. The 100-frame input sequences, though substantial, may not fully capture complex spatiotemporal dynamics across diverse brain networks at rest. Longer inputs could better model these interactions. Additionally, Transformers thrive on large datasets - the scale here, though sizable by fMRI standards, is small relative to typical Transformer applications, and Autistic individuals were underrepresented in the HBN dataset. Given the phenotypic heterogeneity and inter-individual variability inherent to Autism, larger, balanced datasets will be key for future investigations. In a multi-task learning framework, data must be representative for each of the tasks involved, hence increasing the scale. In future work, assessing data sufficiency could involve evaluating the complexity of the tasks (e.g. by looking at the distribution of input and target variables, studying the training convergence, optimising model architectures…), the representativeness of the dataset (e.g. by looking at clinical and demographical variables), and model performance on validation sets (e.g. by using k-fold cross-validation technique). Traut et al. (2022) suggested estimating data sufficiency by analysing training performance as a function of the number of training subjects. High-dimensional tasks and those with significant variability demand larger, more comprehensive datasets. In cases of data limitations, strategies such as data augmentation, synthetic data generation, and transfer learning could be employed. Pooling resources with other research entities to expand datasets may be beneficial too. Dimension reduction could also help in reducing the complexity of the tasks. Future experiments could be inspired by the work of Iwana & Uchida (2021) who built a taxonomy of time-series data augmentation (totalling 4 categories and 12 techniques) and highlighted pros and cons of every technique by leading a comparative study over 128 datasets and 6 types of ANN.

There are also open questions related to input data characteristics. Our ABIDE preprocessing retained most scans, but some artefacts, particularly related to motion, likely remain. Spatial normalisation to template space may also distort signals. This issue was discussed in **Chapters 1 & 4**. Resting-state alone may not provide sufficient signal. Given my earlier empirical study predicting Autism diagnosis from structural MRI data (**Chapter 4**), multimodal integration (e.g. sMRI, PET, EEG, genetics) could improve prediction accuracy (Traut et al., 2021; Horien et al., 2022). Controlling for factors like acquisition parameters and participant demographics may also be beneficial.

This study presented initial interpretability analyses, but Transformer model explanations remain challenging due to their multiple layers and attention heads. I visualised the final encoder representations to extract region-to-region relationships. However, studying intermediate representations could provide additional insights into how spatial patterns and dynamics evolve through the network. LIME highlighted influential regions for single-subject predictions, but generalising these local explanations across the whole dataset is an important next step. The fully-connected block offers limited interpretability; replacing it with more interpretable algorithms like regression or decision trees is an interesting idea.

Compared to recent Transformer studies, my models achieved lower performance, though Kan et al. (2022) noted stability challenges with ABIDE and addressed this via stratified splitting by site, age, and gender. While I balanced age and gender overall, fold-level stratification could improve robustness. As in (Thomas et al. 2023), more extensive hyperparameter optimization is needed. Positional encoding choices also strongly impact models - (Kan et al., 2022) found adjacency matrices superior to the original formulation provided by Vaswani et al. (2017). Ultimately, multimodal fusion of structural and functional data may hold the most promise, as the top model in (Traut et al., 2021) combined sMRI and rs-fMRI. Incorporating complementary anatomical patterns could enhance accuracy and interpretability

In summary, this work implemented a reasonable starting point for Transformer model interpretation in this novel application area. As a foundation for future research, several promising directions were identified such as: analysing intermediate representations, aggregating local explanations, and substituting alternative interpretable modules. Enhancing interpretation methods will lead to greater knowledge of how these models encode predictive fMRI signals related to Autism, advancing applicability in healthcare settings. Expanding the datasets, input features, model capacity, and controllable variables represent exciting opportunities to build on these foundations in future studies. Leveraging the full breadth of neuroimaging, clinical, and demographic data could ultimately yield more robust and generalisable models.

5.6. Conclusion

This study introduced innovative DL architectures for decoding brain dynamics, laying the groundwork for the development of robust and explainable AI systems that leverage diverse neuroimaging, clinical, and demographic data. Applied to Autism, this approach could offer new avenues towards reproducible biomarkers that will advance autism prediction, personalised diagnosis, and treatment in healthcare applications.

6. Side projects: improving Reproducibility in Neuroimaging

Alongside the empirical research conducted for this thesis, I participated in various complementary projects, largely aimed at enhancing reproducibility in neuroimaging. I describe these various activities below, through which I gained well-rounded experiences in skills like scientific communication, collaboration, and outreach.

6.1. Co-leading a Journal Club on Reproducibility

During the first year of my PhD, I was strongly advised to use reproducible methods in my research, especially by Dr. Clare Kelly who, as an expert in neuroimaging, had flagged the main global concerns in the field. With my colleague Jivesh Ramduny, and under the supervision of Dr. Clare Kelly, we started a journal club (JC) at Trinity College Institute of Neuroscience in order to build awareness of that matter in the local neuroscientific community.

We launched the first ReproducibiliTea JC in Ireland. The JC consisted of regular meetings to discuss and debate on scientific practices around reproducibility, including, for instance, open data, open code, paper transparency on methods and outcomes, and existing tools that are being developed to improve the reproducibility and replicability of studies in neuroscience.

Despite the lockdown situation in 2020 due to Covid-19, we continued to organise online meetings and presentations. We ended the year 2020 by co-authoring an article on the open platform Medium that summarised everything we had talked about during the year: https://ramdunyj.medium.com/one-year-of-irelands-first-reproducibilitea-journal-club-a4c217767480 .

6.2. Works on BIDS and BIDS-apps

In 2020, thanks to the regular ReproducibiliTea JC sessions, I discovered many new tools and norms in neuroimaging. The Brain Imaging Data Structure (BIDS) was one that appeared particularly useful to me.

“With the Brain Imaging Data Structure (BIDS), we describe a simple and easy to adopt way of organizing neuroimaging and behavioural data” (from https://bids.neuroimaging.io/).

Standardising the way to organise datasets has many positive effects on reproducibility, including the possibility of developing applications that are generically reusable on any BIDS dataset.

I seized the opportunity to make BIDS-apps for my first PhD project on the automatic quality control of sMRI scans (**Chapter 3**) to dramatically enhance the reusability of the tool within the neuroimaging community. The project and the code of the various apps are available on GitHub : https://github.com/garciaml/BrainQCNet .

In addition to developing BIDS-apps, I was offered the opportunity to participate in co-writing a book on methods for analysing large neuroimaging datasets (https://osf.io/d9r3x/), in which I shared an overview of BIDS and a tutorial on BIDS-apps. The preprint is available here: https://osf.io/rcxg8/ .

6.3. Git and GitHub tutorials

I was invited to make a second contribution to the book on methods for analysing large neuroimaging datasets (https://osf.io/d9r3x/): a tutorial on git and GitHub. git and GitHub dramatically increase the reusability of the code of any project and foster collaborations. The preprint is available here: https://osf.io/jqwpv/ .

6.4. Reviewing (Horien et al., 2022)

At the end of 2021, Dr. Clare Kelly offered me the opportunity to review a paper for the journal Biological Psychiatry. The article reviewed predictive modelling methods of ASD based on fMRI data.

The published article is available here: Horien, C., et al., 2022. Functional Connectome–Based Predictive Modeling in Autism. Biological Psychiatry. https://doi.org/10.1016/j.biopsych.2022.04.008 .

My review on the preprint is presented in **Appendix 4**.

6.5. Teaching

From the second year of the PhD to the last one, I was a teaching assistant and co-ran labs on Python programming and statistics for first-year undergraduate Psychology students. I experienced significant progression in teaching and communication skills over the course of these three years. Skills such as oral communication, English language, class management, leadership, empathy, all grew significantly through this experience.

Overall, all these extra-curricular activities provided a dynamic and multidimensional complement to my core PhD research, opening my eyes to new perspectives and possibilities for alternative pursuits to explore in my career as a researcher.

7. Training courses

Along with taking part in side projects, it has been very important for me to continuously update my knowledge, to engage my curiosity, and to keep learning new things. I took every relevant opportunity I had during the PhD to engage in training. Below, I describe the various summer schools I attended.

7.1. ENERGHY 2021

In July 2021, I participated in the 3-week ENERGHY Summer School - Energising Global Health Innovation and Entrepreneurship. The Summer School took place remotely, due to the pandemic. During the Summer School, I encountered many concepts and useful practices related to social entrepreneurship, taught by international experts on the topic. I also worked on a project as part of a team (randomly allocated) for the Social Purpose Organisation FRIENDSHIP, based in Bangladesh, as well as for the SANOFI Espoir foundation (a main FRIENDSHIP partner).

The project was to build a business model (BM) to underpin the distribution of a new cooking stove in the most rural and remote areas in Bangladesh. Severe burns (requiring reconstruction surgeries) and chronic respiratory diseases are frequently experienced by women and children in these areas, due to very poor infrastructures and houses, in particular, substandard kitchens. The new stove design, made by FRIENDSHIP, was designed to be safer, and took into account the local availability and price of the materials, and the ease of fabrication.

During the summer school, my team suggested a BM inspired by Venture Philanthropy instead of a Charity model that would make people dependent on big funders. The stove could be sold at a very low price (taking into account the minimum revenue of the families in these areas), and this new activity could foster the development of local economies while reinforcing social bonds and community solidarity. Our pitch video can be viewed here:

https://drive.google.com/file/d/19uHAlzREA8nlo_IOX_ybcAQNOzsfgxYR/view?usp=sharing .

We won the Summer School challenge and were offered an opportunity to participate in a 6 month programme on entrepreneurship in Health with the University of Barcelona. The goal was to continue building the business model, mentored by two experts in social entrepreneurship, and to pitch it to a jury at the end.

To help advance the project after the end of the Summer School, I took the lead on managing the project and the team. It was a good exercise - I did my best to organise our tasks in the most efficient way, and to facilitate as much as possible the implementation of work. I learned how to keep communicating with people and to motivate them, keeping in mind that what we had undertaken was voluntary work.

Overall, in my view, learning how to build a social BM, and how to lead and manage a team and a project in the context of volunteering was a great experience.

7.2. OxML 2021 & 2023

The Oxford Machine Learning summer school (OxML) takes place every year, covering the most up-to-date machine learning and deep learning techniques. It covers some of the most important topics in machine learning (ML) and deep learning (DL) in which the field is showing a growing interest (e.g., statistical/probabilistic ML, representation learning, reinforcement learning, causal inference, vision & NLP, geometrical DL) and their application to sustainable development goals (SDGs) (https://www.oxfordml.school/).

I attended the school online in 2021 and in-person in 2023 for the part on Health.

These experiences enabled me to continue developing my skills in machine learning and deep learning, as well as to keep up to date with practices and use-cases across many fields.

7.3. RYLA 2022, sponsored by the Rotary Club Paris Concorde

“Rotary Youth Leadership Awards (RYLA) is an intensive leadership experience organised by Rotary clubs and districts where you develop your skills as a leader while having fun and making connections.” (https://www.rotary.org/en/our-programs/rotary-youth-leadership-awards)

I had the chance to be sponsored by the Rotary Club Paris Concorde to participate in the RYLA edition 2022 organised by the district, from the 21st to the 25th of February 2022.

I was one of 16 young adults from various backgrounds who had been selected based on my CV and cover letter.

The training lasted a week, during which I developed skills like management, leadership, communication, as well as ethics. I also made new connections and friends.

The RYLA was a great opportunity to progress on non-technical aspects that are important for my career in science.

7.4. ECNP Immuno-NeuroPsychiatry Bordeaux Summer School 2022

From the 18th to the 22nd of July 2022, I had the opportunity to participate in the ECNP (European College of Neuropsychopharmacology) Immuno-NeuroPsychiatry Bordeaux Summer School 2022. The week featured a host of international speakers who introduced their work on the emergent topic of Immuno-NeuroPsychiatry. As a young researcher willing to continuously self-train and be up-to-date, it was exciting to discover this area of biological psychiatry. I also made new connections and friends that opened potential future collaborations.

7.5. ARAPI 2022

The ARAPI - “Association pour la Recherche sur l’Autisme et la Prévention des Inadaptations” - is a French association that organises a week-long autumn school on Autism every two years. The goal is to update all the practitioners, parents, associations, and autistic people on the recent advances in research, law, or tools developed to improve the living conditions of Autistic people. I attended the school from the 3rd to the 7th of October 2022, where many international speakers described their research. Discovering the view of the educators and parents asking researchers practical questions stimulated a constructive critical self-appraisal of my own research.

7.6. NeuroHackademy 2023

“NeuroHackademy is a summer school in neuroimaging and data science, held at the University of Washington eScience Institute.”

(https://neurohackademy.org/)

I attended the first part of the school online from the 7th to the 11th of August 2023. As a young researcher, NeuroHackademy contributed significantly to my goal to stay updated on existing tools in neuroimaging and data science - I would strongly recommend any PhD student in the field attend the summer school. I was positively surprised by the broad range of classes and activities proposed.

By participating in the school, I improved my skills in software development, in collaborating with git and GitHub, and in various applications in neuroimaging. I also learned about AI-assisted programming, data governance, and careers in neuroimaging and data science.

I was not able to spend much time socialising virtually, despite the fact that there were many interesting planned activities. Unfortunately I was unable to attend the second week, which featured a team-based hackathon. I was sorry to miss this, as such events provide invaluable opportunities to meet new people from the community and to make connections that are vital to the career of a young researcher.

7.7. Graduate Teaching Assistant

During summer 2022, I attended a course at Trinity College Dublin to improve my skills as a Graduate Teaching Assistant (GTA). Seven learning blocks introduced the role of a GTA, the various ways in which students learn, communication and coping strategies, session planning, the importance and means of assessment and feedback, and how to design an online activity. I was also asked to reflect on and to evaluate my teaching.

These classes helped me to evolve the way I teach and gave me more confidence as a GTA. I found it very useful to write a teaching philosophy statement and to construct a session plan, taking into account the concepts and tools learned during the training.

7.8. Research Integrity

The first year of the PhD, I undertook a mandatory but highly relevant course at Trinity College Dublin on Research Integrity and Impact in an Open Scholarship Era. I had no particular knowledge about the topic, except my own conceptions on what constitutes ethical research.

This course covered a broad range of topics, including copyright and data protection, data management and security, scholarly communication and open research, research evaluation and research impact.

Learning about these at the beginning of the PhD was very helpful, since I subsequently had to deal with many of the principles and tools over the course of my PhD.

Overall, attending these schools has made me deeply sensitised to principles of reproducibility in research, emerging topics spanning psychiatry, neuroimaging, and DL, a diverse array of soft skills, and has opened my mind to the possibility of exploring alternative paths like social entrepreneurship.

8. Discussion and Conclusion

8.1. Summary

This PhD thesis sought to develop interpretable DL models to identify neuroimaging biomarkers of Autism spectrum disorder (ASD). Three core projects focused on structural MRI quality control, structural MRI biomarker discovery, and functional MRI analysis using Transformer models. These efforts have contributed both methodological advances and new evidence to support the use of biologically-grounded AI to elucidate neural markers for Autism. Additional open science contributions were made through developing new tools, standards, and by educating peers. While several limitations motivate recommendations to address challenges related to model interpretation, biases, and optimization, overall, this PhD thesis has successfully developed an explainable imaging analysis framework that, with further refinement, has the potential to elucidate and quantify the heterogeneous neurobiological underpinnings of ASD in a clinically meaningful way.

8.1.1. Project 1

Manual quality control of structural MRI data is essential but time-consuming. To address this, I developed an interpretable DL model called BrainQCNet to automatically detect artefacts in structural brain scans. After manually annotating 980 scans from the ABIDE 1 dataset, the model was trained, validated (during training) and tested (after training), achieving over 90% accuracy on this initial testing set. The optimised BrainQCNet model was then evaluated on three large-scale independent datasets - ABCD (2141 scans), ADHD200 (750 scans), and ABIDE II (799 scans) - demonstrating excellent performance, with 91.4% sensitivity for detecting artefacts in the ABCD dataset, compared to human raters. Critically, BrainQCNet showed higher sensitivity than previous methods while requiring no intensive scan preprocessing, such as normalisation to template space. However, some patterns require further examination to determine their clinical relevance. In particular, at a local level, it was not clear whether all the patterns detected by the model were relevant to the prediction. Future work will pursue these open questions. To support open adoption, several BIDS apps implementing BrainQCNet on GPU and CPU systems were developed and all code was publicly released on GitHub under an open licence. In sum, this project showed DL can rapidly automate and enhance sMRI quality control to improve the reliability of downstream analysis.

8.1.2. Project 2

Standard neuroimaging pipelines rely on intensive preprocessing like spatial normalisation to template space, which may obscure subtle brain patterns associated with Autism. To avoid this, a DL approach using 3D CNNs to predict and interpret Autism from structural MRI scans without spatial normalisation was developed. Two CNN architectures, DenseNet121 and ResNet50, were trained and tested and compared across multiple datasets including ABIDE 1 and 2, and ADHD200. This cross-dataset convergence provided more robust results. The models achieved 50-70% prediction accuracy for Autism; lower prediction accuracy was achieved for participants with comorbid diagnoses. Using guided grad-CAM visualisation, replicable predictive brain regions across models and datasets were identified, including frontal, limbic, and cingulate areas. The importance of these regions aligns with current Autism neuroscience findings. Granular analysis also revealed some differences in predictive regions by gender and age. Critically, models did not rely on non-brain background for prediction. By avoiding potentially biassed preprocessing while revealing interpretable neuroimaging patterns, this work provides clinically-grounded DL biomarkers for Autism. Integration and validation across datasets bolsters generalisability. Sharing the whole code contributes to the wide reuse and development of new approaches based on my work. Future directions include improving the models, performing multimodal analysis, integrating confounds (e.g. age, gender) as predictive variables to the model. Overall, the project advances biologically-informed ML for Autism diagnosis while mitigating risks of standard preprocessing pipelines.

8.1.3. *Project 3*

Transformers have shown promise for sequential data modelling. In this last project, I applied Transformer architecture to resting-state fMRI data from the ABIDE 1 and HBN datasets, using 2035 subjects to classify Autism and capture complex spatiotemporal patterns. The data was preprocessed with C-PAC pipelines and parcellated into Craddock 200 atlas regions, from which mean time series were extracted. Multiple Transformer configurations were tested using 100-fold cross-validation, including pioneering multitask models that also predicted gender and age. Cross-validation constrained model exploration but improved evaluation. All models achieved approximately 64.4% to 79.4% (standard deviation between 0.078 to 0.180) of defined metrics - accuracy and AUROC summed and averaged - but overfitted training data, likely from limited data, imbalanced representation of Autistic participants, and model complexity. To interpret learned representations, chord diagrams showed models partially captured functional connections. I implemented LIME, a local explanation method, to explain individual predictions. However, LIME explanation was limited by the high dimensionality of fMRI data, preventing quantitative analysis across subjects. Further hyperparameter optimization and regularisation may reduce overfitting and improve generalisability. Though predictive performance was modest, this novel application of Transformers with multitask learning to fMRI data demonstrated potential for discovering non-obvious imaging biomarkers of Autism informed by neuroscience priors.

8.1.4. *Side projects and trainings*

Early in the PhD, I engaged directly in efforts to address the reproducibility crisis in neuroimaging. I took action to educate myself and peers through initiatives like co-leading the first Reproducibility Journal Club in Ireland. Throughout my PhD, I have also championed open science tools such as BIDS for standardised data organisation - I have contributed several BIDS tutorials and I have developed multiple BIDS apps to enhance reusability of the projects, including the first BIDS app to use DL. I have also created and shared educational materials on using GitHub for transparent, collaborative coding. These multifaceted efforts to promote open science led to invitations to author book chapters providing practical guidance on BIDS and GitHub for the wider neuroscience community, and to give a hands-on tutorial at the Analysing Large Neuroimaging Datasets Workshop at OHBM Glasgow, 2022.

In addition to these efforts, I actively pursued professional development by attending numerous specialised schools and programs including the ENERGHY social entrepreneurship program, where I helped develop a business model for aid distribution and honed project leadership abilities; the OxML summer school to stay updated on the latest ML and DL methods; the RYLA leadership program to build management and communication competencies; the ECNP neuropsychiatry program to learn about an emergent research area; the ARAPI, an Autism research conference, to connect with diverse stakeholders; and NeuroHackademy 2023, to expand data science skills in Neuroscience. These training programs helped me to grow my knowledge and skills, have provided invaluable networking opportunities, and have renewed my commitment to lifelong knowledge growth.

8.2. Interpretations

Several key hypotheses motivated the work on developing predictive models for ASD using neuroimaging data and DL. Based on the context exposed in **Chapter 1**, a first hypothesis was that structural MRI data alone contains sufficient precision to build a predictive model of ASD. A second hypothesis was that functional MRI data alone also holds adequate specificity for building an ASD prediction model. A third hypothesis was that the brain is a relevant variable for studying ASD, with diagnostic neuroimaging biomarkers detectable through ML. A fourth hypothesis was that while autistic individuals differ in terms of the severity of their symptoms, they share common characteristic patterns in the brain that can be captured by models. A final hypothesis was that current neuroimaging data variability is sufficient to train a robust ASD prediction algorithm that is generalisable to unseen individuals. These five key hypotheses were tested through studies using structural and functional MRI datasets with multiple predictive modelling architectures. The results provided insight into which hypotheses were supported and which should be reconsidered, and suggested additional experiments to further evaluate the potential for brain-based ASD prediction.

The first study demonstrated that DL can rapidly automate MRI quality control - a crucial preprocessing step, which, when performed manually, is somewhat subjective, repetitive, and time-consuming. My attention-based BrainQCNet model achieved excellent global performance for artefact detection, demonstrating how DL can augment human annotators for simple but tedious neuroimaging tasks. Interpretation remains challenging, however - while overall accuracy was high, local model behaviour showed both realistic and unrealistic patterns, highlighting the difficulty of explaining complex Artificial Neural Networks. Additional optimisation and experiments are needed to improve local-level accuracy and model understanding. Nevertheless, initial results were promising - the ample training data yielded a fairly robust global classifier, though broader artefact diversity could further enhance generalisability and reduce False Negatives. On the whole, this study established DL, and attention-based architectures in particular, as a viable approach to the automation of certain MRI preprocessing steps, paving the way for larger investments to tackle more complex analyses.

This argument is reinforced by the fact that other studies showed DL is a useful tool for automating preprocessing pipelines (Isensee et al., 2019; Tanno et al., 2017; Zhang et al., 2017). For instance, Zhang et al. (2017) showed that DL can be used to perform noise reduction, which is a critical preprocessing step in MRI. Tanno et al. (2017) proposed the use of CNNs for super-resolution in diffusion MRI (dMRI), automating the enhancement of MRI resolution. Isensee et al. (2019) described a DL approach for automated brain extraction, which is a fundamental step in many MRI-based studies. Taken together with my work, these studies provide support for the idea that DL is not only a viable approach for automating MRI preprocessing steps but is already achieving state-of-the-art results in various tasks. Moving forward, pairing strong global performance with granular model interpretability remains a key challenge that must be addressed as I scale up DL for enhanced neuroimaging workflows.

While showing promise, the predictive modelling studies in this thesis revealed limits on the efficacy of structural and functional MRI alone for robust Autism detection. Despite finding some consistent regional patterns that are aligned with Autism neuroscience (e.g., left-hemisphere language regions contributed to sMRI-based predictions), overall performance across multiple algorithms was modest, with classification accuracy ranging from 50-70%. There was also significant variability associated with factors such as age, gender, and site. Globally, studies with a comparable number of participants and using an external independent testing dataset have yielded relatively better performance, with accuracies ranging from 65-79% (Heinsfield et al., 2018; Kan et al., 2022; Traut et al., 2021). Rafiee et al. (2022) mentions that Heinsfield et al. (2018) achieved a classification accuracy of 70% implementing a DL algorithm on rs-fMRI data of 505 ASD and 530 TD subjects. Their DNN 70% accuracy was higher than the calculated accuracy for random forest (0.63) or SVM (0.65) methods. Conversely, Traut et al. (2021) worked with >2000 participants and achieved the best AUROC scores with models using traditional ML over DL algorithms. Traut et al. (2021) also showed that the performance of the best model dropped on an external independent dataset. This variation highlights the challenges inherent in MRI-based classification. To be clinically useful, a prediction tool must be robust and reproducible, and prediction accuracy must generalise across independent datasets. It is crucial that future work elucidates how accuracy varies based on the methodologies, algorithms, and datasets used.

The results of this thesis and reported in the wider literature underscore the diversity of neural phenotypes in Autism and suggest that current neuroimaging biomarkers lack the specificity and precision to generalise broadly, especially amidst highly prevalent comorbid diagnoses. The field is grappling with these challenges. In their review, Ecker et al. (2015) emphasised the heterogeneity of neuroimaging findings in Autism and the need for a more nuanced understanding that captures the diversity of neural phenotypes. Similarly, Haar et al. (2016) illustrated the considerable variability in anatomical findings across different MRI studies of autism, raising questions about the reliability and specificity of potential biomarkers. Finally, Kushki et al. (2019) pointed out the high prevalence of psychiatric comorbidity in autism, suggesting that these comorbidities could confound neuroimaging findings. While this thesis does not offer a definitive answer to the challenge of heterogeneity, it does show that there are opportunities for progress. For example, the work presented in **Chapters 4 and 5** shows that granular analysis of predictive features suggested the presence of shared neural signatures within this heterogeneity. Future work will pursue these clues, using larger datasets with greater representation of both ASD and comorbidities.

Critically, each model architecture yielded expectedly different results, indicating the importance of multi-model convergence to mitigate individual algorithm biases. In isolation, a single model's biases can dominate but combining diverse architectures can reveal more robust generalisable biomarkers. Goodfellow et al. (2016), a foundational book on DL, describes the benefits of ensemble methods in various sections. Long before, Hansen & Salamon (1990) discussed the benefits of using ensembles of ANN to improve generalisation performance, and (Perrone & Cooper, 1993) focused on how ensemble methods can help in situations when individual neural networks provide conflicting outputs. By leveraging the combined strengths and mitigating the individual weaknesses of multiple models, ensemble methods can indeed provide more robust and accurate DL predictions.

8.3. Implications of the PhD project

Alongside the core modelling projects, this thesis work strongly embraced open science practices to maximise research impact. Extensive self-directed training in programming, information technology tools, and artificial intelligence cultivated expertise applicable across studies. Participation in AI and Health summer schools also enabled honing techniques in responsible and interpretable DL.

Significant time and effort was dedicated to developing the BIDS-apps the implement my models on GPU systems using CUDA/CuDNN. This entailed overcoming several substantial technical hurdles to make the shared code fully reproducible. All code has been made openly available on GitHub to facilitate adoption. In addition, I created video tutorials on use of BIDS, Git, and GitHub to lower entry barriers so that more neuroscientists from diverse backgrounds and with limited resources can leverage these open science platforms.

Overall, this thesis demonstrated the feasibility of interpreting DL models and building ethical, responsible AI algorithms aligned with community needs. The integration of open science principles follows FAIR data stewardship, enhancing discovery and collaboration. This multifaceted approach combining methodological advances with openness and ethics showcases how to translate neuroimaging AI to benefit the Autism community. The long-term impact stems not just from novel techniques but also the commitments to openness, outreach, and conscientious application.

8.4. Limitations of the PhD project

My work established a novel methodological foundation but highlights significant remaining challenges for prediction of Autism diagnosis from MRI data.

8.4.1. On interpreting and explaining DL models

While **Chapter 4** developed an initial pipeline for interpretable DL - through the identification of the brain regions that contributed most to prediction - my work revealed significant lingering challenges to the evaluation and quantification of uncertainty within DL models. The high complexity of modern ANN often renders their inner workings opaque and decision-making inscrutable, even for developers. This "black box" nature makes quantitative analysis of algorithm behaviour and predictions difficult. Measuring feature importance and relating model components to underlying mechanisms remains an open pursuit in AI research. Methods like LIME, Shapley values, and integrated gradients help “peek inside” ANN (Ribeiro et al., 2016; Lundberg & Lee, 2017; Sundararajan et al., 2017), but currently lack scalability and standardisation.

Equally crucial is the quantification of uncertainty - conveying when predictions may be unreliable. **Chapter 5**’s extensive cross-validation enabled better estimation of accuracy and enabled the computation of confidence intervals around model performance. Meinke & Hein (2020) shows the importance of uncertainty in DL models, particularly in domains like healthcare, where a wrong decision can have dire consequences. One can imagine that quantification of uncertainty may be of even greater concern for psychiatric diagnoses, which lack specific known biological bases. Systematic and granular uncertainty quantification via Bayesian DL, ensembling, conformal prediction, and related techniques

(Gal & Ghahramani, 2016; Angelopoulos & Bates, 2021) is essential for clinical translation. Both robust evaluation and uncertainty measurement will be critical to developing trustworthy AI systems ready for deployment in medical settings where reliability and transparency are paramount.

By highlighting current gaps in practices, this thesis motivates and informs future work not just to advance the predictive performance of DL in neuroscience but crucially, to also boost model transparency, accountability, and the probabilistic understanding of limitations. Tackling these multifaceted open problems will require cross-disciplinary collaboration, but promises to accelerate responsible translation of AI innovations to improve patient outcomes.

8.4.2. On preprocessing pipelines

My work reveals the complex double-edged impacts of neuroimaging preprocessing on downstream DL analysis. **Chapter 3** demonstrated the value of rigorous quality control by developing a model to accelerate MRI artefact detection. Common preprocessing steps like spatial normalisation, smoothing, and registration make assumptions about typical anatomy that risk distorting or obscuring subtle morphological features associated with ASD. **Chapter 4** use of raw structural MRI data as input avoided such pitfalls. On the other hand, this “minimally processed” data approach subsequently limited the biological interpretability of the learned features and biomarkers driving model predictions, since there was no shared anatomical context.

This illustrates the inherent trade-offs between preserving naturalistic, unbiased brain signatures in the data and gaining specific anatomical meaning needed to aggregate findings across participants and relate these to clinical traits and neuroscientific knowledge. While DL thrives on extracting signals directly from minimally processed data, further work on relating the discovered patterns to tangible biological insights of clinical utility remains essential to practical adoption. Indeed, Topol (2019) discusses the convergence of AI and human intelligence in Medicine and highlights the importance of interpretability and clinical relevance.

The monomodal focus of my studies also constrained the full elucidation of the predictive features and biomarkers detected by models. Multimodal integration of neuroimaging with genetics, cognitive tests, and clinical assessments appears critical for grounding DL models in biological mechanisms relevant to heterogeneous neurodivergences like Autism. Other studies have pointed out the importance of multimodal data. For example, Parikshak et al. (2013) showed the benefits of integrating genomics with functional data to elucidate the pathways and circuits implicated in autism. In a review, Sui et al. (2012) also emphasised the importance of multimodal data fusion, especially in capturing more complex, high-dimensional representations of the brain.

Overall, my work strongly motivates future research into tailored, lightweight preprocessing and fusion techniques that balance performance, interpretability, and scientific value for studying complex neurological conditions. Developing such optimised pipelines will require cross-disciplinary collaboration and community feedback to enable AI that intersects with, rather than diverges from, human-driven neuroscience.

8.4.3. On dataset biases

My work reveals considerable risks of bias amplification and skewed representation in current neuroimaging datasets that DL algorithms could potentially exacerbate. Crossing MRI data with genotypic information and stratifying analyses by genetic clusters is essential to efforts to build equitable models tuned to diverse populations, rather than overrepresented subgroups. However, despite current huge efforts by the community to build large datasets as seen in **Chapter 1**, most of these like ABIDE contain limited, if any, genetic data. The available samples also have potentially problematic demographic compositions—for example, in ABIDE datasets, a preponderance of younger male subjects risks models tuned only to this group. In **Chapter 4**, more robust replicable important brain regions were found for boys than for girls. While this demographic composition may be consistent with wider prevalence as introduced in **Chapter 1**, sensitive and specific models will require better representation of females and Autistic individuals of different ages. While multi-site data pooling has enabled larger samples, variability in MRI data acquisition protocols across sites can further confound analysis. This phenomenon was particularly observed in **Chapter 4**.

More concerning is the limited characterisation of clinical, behavioural, and phenotypic traits alongside neural data. Details on medications, comorbid conditions, symptom profiles, and other variables are often unreported, yet crucial for relating brain patterns to real-world functioning. Such issues likely stem in part from the challenges of scanning people with neurodevelopmental differences, where success can depend heavily on individual factors. Those able to tolerate MRI may represent a narrow subset. As ASD is an evolving, lifelong neurodivergence, studying static data slices in isolation also risks overlooking crucial developmental trajectories.

Ultimately, creating more balanced, representative datasets will require active involvement of autistic participants and advocates in protocol co-design (Heraty et al., 2023). This could fruitfully involve capturing multidimensional data across modalities and timepoints to better encapsulate heterogeneity. Sensitive accommodation of individual needs and preferences would enable inclusion of a broader population. Indeed, neuroimaging may not be ideal or feasible for many. Pursuing such inclusive, ethically-obtained data will allow DL to complement today's small homogeneous samples with equitable insights benefitting the full community.

Overall, my work strongly motivates interweaving cutting-edge modelling with participatory data improvements to ensure neuroimaging AI meaningfully serves diverse populations.

8.4.4. On deep learning

Deep learning's explosive growth in model complexity introduces new challenges in robust training and generalisation. While my datasets were large compared to many existing studies with Autistic participants, several models in **Chapters 4 and 5** still demonstrated overfitting - suggesting that the data contained insufficient diversity to capture heterogeneous neurological divergences. Aligned with my work, multiple recent studies have estimated that sample sizes well into the tens or hundreds of thousands are necessary to reliably train deep neural networks for ASD detection without overfitting (Jiao et al., 2021; Haar et al., 2022). Moreover, the combinatorial breadth of possible configurations across network architecture, hyperparameters, and optimization techniques leads to a vast tuning space. However, exhaustive tuning risks simply overfitting to idiosyncrasies of limited datasets rather than learning generalisable and replicable patterns that transfer robustly to new out-of-sample cases.

My work thus underscores the pressing need for larger, more varied Autism imaging datasets, alongside careful methodology to develop reliable DL biomarkers ready for real-world deployment. Assembling appropriately large and representative training data will require collaboration across multiple research centres and clinics. Crucially, active involvement of autistic community members in data collection and protocol design will help capture the diversity of the spectrum (Heraty et al., 2023). Complementing big data advances with rigorous cross-validation, regularisation, uncertainty quantification, and related techniques will also be key to combating overfitting given the intrinsic complexity of deep nets. Novel unsupervised learning methods could also be beneficial for harnessing the large amounts of unlabeled data (Zaadnoordijk et al., 2022). Guided by both human-centred and technical best practices, DL holds immense potential to uncover reproducible neuroimaging patterns that provide clinically useful insights into heterogeneous neurodivergences like Autism.

Overall, my studies underscored a number of challenges and limitations, but, in doing so, also outlined a research program for progress through bigger, more varied dataset creation, integrated predictive modelling, and grounding in behavioural dimensions.

Though significant challenges remain, the work of this PhD thesis provides both a strong motivational foundation and methodological strategy for the pursuit of biologically-grounded DL as a means to elucidate Autism's complex neural correlates in a clinically meaningful way.

8.5. Recommendations

The limitations discussed in the preceding section prompt several recommendations for future work to advance biologically-grounded AI modelling of ASD using neuroimaging data. I suggest that future studies should:

- Integrate multimodal data beyond MRI, including genetics, cognition, and clinical assessments, to enhance biological interpretation. Fusing neuroimaging with other biological and phenotypic data can help provide learned patterns with tangible clinical significance.
- Contextualise studies with more specific inclusion criteria if dataset size is limited. Focusing on targeted demographic or behavioural factors can reduce heterogeneity and improve characterisation of neural correlates within a defined ASD context.
- Test diverse DL model architectures for any predictive modelling task. Varying approaches mitigates individual algorithm biases and enables convergence on the most robust generalisable patterns.
- Employ statistical methods such as N-fold cross-validation frameworks to rigorously evaluate model performance and uncertainty. However, model exploration time should be balanced with the number of experiments feasible.
- Explore longitudinal data to extract intra-individual patterns over time, alongside inter-individual differences. Modelling developmental changes may reveal key neural trajectories.
- Build interpretation pipelines to explain model reasoning and relate features to neuroscientific mechanisms. Explainable AI is essential for clinical utility and adoption.
- Continually update skills in AI, programming, neuroscience, and psychiatry. All these disciplines are rapidly advancing, requiring lifelong learning to apply them effectively in multidisciplinary research.
- Maintain openness and reproducibility in science. As research rapidly advances in the fields of psychiatry, neuroimaging, and AI, it is crucial to uphold these principles to foster work that is both impactful and reliable.

- Carefully evaluate the ethical implications of AI techniques prior to application in Autism research or care. Ensure models are transparent, fair, reproducible, and designed to safely complement clinicians rather than replace them.

Adhering to these recommendations can promote development of more reliable, interpretable, and clinically useful AI models of ASD using brain imaging and related data. By attending to ethical considerations alongside methodological advances, research should lead to more responsible AI to benefit the Autism community.

8.6. Conclusion

This thesis presented pioneering explorations into developing interpretable deep learning frameworks for elucidating neuroimaging biomarkers and patterns associated with Autism Spectrum Disorder. Through three complementary projects analysing structural and functional MRI data, methods were developed and tested against specific hypotheses related to the viability of brain imaging and DL for prediction of Autism diagnosis in the presence of clinical and phenotypic heterogeneity.

While falling short of robust prediction and subject to limitations, these projects highlighted pathways forward - through integrating diverse data modalities, improving model optimization and evaluation, and applying DL in synergy with neuroscience domain knowledge. Additional open science contributions provided reusable research tools and demonstrated commitments to ethics and rigour.

Overall, this research established a strong motivational foundation and methodological strategy for biologically-grounded AI modelling to quantify and interpret complex neural correlates of Autism traits. The limitations identified motivate a set of recommendations for future work aimed at overcoming current challenges in explainable and equitable neuroimaging analysis. By laying this groundwork and direction for the field, this thesis provides a springboard for future efforts to refine data-driven imaging biomarkers that can translate to enhanced clinical insights and precision care for autistic individuals.

REFERENCES

Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., Litke, S., O'Hagan, B., Andersen, J., Bronstein, B., Bui, A., Bushey, M., Butler, H., Castagna, V., Camacho, N., … Milham, M. P. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, *4*(1), 170181. https://doi.org/10.1038/sdata.2017.181

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., … Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, *166*, 400–424. https://doi.org/10.1016/j.neuroimage.2017.10.034

Allsopp, K., Read, J., Corcoran, R., & Kinderman, P. (2019). Heterogeneity in psychiatric diagnostic classification. *Psychiatry Research*, *279*, 15–22. https://doi.org/10.1016/j.psychres.2019.07.005

Ameis, S. H., & Catani, M. (2015). Altered white matter connectivity as a neural substrate for social impairment in Autism Spectrum Disorder. *Cortex*, *62*, 158–181. https://doi.org/10.1016/j.cortex.2014.10.014

Andrews, D. S., Avino, T. A., Gudbrandsen, M., Daly, E., Marquand, A., Murphy, C. M., Lai, M.-C., Lombardo, M. V., Ruigrok, A. N. V., Williams, S. C., Bullmore, E. T., The MRC AIMS Consortium, Suckling, J., Baron-Cohen, S., Craig, M. C., Murphy, D. G. M., & Ecker, C. (2017). In Vivo Evidence of Reduced Integrity of the Gray–White Matter Boundary in Autism Spectrum Disorder. *Cerebral Cortex (New York, NY)*, *27*(2), 877–887. https://doi.org/10.1093/cercor/bhw404

APA, Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). (2013). American Psychiatric Association Publishing.

Arya, D., Olij, R., Gupta, D. K., Gazzar, A. E., Wingen, G., Worring, M., & Thomas, R. M. (2020). Fusing Structural and Functional MRIs using Graph Convolutional Networks for

Autism Classification. *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, 44–61. https://proceedings.mlr.press/v121/arya20a.html

Backhausen, L. L., Herting, M. M., Buse, J., Roessner, V., Smolka, M. N., & Vetter, N. C. (2016). Quality Control of Structural MRI Images Applied Using FreeSurfer—A Hands-On Workflow to Rate Motion Artifacts. *Frontiers in Neuroscience*, *10*. https://doi.org/10.3389/fnins.2016.00558

Baker, J. T., Dillon, D. G., Patrick, L. M., Roffman, J. L., Brady, R. O., Pizzagalli, D. A., Öngür, D., & Holmes, A. J. (2019). Functional connectomics of affective and psychotic pathology. *Proceedings of the National Academy of Sciences*, *116*(18), 9050–9059. https://doi.org/10.1073/pnas.1820780116

Baron-Cohen, S. (2002). The extreme male brain theory of autism.

Bedel, H., Şıvgın, I., Dalmaz, O., Dar, S. U. H., & Cukur, T. (2022). *BolT: Fused Window Transformers for fMRI Time Series Analysis*. https://doi.org/10.48550/arXiv.2205.11578

Bedford, S. A., Park, M. T. M., Devenyi, G. A., Tullo, S., Germann, J., Patel, R., Anagnostou, E., Baron-Cohen, S., Bullmore, E. T., Chura, L. R., Craig, M. C., Ecker, C., Floris, D. L., Holt, R. J., Lenroot, R., Lerch, J. P., Lombardo, M. V., Murphy, D. G. M., Raznahan, A., … Chakravarty, M. M. (2020). Large-scale analyses of the relationship between sex, age and intelligence quotient heterogeneity and cortical morphometry in autism spectrum disorder. *Molecular Psychiatry*, *25*(3), 614–628. https://doi.org/10.1038/s41380-019-0420-6

Beggiato, A., Peyre, H., Maruani, A., Scheid, I., Rastam, M., Amsellem, F., Gillberg, C. I., Leboyer, M., Bourgeron, T., Gillberg, C., & Delorme, R. (2017). Gender differences in autism spectrum disorders: Divergence among specific core symptoms: Gender differences in ASD. *Autism Research*, *10*(4), 680–689. https://doi.org/10.1002/aur.1715

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S., & Craddock, R. C. (2017a). The Neuro Bureau ADHD-200 Preprocessed repository. *NeuroImage*, *144*, 275–286. https://doi.org/10.1016/j.neuroimage.2016.06.034

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S., & Craddock, R. C. (2017b). The Neuro Bureau ADHD-200 Preprocessed repository. *NeuroImage*, *144*, 275–286. https://doi.org/10.1016/j.neuroimage.2016.06.034

Bengs, M., Gessert, N., & Schlaefer, A. (2020). 4D Spatio-Temporal Deep Learning with 4D fMRI Data for Autism Spectrum Disorder Classification.

Benkarim, O., Paquola, C., Park, B., Kebets, V., Hong, S.-J., Wael, R. V. de, Zhang, S., Yeo, B. T. T., Eickenberg, M., Ge, T., Poline, J.-B., Bernhardt, B. C., & Bzdok, D. (2022). Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. *PLOS Biology*, *20*(4), e3001627. https://doi.org/10.1371/journal.pbio.3001627

Berdejo-Espinola, V., & Amano, T. (2023). AI tools can improve equity in science. *Science*, *379*(6636), 991–991. https://doi.org/10.1126/science.adg9714

Bird, N., & Flint, R. (2019). *Autism: People face 'daily discrimination' in work—BBC News*. https://www.bbc.com/news/uk-wales-49523283

Bottou, L., Curtis, F. E., & Nocedal, J. (2018). *Optimization Methods for Large-Scale Machine Learning* (arXiv:1606.04838). arXiv. http://arxiv.org/abs/1606.04838

Bourgeron, T. (2015). From the genetic architecture to synaptic plasticity in autism spectrum disorder. *Nature Reviews Neuroscience*, *16*(9), 551–563. https://doi.org/10.1038/nrn3992

Bourgeron, T. (2023, January). *Des gènes, des synapses, des autismes | Éditions Odile Jacob*. https://www.odilejacob.fr/catalogue/sciences/genetique/des-genes-des-synapses-des-autismes_9782415003906.php

Buescher, A. V. S., Cidav, Z., Knapp, M., & Mandell, D. S. (2014). Costs of Autism Spectrum Disorders in the United Kingdom and the United States. *JAMA Pediatrics*, *168*(8), 721. https://doi.org/10.1001/jamapediatrics.2014.210

Canario, E., Chen, D., & Biswal, B. (2021). A review of resting-state fMRI and its use to examine psychiatric disorders. *Psychoradiology*, *1*(1), 42–53. https://doi.org/10.1093/psyrad/kkab003

Cardoso, M. J., Modat, M., Wolz, R., Melbourne, A., Cash, D., Rueckert, D., & Ourselin, S. (2015). Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion. *IEEE Transactions on Medical Imaging*, *34*(9), 1976–1988. https://doi.org/10.1109/TMI.2015.2418298

Carper, R. A., Solders, S., Treiber, J. M., Fishman, I., & Müller, R.-A. (2015). Corticospinal Tract Anatomy and Functional Connectivity of Primary Motor Cortex in Autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, *54*(10), 859–867. https://doi.org/10.1016/j.jaac.2015.07.007

Cazalis, F. (2017, July 6). *Ces femmes autistes qui s'ignorent*. The Conversation. http://theconversation.com/ces-femmes-autistes-qui-signorent-75998

Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., & Rudin, C. (2019). This Looks Like That: Deep Learning for Interpretable Image Recognition. *arXiv:1806.10574 [Cs, Stat]*. http://arxiv.org/abs/1806.10574

Chen, S., Ma, K., & Zheng, Y. (2019). Med3D: Transfer Learning for 3D Medical Image Analysis. *arXiv:1904.00625 [Cs]*. http://arxiv.org/abs/1904.00625

Chien, Y.-L., Chen, Y.-C., Chiu, Y.-N., Tsai, W.-C., & Gau, S. S.-F. (2021). A translational exploration of the effects of WNT2 variants on altered cortical structures in autism spectrum disorder. *Journal of Psychiatry & Neuroscience : JPN*, *46*(6), E647–E658. https://doi.org/10.1503/jpn.210022

Christensen, D. L., Maenner, M. J., Bilder, D., Constantino, J. N., Daniels, J., Durkin, M. S., Fitzgerald, R. T., Kurzius-Spencer, M., Pettygrove, S. D., Robinson, C., Shenouda, J., White, T., Zahorodny, W., Pazol, K., & Dietz, P. (2019). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 4 Years—Early Autism and Developmental Disabilities Monitoring Network, Seven Sites, United States, 2010, 2012, and 2014. *MMWR. Surveillance Summaries*, *68*(2), 1–19. https://doi.org/10.15585/mmwr.ss6802a1

Cieslak, M., Cook, P. A., He, X., Yeh, F.-C., Dhollander, T., Adebimpe, A., Aguirre, G. K., Bassett, D. S., Betzel, R. F., Bourque, J., Cabral, L. M., Davatzikos, C., Detre, J. A., Earl, E., Elliott, M. A., Fadnavis, S., Fair, D. A., Foran, W., Fotiadis, P., … Satterthwaite, T. D. (2021).

QSIPrep: An integrative platform for preprocessing and reconstructing diffusion MRI data. *Nature Methods*, *18*(7), Article 7. https://doi.org/10.1038/s41592-021-01185-5

Clark, D. B., Fisher, C. B., Bookheimer, S., Brown, S. A., Evans, J. H., Hopfer, C., Hudziak, J., Montoya, I., Murray, M., Pfefferbaum, A., & Yurgelun-Todd, D. (2018). Biomedical ethics and clinical oversight in multisite observational neuroimaging studies with children and adolescents: The ABCD experience. *Developmental Cognitive Neuroscience*, *32*, 143–154. https://doi.org/10.1016/j.dcn.2017.06.005

Clark, M. L. E., Vinen, Z., Barbaro, J., & Dissanayake, C. (2018). School Age Outcomes of Children Diagnosed Early and Later with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *48*(1), 92–102. https://doi.org/10.1007/s10803-017-3279-x

Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., & Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, *33*(8), 1914–1928. https://doi.org/10.1002/hbm.21333

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., & Varoquaux, G. (2019). Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage*, *192*, 115–134. https://doi.org/10.1016/j.neuroimage.2019.02.062

Dawson, G., & Burner, K. (2011). Behavioral interventions in children and adolescents with autism spectrum disorder: A review of recent findings. *Current Opinion in Pediatrics*, *23*(6), 616–620. https://doi.org/10.1097/MOP.0b013e32834cf082

Dean, M., Harwood, R., & Kasari, C. (2017). The art of camouflage: Gender differences in the social behaviors of girls and boys with autism spectrum disorder. *Autism*, *21*(6), 678–689. https://doi.org/10.1177/1362361316671845

Dekhil, O., Ali, M., Haweel, R., Elnakib, Y., Ghazal, M., Hajjdiab, H., Fraiwan, L., Shalaby, A., Soliman, A., Mahmoud, A., Keynton, R., Casanova, M. F., Barnes, G., & El-Baz, A. (2020). A Comprehensive Framework for Differentiating Autism Spectrum Disorder From Neurotypicals by Fusing Structural MRI and Resting State Functional MRI. *Seminars in Pediatric Neurology*, *34*, 100805. https://doi.org/10.1016/j.spen.2020.100805

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L. M. E., Bookheimer, S. Y., Braden, B. B., Byrge, L., Castellanos, F. X., Dapretto, M., Delorme, R., Fair, D. A., Fishman, I., … Milham, M. P. (2017a). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, *4*(1), 170010. https://doi.org/10.1038/sdata.2017.10

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L. M. E., Bookheimer, S. Y., Braden, B. B., Byrge, L., Castellanos, F. X., Dapretto, M., Delorme, R., Fair, D. A., Fishman, I., … Milham, M. P. (2017b). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, *4*(1), Article 1. https://doi.org/10.1038/sdata.2017.10

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., … Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, *19*(6), Article 6. https://doi.org/10.1038/mp.2013.78

Dickie, E. W., Ameis, S. H., Shahab, S., Calarco, N., Smith, D. E., Miranda, D., Viviano, J. D., & Voineskos, A. N. (2018). Personalized Intrinsic Network Topography Mapping and Functional Connectivity Deficits in Autism Spectrum Disorder. *Biological Psychiatry*, *84*(4), 278–286. https://doi.org/10.1016/j.biopsych.2018.02.1174

Dvornek, N. C., Ventola, P., Pelphrey, K. A., & Duncan, J. S. (2017). Identifying Autism from Resting-State fMRI Using Long Short-Term Memory Networks. *Machine Learning in Medical Imaging. MLMI (Workshop)*, *10541*, 362–370. https://doi.org/10.1007/978-3-319-67389-9_42

Ecker, C., Bookheimer, S. Y., & Murphy, D. G. M. (2015). Neuroimaging in autism spectrum disorder: Brain structure and function across the lifespan. *The Lancet Neurology*, *14*(11), 1121–1134. https://doi.org/10.1016/S1474-4422(15)00050-2

Elibol, H. M., Nguyen, V., Linderman, S., Johnson, M., Hashmi, A., & Doshi-Velez, F. (2016). Cross-Corpora Unsupervised Learning of Trajectories in Autism Spectrum Disorders. *Journal of Machine Learning Research*, *17*(133), 1–38.

Emerson, R. W., Adams, C., Nishino, T., Hazlett, H. C., Wolff, J. J., Zwaigenbaum, L., Constantino, J. N., Shen, M. D., Swanson, M. R., Elison, J. T., Kandala, S., Estes, A. M., Botteron, K. N., Collins, L., Dager, S. R., Evans, A. C., Gerig, G., Gu, H., McKinstry, R. C., … Piven, J. (2017). Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. *Science Translational Medicine*, *9*(393), eaag2882. https://doi.org/10.1126/scitranslmed.aag2882

Eslami, T., Mirjalili, V., Fong, A., Laird, A. R., & Saeed, F. (2019). ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data. *Frontiers in Neuroinformatics*, *13*. https://www.frontiersin.org/articles/10.3389/fninf.2019.00070

Eslami, T., Raiker, J. S., & Saeed, F. (2021). Chapter 4—Explainable and scalable machine learning algorithms for detection of autism spectrum disorder using fMRI data. In A. S. El-Baz & J. S. Suri (Eds.), *Neural Engineering Techniques for Autism Spectrum Disorder* (pp. 39–54). Academic Press. https://doi.org/10.1016/B978-0-12-822822-7.00004-1

Eslami, T., & Saeed, F. (2019). Auto-ASD-Network: A Technique Based on Deep Learning and Support Vector Machines for Diagnosing Autism Spectrum Disorder using fMRI Data. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 646–651. https://doi.org/10.1145/3307339.3343482

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE*, *12*(9), e0184661. https://doi.org/10.1371/journal.pone.0184661

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*(11), 1664–1671. https://doi.org/10.1038/nn.4135

Fishman, I., Linke, A. C., Hau, J., Carper, R. A., & Müller, R.-A. (2018). Atypical Functional Connectivity of Amygdala Related to Reduced Symptom Severity in Children With Autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, *57*(10), 764-774.e3. https://doi.org/10.1016/j.jaac.2018.06.015

Floris, D. L., Lai, M.-C., Auer, T., Lombardo, M. V., Ecker, C., Chakrabarti, B., Wheelwright, S. J., Bullmore, E. T., Murphy, D. G. M., Baron-Cohen, S., & Suckling, J. (2016). Atypically rightward cerebral asymmetry in male adults with autism stratifies individuals with and without language delay. *Human Brain Mapping*, *37*(1), 230–253. https://doi.org/10.1002/hbm.23023

Fombonne, E. (2009). Epidemiology of Pervasive Developmental Disorders. *Pediatric Research*, *65*(6), 591–598. https://doi.org/10.1203/PDR.0b013e31819e7203

Garcia, M., Dosenbach, N., & Kelly, C. (2022). *BrainQCNet: A Deep Learning attention-based model for multi-scale detection of artifacts in brain structural MRI scans* (p. 2022.03.11.483983). bioRxiv. https://doi.org/10.1101/2022.03.11.483983

Ghaziuddin, M., Weidmer-Mikhail, E., & Ghaziuddin, N. (2002). Comorbidity of Asperger syndrome: A preliminary report. *Journal of Intellectual Disability Research*, *42*(4), 279–283. https://doi.org/10.1111/j.1365-2788.1998.tb01647.x

Gillberg, C. (2010). The ESSENCE in child psychiatry: Early Symptomatic Syndromes Eliciting Neurodevelopmental Clinical Examinations. *Research in Developmental Disabilities*, *31*(6), 1543–1551. https://doi.org/10.1016/j.ridd.2010.06.002

Gillberg, I. C., Helles, A., Billstedt, E., & Gillberg, C. (2016). Boys with Asperger Syndrome Grow Up: Psychiatric and Neurodevelopmental Disorders 20 Years After Initial Diagnosis.

Journal of Autism and Developmental Disorders, *46*(1), 74–82. https://doi.org/10.1007/s10803-015-2544-0

Gilmore, A., Buser, N., & Hanson, J. L. (2019). *Variations in Structural MRI Quality Significantly Impact Commonly-Used Measures of Brain Anatomy* [Preprint]. Neuroscience. https://doi.org/10.1101/581876

Girault, J. B., Donovan, K., Hawks, Z., Talovic, M., Forsen, E., Elison, J. T., ... & IBIS Network. (2022). Infant visual brain development and inherited genetic liability in autism. *American Journal of Psychiatry*, *179*(8), 573-585.

Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L. R., Auerbach, E. J., Behrens, T. E. J., Coalson, T. S., Harms, M. P., Jenkinson, M., Moeller, S., Robinson, E. C., Sotiropoulos, S. N., Xu, J., Yacoub, E., Ugurbil, K., & Van Essen, D. C. (2016). The Human Connectome Project's neuroimaging approach. *Nature Neuroscience*, *19*(9), 1175–1187. https://doi.org/10.1038/nn.4361

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. https://www.deeplearningbook.org/

Gorgolewski, K. J., Alfaro-Almagro, F., Auer, T., Bellec, P., Capotă, M., Chakravarty, M. M., Churchill, N. W., Cohen, A. L., Craddock, R. C., Devenyi, G. A., Eklund, A., Esteban, O., Flandin, G., Ghosh, S. S., Guntupalli, J. S., Jenkinson, M., Keshavan, A., Kiar, G., Liem, F., … Poldrack, R. A. (2017). BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLOS Computational Biology*, *13*(3), e1005209. https://doi.org/10.1371/journal.pcbi.1005209

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 160044. https://doi.org/10.1038/sdata.2016.44

Guthrie, W., Wallis, K., Bennett, A., Brooks, E., Dudley, J., Gerdes, M., Pandey, J., Levy, S. E., Schultz, R. T., & Miller, J. S. (2019). Accuracy of Autism Screening in a Large Pediatric Network. *Pediatrics*, *144*(4), e20183963. https://doi.org/10.1542/peds.2018-3963

Ha, S., Sohn, I.-J., Kim, N., Sim, H. J., & Cheon, K.-A. (2015). Characteristics of Brains in Autism Spectrum Disorder: Structure, Function and Connectivity across the Lifespan. *Experimental Neurobiology*, *24*(4), 273–284. https://doi.org/10.5607/en.2015.24.4.273

Haar, S., Berman, S., Behrmann, M., & Dinstein, I. (2016). Anatomical Abnormalities in Autism? *Cerebral Cortex (New York, N.Y.: 1991)*, *26*(4), 1440–1452. https://doi.org/10.1093/cercor/bhu242

Hagler, D. J., Hatton, SeanN., Cornejo, M. D., Makowski, C., Fair, D. A., Dick, A. S., Sutherland, M. T., Casey, B. J., Barch, D. M., Harms, M. P., Watts, R., Bjork, J. M., Garavan, H. P., Hilmer, L., Pung, C. J., Sicat, C. S., Kuperman, J., Bartsch, H., Xue, F., … Dale, A. M. (2019). Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage*, *202*, 116091. https://doi.org/10.1016/j.neuroimage.2019.116091

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. https://doi.org/10.1109/34.58871

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [Cs]*. http://arxiv.org/abs/1512.03385

He, Y., Byrge, L., & Kennedy, D. P. (2020). Nonreplication of functional connectivity differences in autism spectrum disorder across multiple sites and denoising strategies. *Human Brain Mapping*, *41*(5), 1334–1350. https://doi.org/10.1002/hbm.24879

Heasman, B. (2017, July 31). Employers may discriminate against autism without realising. *LSE Business Review*. https://blogs.lse.ac.uk/businessreview/2017/07/31/employers-may-discriminate-against-autism-without-realising/

Heasman, B., & Gillespie, A. (2018). Perspective-taking is two-sided: Misunderstandings between people with Asperger's syndrome and their family members. *Autism*, *22*(6), 740–750. https://doi.org/10.1177/1362361317708287

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2017). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage : Clinical*, *17*, 16–23. https://doi.org/10.1016/j.nicl.2017.08.017

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, *17*, 16–23. https://doi.org/10.1016/j.nicl.2017.08.017

Heraty, S., Lautarescu, A., Belton, D., Boyle, A., Cirrincione, P., Doherty, M., Douglas, S., Plas, J. R. D., Bosch, K. V. D., Violland, P., Tercon, J., Ruigrok, A., Murphy, D. G. M., Bourgeron, T., Chatham, C., Loth, E., Oakley, B., McAlonan, G. M., Charman, T., … Jones, E. J. H. (2023). Bridge-building between communities: Imagining the future of biomedical autism research. *Cell*, *186*(18), 3747–3752. https://doi.org/10.1016/j.cell.2023.08.004

Hong, S.-J., Valk, S. L., Di Martino, A., Milham, M. P., & Bernhardt, B. C. (2018). Multidimensional Neuroanatomical Subtyping of Autism Spectrum Disorder. *Cerebral Cortex (New York, N.Y.: 1991)*, *28*(10), 3578–3588. https://doi.org/10.1093/cercor/bhx229

Hong, S.-J., Vogelstein, J. T., Gozzi, A., Bernhardt, B. C., Yeo, B. T. T., Milham, M. P., & Di Martino, A. (2020). Toward Neurosubtypes in Autism. *Biological Psychiatry*, *88*(1), 111–128. https://doi.org/10.1016/j.biopsych.2020.03.022

Horien, C., Floris, D. L., Greene, A. S., Noble, S., Rolison, M., Tejavibulya, L., O'Connor, D., McPartland, J. C., Scheinost, D., Chawarska, K., Lake, E. M. R., & Constable, R. T. (2022). Functional Connectome–Based Predictive Modeling in Autism. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2022.04.008

Hu, J., Cao, L., Li, T., Liao, B., Dong, S., & Li, P. (2020). Interpretable Learning Approaches in Resting-State Functional Connectivity Analysis: The Case of Autism Spectrum Disorder. *Computational and Mathematical Methods in Medicine*, *2020*, 1–12. https://doi.org/10.1155/2020/1394830

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). *Densely Connected Convolutional Networks*. 4700–4708. https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). Densely Connected Convolutional Networks. *arXiv:1608.06993 [Cs]*. http://arxiv.org/abs/1608.06993

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K. H., & Kickingereder, P. (2019). Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, *40*(17), 4952–4964. https://doi.org/10.1002/hbm.24750

Iwana, B. K., & Uchida, S. (2021). An empirical survey of data augmentation for time series classification with neural networks. *PLOS ONE*, *16*(7), e0254841. https://doi.org/10.1371/journal.pone.0254841

Jiang, R., Calhoun, V. D., Zuo, N., Lin, D., Li, J., Fan, L., Qi, S., Sun, H., Fu, Z., Song, M., Jiang, T., & Sui, J. (2018a). Connectome-based individualized prediction of temperament trait scores. *NeuroImage*, *183*, 366–374. https://doi.org/10.1016/j.neuroimage.2018.08.038

Jiang, R., Calhoun, V. D., Zuo, N., Lin, D., Li, J., Fan, L., Qi, S., Sun, H., Fu, Z., Song, M., Jiang, T., & Sui, J. (2018b). Connectome-based individualized prediction of temperament trait scores. *NeuroImage*, *183*, 366–374. https://doi.org/10.1016/j.neuroimage.2018.08.038

Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., & Yang, C. (2022). *Brain Network Transformer* (arXiv:2210.06681). arXiv. https://doi.org/10.48550/arXiv.2210.06681

Karcher, N. R., & Barch, D. M. (2021). The ABCD study: Understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology*, *46*(1), 131–142. https://doi.org/10.1038/s41386-020-0736-6

Kelly, C., Uddin, L. Q., Shehzad, Z., Margulies, D. S., Castellanos, F. X., Milham, M. P., & Petrides, M. (2010). Broca's region: Linking human brain functional connectivity data and non-human primate tracing anatomy studies. *The European Journal of Neuroscience*, *32*(3), 383–398. https://doi.org/10.1111/j.1460-9568.2010.07279.x

Keshavan, A., Yeatman, J. D., & Rokem, A. (2019). Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging. *Frontiers in Neuroinformatics*, *13*, 29. https://doi.org/10.3389/fninf.2019.00029

Khosla, M., Jamison, K., Kuceyeski, A., & Sabuncu, M. R. (2018). 3D Convolutional Neural Networks for Classification of Functional Connectomes. In D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. S. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, & A. Madabhushi (Eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 137–145). Springer International Publishing. https://doi.org/10.1007/978-3-030-00889-5_16

Khosla, M., Jamison, K., Kuceyeski, A., & Sabuncu, M. R. (2019). Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. *NeuroImage*, *199*, 651–662. https://doi.org/10.1016/j.neuroimage.2019.06.012

King, J. B., Prigge, M. B. D., King, C. K., Morgan, J., Weathersby, F., Fox, J. C., Dean, D. C., Freeman, A., Villaruz, J. A. M., Kane, K. L., Bigler, E. D., Alexander, A. L., Lange, N., Zielinski, B., Lainhart, J. E., & Anderson, J. S. (2019). Generalizability and reproducibility of functional connectivity in autism. *Molecular Autism*, *10*(1), 27. https://doi.org/10.1186/s13229-019-0273-5

Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization* (arXiv:1412.6980). arXiv. https://doi.org/10.48550/arXiv.1412.6980

Kirkovski, M., Enticott, P. G., & Fitzgerald, P. B. (2013). A Review of the Role of Female Gender in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *43*(11), 2584–2603. https://doi.org/10.1007/s10803-013-1811-1

Kishida, K. T., De Asis-Cruz, J., Treadwell-Deering, D., Liebenow, B., Beauchamp, M. S., & Montague, P. R. (2019). Diminished single-stimulus response in vmPFC to favorite people in children diagnosed with Autism Spectrum Disorder. *Biological Psychology*, *145*, 174–184. https://doi.org/10.1016/j.biopsycho.2019.04.009

Kohli, J. S., Kinnear, M. K., Fong, C. H., Fishman, I., Carper, R. A., & Müller, R.-A. (2019). Local Cortical Gyrification is Increased in Children With Autism Spectrum Disorders, but

Decreases Rapidly in Adolescents. *Cerebral Cortex (New York, N.Y.: 1991)*, *29*(6), 2412–2423. https://doi.org/10.1093/cercor/bhy111

Kunda, M., Zhou, S., Gong, G., & Lu, H. (2023). Improving Multi-Site Autism Classification via Site-Dependence Minimization and Second-Order Functional Connectivity. *IEEE Transactions on Medical Imaging*, *42*(1), 55–65. https://doi.org/10.1109/TMI.2022.3203899

Kushki, A., Anagnostou, E., Hammill, C., Duez, P., Brian, J., Iaboni, A., Schachar, R., Crosbie, J., Arnold, P., & Lerch, J. P. (2019). Examining overlap and homogeneity in ASD, ADHD, and OCD: A data-driven, diagnosis-agnostic approach. *Translational Psychiatry*, *9*(1), 318. https://doi.org/10.1038/s41398-019-0631-2

Laird, A. R., Fox, P. M., Eickhoff, S. B., Turner, J. A., Ray, K. L., McKay, D. R., Glahn, D. C., Beckmann, C. F., Smith, S. M., & Fox, P. T. (2011). Behavioral interpretations of intrinsic connectivity networks. *Journal of Cognitive Neuroscience*, *23*(12), 4022–4037. https://doi.org/10.1162/jocn_a_00077

Lake, E. M. R., Finn, E. S., Noble, S. M., Vanderwal, T., Shen, X., Rosenberg, M. D., Spann, M. N., Chun, M. M., Scheinost, D., & Constable, R. T. (2019). The Functional Brain Organization of an Individual Allows Prediction of Measures of Social Abilities Transdiagnostically in Autism and Attention-Deficit/Hyperactivity Disorder. *Biological Psychiatry*, *86*(4), 315–326. https://doi.org/10.1016/j.biopsych.2019.02.019

Lam, P., Zhu, A. H., Gari, I. B., Jahanshad, N., & Thompson, P. M. (2020). 3D Grid-Attention Networks for Interpretable Age and Alzheimer’s Disease Prediction from Structural MRI. *arXiv:2011.09115 [Eess, q-Bio]*. http://arxiv.org/abs/2011.09115

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

LeCun, Y., Haffner, P., Bottou, L., & Bengio, Y. (1999). Object Recognition with Gradient-Based Learning. In D. A. Forsyth, J. L. Mundy, V. di Gesú, & R. Cipolla, *Shape, Contour and Grouping in Computer Vision* (Vol. 1681, pp. 319–345). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-46805-6_19

Lee, J. K., Andrews, D. S., Ozonoff, S., Solomon, M., Rogers, S., Amaral, D. G., & Nordahl, C. W. (2021). Longitudinal Evaluation of Cerebral Growth Across Childhood in Boys and Girls With Autism Spectrum Disorder. *Biological Psychiatry*, *90*(5), 286–294. https://doi.org/10.1016/j.biopsych.2020.10.014

Lee, P. H., Anttila, V., Won, H., Feng, Y.-C. A., Rosenthal, J., Zhu, Z., Tucker-Drob, E. M., Nivard, M. G., Grotzinger, A. D., Posthuma, D., Wang, M. M.-J., Yu, D., Stahl, E. A., Walters, R. K., Anney, R. J. L., Duncan, L. E., Ge, T., Adolfsson, R., Banaschewski, T., … Smoller, J. W. (2019). Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell*, *179*(7), 1469-1482.e11. https://doi.org/10.1016/j.cell.2019.11.020

Leming, M., Górriz, J. M., & Suckling, J. (2020). Ensemble Deep Learning on Large, Mixed-Site fMRI Datasets in Autism and Other Tasks. *International Journal of Neural Systems*, *30*(07), 2050012. https://doi.org/10.1142/S0129065720500124

Leming, M. J., Baron-Cohen, S., & Suckling, J. (2021). Single-participant structural similarity matrices lead to greater accuracy in classification of participants than function in autism in MRI. *Molecular Autism*, *12*(1), 34. https://doi.org/10.1186/s13229-021-00439-5

Leming, M., & Suckling, J. (2019). Deep Learning on Brain Images in Autism: What Do Large Samples Reveal of Its Complexity? In J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López, J. Toledo Moreo, & H. Adeli (Eds.), *Understanding the Brain Function and Emotions* (pp. 389–402). Springer International Publishing. https://doi.org/10.1007/978-3-030-19591-5_40

Li, J., Wang, F., Pan, J., & Wen, Z. (2021). Identification of Autism Spectrum Disorder With Functional Graph Discriminative Network. *Frontiers in Neuroscience*, *15*, 729937. https://doi.org/10.3389/fnins.2021.729937

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M. J., & Vercauteren, T. (2017). On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task. In M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.-T. Yap, & D. Shen (Eds.), *Information Processing in Medical Imaging* (pp. 348–360). Springer International Publishing. https://doi.org/10.1007/978-3-319-59050-9_28

Li, X., Dvornek, N. C., Papademetris, X., Zhuang, J., Staib, L. H., Ventola, P., & Duncan, J. S. (2018). 2-Channel convolutional 3D deep neural network (2CC3D) for fMRI analysis: ASD classification and feature learning. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 1252–1255. https://doi.org/10.1109/ISBI.2018.8363798

Li, X., Dvornek, N. C., Zhuang, J., Ventola, P., & Duncan, J. S. (2018). Brain Biomarker Interpretation in ASD Using Deep Learning and fMRI. *Medical Image Computing and Computer-Assisted Intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, *11072*, 206–214. https://doi.org/10.1007/978-3-030-00931-1_24

Lindell, A. K., & Hudry, K. (2013). Atypicalities in cortical structure, handedness, and functional lateralization for language in autism spectrum disorders. *Neuropsychology Review*, *23*(3), 257–270. https://doi.org/10.1007/s11065-013-9234-5

Lombardo, M. V., Pierce, K., Eyler, L., Barnes, C. C., Ahrens-Barbeau, C., Solso, S., Campbell, K., & Courchesne, E. (2015). Different functional neural substrates for good and poor language outcome in autism. *Neuron*, *86*(2), 567–577. https://doi.org/10.1016/j.neuron.2015.03.023

Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(6), 466–474. https://doi.org/10.1016/j.jaac.2017.03.013

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, *19*(2), 185–212. https://doi.org/10.1007/BF02211841

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *24*(5), 659–685. https://doi.org/10.1007/BF02172145

Lu, H., Liu, S., Wei, H., & Tu, J. (2020). Multi-kernel fuzzy clustering based on auto-encoder for fMRI functional network. *Expert Systems with Applications*, *159*, 113513. https://doi.org/10.1016/j.eswa.2020.113513

Luo, Y., Wang, Y., Zu, C., Zhan, B., Wu, X., Zhou, J., Shen, D., & Zhou, L. (2021). 3D Transformer-GAN for High-Quality PET Reconstruction. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, & C. Essert (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (pp. 276–285). Springer International Publishing. https://doi.org/10.1007/978-3-030-87231-1_27

Maenner, M. J. (2023). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *MMWR. Surveillance Summaries*, *72*. https://doi.org/10.15585/mmwr.ss7202a1

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), 1014–1019. https://doi.org/10.1038/s41593-022-01114-5

Malkiel, I., Rosenman, G., Wolf, L., & Hendler, T. (2022). *Self-Supervised Transformers for fMRI representation* (arXiv:2112.05761). arXiv. http://arxiv.org/abs/2112.05761

Marcus, D. S., Harms, M. P., Snyder, A. Z., Jenkinson, M., Wilson, J. A., Glasser, M. F., Barch, D. M., Archie, K. A., Burgess, G. C., Ramaratnam, M., Hodge, M., Horton, W., Herrick, R., Olsen, T., McKay, M., House, M., Hileman, M., Reid, E., Harwell, J., … Van Essen, D. C. (2013). Human Connectome Project informatics: Quality control, database services, and data visualization. *NeuroImage*, *80*, 202–219. https://doi.org/10.1016/j.neuroimage.2013.05.077

McAvoy, M., Mitra, A., Coalson, R. S., d'Avossa, G., Keidel, J. L., Petersen, S. E., & Raichle, M. E. (2016). Unmasking Language Lateralization in Human Brain Intrinsic Activity. *Cerebral Cortex (New York, N.Y.: 1991)*, *26*(4), 1733–1746. https://doi.org/10.1093/cercor/bhv007

McKinnon, C. J., Eggebrecht, A. T., Todorov, A., Wolff, J. J., Elison, J. T., Adams, C. M., Snyder, A. Z., Estes, A. M., Zwaigenbaum, L., Botteron, K. N., McKinstry, R. C., Marrus, N., Evans, A., Hazlett, H. C., Dager, S. R., Paterson, S. J., Pandey, J., Schultz, R. T., Styner, M. A., … Pruett, J. R. (2019). Restricted and Repetitive Behavior and Brain Functional Connectivity in Infants at Risk for Developing Autism Spectrum Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*(1), 50–61. https://doi.org/10.1016/j.bpsc.2018.09.008

Meinke, A., & Hein, M. (2020). *Towards neural networks that provably know when they don't know* (arXiv:1909.12180). arXiv. https://doi.org/10.48550/arXiv.1909.12180

Miller, M., Musser, E. D., Young, G. S., Olson, B., Steiner, R. D., & Nigg, J. T. (2019). Sibling Recurrence Risk and Cross-aggregation of Attention-Deficit/Hyperactivity Disorder and Autism Spectrum Disorder. *JAMA Pediatrics*, *173*(2), 147. https://doi.org/10.1001/jamapediatrics.2018.4076

Milton, D. E. M., Heasman, B., & Sheppard, E. (2018). Double Empathy. In F. R. Volkmar (Ed.), *Encyclopedia of Autism Spectrum Disorders* (pp. 1–8). Springer New York. https://doi.org/10.1007/978-1-4614-6435-8_102273-1

MRC AIMS Consortium, Bedford, S. A., Park, M. T. M., Devenyi, G. A., Tullo, S., Germann, J., Patel, R., Anagnostou, E., Baron-Cohen, S., Bullmore, E. T., Chura, L. R., Craig, M. C., Ecker, C., Floris, D. L., Holt, R. J., Lenroot, R., Lerch, J. P., Lombardo, M. V., Murphy, D. G. M., … Chakravarty, M. M. (2020). Large-scale analyses of the relationship between sex, age and intelligence quotient heterogeneity and cortical morphometry in autism spectrum disorder. *Molecular Psychiatry*, *25*(3), 614–628. https://doi.org/10.1038/s41380-019-0420-6

Mundy, P. (2003). Annotation: The neural basis of social impairments in autism: the role of the dorsal medial-frontal cortex and anterior cingulate system. *Journal of Child Psychology and Psychiatry*, *44*(6), 793–809. https://doi.org/10.1111/1469-7610.00165

Nakagawa, N., Plestant, C., Yabuno-Nakagawa, K., Li, J., Lee, J., Huang, C.-W., Lee, A., Krupa, O., Adhikari, A., Thompson, S., Rhynes, T., Arevalo, V., Stein, J. L., Molnár, Z., Badache, A., & Anton, E. S. (2019). Memo1-Mediated Tiling of Radial Glial Cells Facilitates Cerebral Cortical Development. *Neuron*, *103*(5), 836-852.e5. https://doi.org/10.1016/j.neuron.2019.05.049

Nebel, M. B., Eloyan, A., Barber, A. D., & Mostofsky, S. H. (2014). Precentral gyrus functional connectivity signatures of autism. *Frontiers in Systems Neuroscience*, *8*, 80. https://doi.org/10.3389/fnsys.2014.00080

Nguyen, S., Ng, B., Kaplan, A. D., & Ray, P. (2020). Attend and Decode: 4D fMRI Task State Decoding Using Attention Models.

Nordahl, C. W., Dierker, D., Mostafavi, I., Schumann, C. M., Rivera, S. M., Amaral, D. G., & Van Essen, D. C. (2007). Cortical folding abnormalities in autism revealed by surface-based morphometry. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(43), 11725–11735. https://doi.org/10.1523/JNEUROSCI.0777-07.2007

Nordahl, C. W., Mello, M., Shen, A. M., Shen, M. D., Vismara, L. A., Li, D., Harrington, K., Tanase, C., Goodlin-Jones, B., Rogers, S., Abbeduto, L., & Amaral, D. G. (2016). Methods for acquiring MRI data in children with autism spectrum disorder and intellectual impairment without the use of sedation. *Journal of Neurodevelopmental Disorders*, *8*(1), 20. https://doi.org/10.1186/s11689-016-9154-9

O'Regan, E. (2023, March 1). *Children with autism 'harmed by delays in disability services', Oireachtas committee told*. Independent.Ie. https://www.independent.ie/irish-news/children-with-autism-harmed-by-delays-in-disability-services-oireachtas-committee-told/42364218.html

Pagani, M., Bertero, A., Liska, A., Galbusera, A., Sabbioni, M., Barsotti, N., Colenbier, N., Marinazzo, D., Scattoni, M. L., Pasqualetti, M., & Gozzi, A. (2019). Deletion of Autism Risk Gene Shank3 Disrupts Prefrontal Connectivity. *The Journal of Neuroscience*, *39*(27), 5299–5310. https://doi.org/10.1523/JNEUROSCI.2529-18.2019

Pagnozzi, A. M., Conti, E., Calderoni, S., Fripp, J., & Rose, S. E. (2018). A systematic review of structural MRI biomarkers in autism spectrum disorder: A machine learning perspective.

International Journal of Developmental Neuroscience: The Official Journal of the International Society for Developmental Neuroscience, *71*, 68–82. https://doi.org/10.1016/j.ijdevneu.2018.08.010

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., Horvath, S., & Geschwind, D. H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, *155*(5), 1008–1021. https://doi.org/10.1016/j.cell.2013.10.031

Patriquin, M. A., DeRamus, T., Libero, L. E., Laird, A., & Kana, R. K. (2016). Neuroanatomical and neurofunctional markers of social cognition in autism spectrum disorder. *Human Brain Mapping*, *37*(11), 3957–3978. https://doi.org/10.1002/hbm.23288

Pereira, A. M., Campos, B. M., Coan, A. C., Pegoraro, L. F., de Rezende, T. J. R., Obeso, I., Dalgalarrondo, P., da Costa, J. C., Dreher, J.-C., & Cendes, F. (2018). Differences in Cortical Structure and Functional MRI Connectivity in High Functioning Autism. *Frontiers in Neurology*, *9*. https://www.frontiersin.org/article/10.3389/fneur.2018.00539

Perrone, M., & Cooper, L. (1993). When Networks Disagree: Ensemble Methods for Hybrid Neural Networks. *Neural Networks for Speech and Image Processing*. https://doi.org/10.1142/9789812795885_0025

Pierce, K., Gazestani, V. H., Bacon, E., Barnes, C. C., Cha, D., Nalabolu, S., Lopez, L., Moore, A., Pence-Stophaeros, S., & Courchesne, E. (2019). Evaluation of the Diagnostic Stability of the Early Autism Spectrum Disorder Phenotype in the General Population Starting at 12 Months. *JAMA Pediatrics*, *173*(6), 578. https://doi.org/10.1001/jamapediatrics.2019.0624

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quidé, Y., Green, M. J., Weickert, C. S., Weickert, T., Bruggemann, J., Kircher, T., Nenadić, I., Cairns, M. J., Seal, M., Schall, U., Henskens, F., Fullerton, J. M., Mowry, B., Pantelis, C., Lenroot, R., … Pineda-Zapata, J. (2020). Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. *NeuroImage*, *218*, 116956. https://doi.org/10.1016/j.neuroimage.2020.116956

Rafiee, F., Rezvani Habibabadi, R., Motaghi, M., Yousem, D. M., & Yousem, I. J. (2022). Brain MRI in Autism Spectrum Disorder: Narrative Review and Recent Advances. *Journal of*

Magnetic Resonance Imaging: JMRI, *55*(6), 1613–1624. https://doi.org/10.1002/jmri.27949

Rakić, M., Cabezas, M., Kushibar, K., Oliver, A., & Lladó, X. (2020). Improving the detection of autism spectrum disorder by combining structural and functional MRI information. *NeuroImage. Clinical*, *25*, 102181. https://doi.org/10.1016/j.nicl.2020.102181

Rauch, S. L. (2005). Neuroimaging and Attention-Deficit/Hyperactivity Disorder in the 21st Century: What to Consider and How to Proceed. *Biological Psychiatry*, *57*(11), 1261–1262. https://doi.org/10.1016/j.biopsych.2005.02.014

Raznahan, A., Wallace, G. L., Antezana, L., Greenstein, D., Lenroot, R., Thurm, A., Gozzi, M., Spence, S., Martin, A., Swedo, S. E., & Giedd, J. N. (2013). Compared to what? Early brain overgrowth in autism and the perils of population norms. *Biological Psychiatry*, *74*(8), 563–575. https://doi.org/10.1016/j.biopsych.2013.03.022

Reports on the prevalence of autism in Ireland and a review of the services for people with autism. (2018, December 5). https://www.gov.ie/en/publication/0cc791-reports-on-the-prevalence-of-autism-in-ireland-and-a-review-of-the-s/?referrer=http://www.health.gov.ie/blog/publications/reports-on-the-prevalence-of-autism-in-ireland-and-a-review-of-the-services-for-people-with-autism/

Retico, A., Giuliano, A., Tancredi, R., Cosenza, A., Apicella, F., Narzisi, A., Biagi, L., Tosetti, M., Muratori, F., & Calderoni, S. (2016). The effect of gender on the neuroanatomy of children with autism spectrum disorders: A support vector machine case-control study. *Molecular Autism*, *7*(1), 5. https://doi.org/10.1186/s13229-015-0067-3

Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., Van Der Kouwe, A. J. W., & Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage*, *107*, 107–115. https://doi.org/10.1016/j.neuroimage.2014.12.006

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *'Why Should I Trust You?': Explaining the Predictions of Any Classifier* (arXiv:1602.04938). arXiv. https://doi.org/10.48550/arXiv.1602.04938

Riddle, K., Cascio, C. J., & Woodward, N. D. (2017). Brain structure in autism: A voxel-based morphometry analysis of the Autism Brain Imaging Database Exchange (ABIDE). *Brain Imaging and Behavior*, *11*(2), 541–551. https://doi.org/10.1007/s11682-016-9534-5

Rogers, S. J., Vismara, L., Wagner, A. L., McCormick, C., Young, G., & Ozonoff, S. (2014). Autism Treatment in the First Year of Life: A Pilot Study of Infant Start, a Parent-Implemented Intervention for Symptomatic Infants. *Journal of Autism and Developmental Disorders*, *44*(12), 2981–2995. https://doi.org/10.1007/s10803-014-2202-y

Ruder, S. (2017). *An overview of gradient descent optimization algorithms* (arXiv:1609.04747). arXiv. http://arxiv.org/abs/1609.04747

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning Internal Representations by Error Propagation. In *Readings in Cognitive Science* (pp. 399–421). Elsevier. https://doi.org/10.1016/B978-1-4832-1446-7.50035-2

Rutter M, Bailey A, Berument S, et al. (2003). Social Communication Questionnaire (SCQ). *Western Psychological Services.*

Ruzzo, E. K., Pérez-Cano, L., Jung, J.-Y., Wang, L., Kashef-Haghighi, D., Hartl, C., Singh, C., Xu, J., Hoekstra, J. N., Leventhal, O., Leppä, V. M., Gandal, M. J., Paskov, K., Stockham, N., Polioudakis, D., Lowe, J. K., Prober, D. A., Geschwind, D. H., & Wall, D. P. (2019). Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. *Cell*, *178*(4), 850-866.e26. https://doi.org/10.1016/j.cell.2019.07.015

Sacco, R., Camilleri, N., Eberhardt, J., Umla-Runge, K., & Newbury-Birch, D. (2022). *The Prevalence of Autism Spectrum Disorder in Europe*. https://doi.org/10.5772/intechopen.108123

Sanders, S. J. (2015). First glimpses of the neurobiology of autism spectrum disorder. *Current Opinion in Genetics & Development*, *33*, 80–92. https://doi.org/10.1016/j.gde.2015.10.002

Saragosa-Harris, N. M., Chaku, N., MacSweeney, N., Guazzelli Williamson, V., Scheuplein, M., Feola, B., Cardenas-Iniguez, C., Demir-Lira, E., McNeilly, E. A., Huffman, L. G., Whitmore, L., Michalska, K. J., Damme, K. S., Rakesh, D., & Mills, K. L. (2022). A practical guide for researchers and reviewers using the ABCD Study and other large longitudinal datasets. *Developmental Cognitive Neuroscience*, *55*, 101115. https://doi.org/10.1016/j.dcn.2022.101115

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., … Walters, R. K. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, *180*(3), 568-584.e23. https://doi.org/10.1016/j.cell.2019.12.036

Schork, A. J., Won, H., Appadurai, V., Nudel, R., Gandal, M., Delaneau, O., Revsbech Christiansen, M., Hougaard, D. M., Bækved-Hansen, M., Bybjerg-Grauholm, J., Giørtz Pedersen, M., Agerbo, E., Bøcker Pedersen, C., Neale, B. M., Daly, M. J., Wray, N. R., Nordentoft, M., Mors, O., Børglum, A. D., … Werge, T. (2019). A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nature Neuroscience*, *22*(3), 353–361. https://doi.org/10.1038/s41593-018-0320-0

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv:1610.02391 [Cs]*. https://doi.org/10.1007/s11263-019-01228-7

Sha, Z., Wager, T. D., Mechelli, A., & He, Y. (2019). Common Dysfunction of Large-Scale Neurocognitive Networks Across Psychiatric Disorders. *Biological Psychiatry*, *85*(5), 379–388. https://doi.org/10.1016/j.biopsych.2018.11.011

Sharda, M., Khundrakpam, B. S., Evans, A. C., & Singh, N. C. (2016). Disruption of structural covariance networks for language in autism is modulated by verbal ability. *Brain Structure & Function*, *221*(2), 1017–1032. https://doi.org/10.1007/s00429-014-0953-z

Shehzad, Z., Giavasis, S., Li, Q., Yassine, Y., Yan, C., Liu, Z., Milham, M., Bellec, P., & Craddock, C. (2015). The Preprocessed Connectomes Project Quality Assessment Protocol—A resource for measuring the quality of MRI data. *Frontiers in Neuroscience*, *9*. https://doi.org/10.3389/conf.fnins.2015.91.00047

Sherkatghanad, Z., Akhondzadeh, M., Salari, S., Zomorodi-Moghadam, M., Abdar, M., Acharya, U. R., Khosrowabadi, R., & Salari, V. (2020). Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network. *Frontiers in Neuroscience*, *13*. https://www.frontiersin.org/articles/10.3389/fnins.2019.01325

Silva, A. I., Haddon, J. E., Ahmed Syed, Y., Trent, S., Lin, T.-C. E., Patel, Y., Carter, J., Haan, N., Honey, R. C., Humby, T., Assaf, Y., Owen, M. J., Linden, D. E. J., Hall, J., & Wilkinson, L. S. (2019). Cyfip1 haploinsufficient rats show white matter changes, myelin thinning, abnormal oligodendrocytes and behavioural inflexibility. *Nature Communications*, *10*(1), 3455. https://doi.org/10.1038/s41467-019-11119-7

Simonoff, E., Pickles, A., Charman, T., Chandler, S., Loucas, T., & Baird, G. (2008). Psychiatric Disorders in Children With Autism Spectrum Disorders: Prevalence, Comorbidity, and Associated Factors in a Population-Derived Sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*(8), 921–929. https://doi.org/10.1097/CHI.0b013e318179964f

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [Cs]*. http://arxiv.org/abs/1409.1556

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., & Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, *106*(31), 13040–13045. https://doi.org/10.1073/pnas.0905267106

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). Striving for Simplicity: The All Convolutional Net. *arXiv:1412.6806 [Cs]*. http://arxiv.org/abs/1412.6806

Subah, F. Z., Deb, K., Dhar, P. K., & Koshiba, T. (2021). A Deep Learning Approach to Predict Autism Spectrum Disorder Using Multisite Resting-State fMRI. *Applied Sciences*, *11*(8), Article 8. https://doi.org/10.3390/app11083636

Subbaraju, V., Suresh, M. B., Sundaram, S., & Narasimhan, S. (2017). Identifying differences in brain activities and an accurate detection of autism spectrum disorder using resting state functional-magnetic resonance imaging: A spatial filtering approach. *Medical Image Analysis*, *35*, 375–389. https://doi.org/10.1016/j.media.2016.08.003

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Sui, J., Adali, T., Yu, Q., & Calhoun, V. D. (2012). A Review of Multivariate Methods for Multimodal Fusion of Brain Imaging Data. *Journal of Neuroscience Methods*, *204*(1), 68–81. https://doi.org/10.1016/j.jneumeth.2011.10.031

Sujit, S. J., Coronado, I., Kamali, A., Narayana, P. A., & Gabr, R. E. (2019). Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. *Journal of Magnetic Resonance Imaging*, *50*(4), 1260–1267. https://doi.org/10.1002/jmri.26693

Szatmari, P., Georgiades, S., Duku, E., Bennett, T. A., Bryson, S., Fombonne, E., Mirenda, P., Roberts, W., Smith, I. M., Vaillancourt, T., Volden, J., Waddell, C., Zwaigenbaum, L., Elsabbagh, M., Thompson, A., & Pathways in ASD Study Team. (2015). Developmental trajectories of symptom severity and adaptive functioning in an inception cohort of preschool children with autism spectrum disorder. *JAMA Psychiatry*, *72*(3), 276–283. https://doi.org/10.1001/jamapsychiatry.2014.2463

Tager-Flusberg, H., & Kasari, C. (2013). Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum. *Autism Research: Official Journal of the International Society for Autism Research*, *6*(6), 468–478. https://doi.org/10.1002/aur.1329

Tanno, R., Worrall, D., Ghosh, A., Kaden, E., Sotiropoulos, S., Criminisi, A., & Alexander, D. (2017). *Bayesian Image Quality Transfer with CNNs: Exploring Uncertainty in dMRI Super-Resolution*. 611–619. https://doi.org/10.1007/978-3-319-66182-7_70

The Brainstorm Consortium, Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G. J., Gormley, P., Malik, R., Patsopoulos, N. A., Ripke, S., Wei, Z., Yu, D., Lee, P. H., Turley, P., Grenier-Boley, B., Chouraki, V., … Neale,

B. M. (2018). Analysis of shared heritability in common disorders of the brain. *Science*, *360*(6395), eaap8757. https://doi.org/10.1126/science.aap8757

Thomas, A. W., Ré, C., & Poldrack, R. A. (2023). *Self-Supervised Learning of Brain Dynamics from Broad Neuroimaging Data* (arXiv:2206.11417). arXiv. http://arxiv.org/abs/2206.11417

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

Thompson, P. M., Jahanshad, N., Ching, C. R. K., Salminen, L. E., Thomopoulos, S. I., Bright, J., Baune, B. T., Bertolín, S., Bralten, J., Bruin, W. B., Bülow, R., Chen, J., Chye, Y., Dannlowski, U., De Kovel, C. G. F., Donohoe, G., Eyler, L. T., Faraone, S. V., Favre, P., … for the ENIGMA Consortium. (2020). ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Translational Psychiatry*, *10*(1), 100. https://doi.org/10.1038/s41398-020-0705-1

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Traut, N., Heuer, K., Lemaître, G., Beggiato, A., Germanaud, D., Elmaleh, M., Bethegnies, A., Bonnasse-Gahot, L., Cai, W., Chambon, S., Cliquet, F., Ghriss, A., Guigui, N., de Pierrefeu, A., Wang, M., Zantedeschi, V., Boucaud, A., van den Bossche, J., Kegl, B., … Varoquaux, G. (2021). *Insights from an autism imaging biomarker challenge: Promises and threats to biomarker discovery* [Preprint]. Radiology and Imaging. https://doi.org/10.1101/2021.11.24.21266768

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978 van Rooij, D., Anagnostou, E., Arango, C., Auzias, G., Behrmann, M., Busatto, G. F., Calderoni, S., Daly, E., Deruelle, C., Di Martino, A., Dinstein, I., Duran, F. L. S., Durston, S., Ecker, C., Fair, D., Fedor, J., Fitzgerald, J., Freitag, C. M., Gallagher, L., … Buitelaar, J. K. (2018). Cortical and Subcortical Brain Morphometry Differences Between Patients With Autism Spectrum Disorder and Healthy Individuals Across the Lifespan: Results From the ENIGMA ASD Working Group. *The American Journal of Psychiatry*, *175*(4), 359–369. https://doi.org/10.1176/appi.ajp.2017.17010100

Van 'T Hof, M., Tisseur, C., Van Berckelear-Onnes, I., Van Nieuwenhuyzen, A., Daniels, A. M., Deen, M., Hoek, H. W., & Ester, W. A. (2021). Age at autism spectrum disorder diagnosis: A systematic review and meta-analysis from 2012 to 2019. *Autism*, *25*(4), 862–873. https://doi.org/10.1177/1362361320971107

Van Wijngaarden-Cremers, P. J. M., Van Eeten, E., Groen, W. B., Van Deurzen, P. A., Oosterling, I. J., & Van Der Gaag, R. J. (2014). Gender and Age Differences in the Core Triad of Impairments in Autism Spectrum Disorders: A Systematic Review and Meta-analysis. *Journal of Autism and Developmental Disorders*, *44*(3), 627–635. https://doi.org/10.1007/s10803-013-1913-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., Pérez-Stable, E. J., Riley, W. T., Bloch, M. H., Conway, K., Deeds, B. G., Dowling, G. J., Grant, S., Howlett, K. D., Matochik, J. A., Morgan, G. D., Murray, M. M., Noronha, A., Spong, C. Y., … Weiss, S. R. B. (2018). The conception of the ABCD study: From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, *32*, 4–7. https://doi.org/10.1016/j.dcn.2017.10.002

Walbrin, J., Downing, P., & Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia*, *112*, 31–39. https://doi.org/10.1016/j.neuropsychologia.2018.02.023

Wang, Y., Wang, J., Wu, F.-X., Hayrat, R., & Liu, J. (2020). AIMAFE: Autism spectrum disorder identification with multi-atlas deep feature representation and ensemble learning. *Journal of Neuroscience Methods*, *343*, 108840. https://doi.org/10.1016/j.jneumeth.2020.108840

Werling, D. M., & Geschwind, D. H. (2013). Sex differences in autism spectrum disorders. *Current Opinion in Neurology*, *26*(2), 146–153. https://doi.org/10.1097/WCO.0b013e32835ee548

Whelan, C. D., Altmann, A., Botía, J. A., Jahanshad, N., Hibar, D. P., Absil, J., Alhusaini, S., Alvim, M. K. M., Auvinen, P., Bartolini, E., Bergo, F. P. G., Bernardes, T., Blackmon, K., Braga, B., Caligiuri, M. E., Calvo, A., Carr, S. J., Chen, J., Chen, S., … Sisodiya, S. M. (2018). Structural brain abnormalities in the common epilepsies assessed in a worldwide ENIGMA study. *Brain*, *141*(2), 391–408. https://doi.org/10.1093/brain/awx341

White, T., Jansen, P. R., Muetzel, R. L., Sudre, G., El Marroun, H., Tiemeier, H., Qiu, A., Shaw, P., Michael, A. M., & Verhulst, F. C. (2018). Automated quality assessment of structural magnetic resonance images in children: Comparison with visual inspection and surface-based reconstruction. *Human Brain Mapping*, *39*(3), 1218–1231. https://doi.org/10.1002/hbm.23911

Wolfers, T., Floris, D. L., Dinga, R., Van Rooij, D., Isakoglou, C., Kia, S. M., Zabihi, M., Llera, A., Chowdanayaka, R., Kumar, V. J., Peng, H., Laidi, C., Batalle, D., Dimitrova, R., Charman, T., Loth, E., Lai, M.-C., Jones, E., Baumeister, S., … Beckmann, C. F. (2019). From pattern classification to stratification: Towards conceptualizing the heterogeneity of Autism Spectrum Disorder. *Neuroscience & Biobehavioral Reviews*, *104*, 240–254. https://doi.org/10.1016/j.neubiorev.2019.07.010

Yang, D. Y.-J., Beam, D., Pelphrey, K. A., Abdullahi, S., & Jou, R. J. (2016). Cortical morphological markers in children with autism: A structural magnetic resonance imaging study of thickness, area, volume, and gyrification. *Molecular Autism*, *7*(1), 11. https://doi.org/10.1186/s13229-016-0076-x

Yang, M., Cao, M., Chen, Y., Chen, Y., Fan, G., Li, C., Wang, J., & Liu, T. (2021). Large-Scale Brain Functional Network Integration for Discrimination of Autism Using a 3-D Deep

Learning Model. *Frontiers in Human Neuroscience*, *15*, 687288. https://doi.org/10.3389/fnhum.2021.687288

Yao, Y., Rosasco, L., & Caponnetto, A. (2007). On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, *26*(2), 289–315. https://doi.org/10.1007/s00365-006-0663-2

Yoon, S., Parnell, E., Kasherman, M., Forrest, M. P., Myczek, K., Premarathne, S., Sanchez Vega, M. C., Piper, M., Burne, T. H. J., Jolly, L. A., Wood, S. A., & Penzes, P. (2020). Usp9X Controls Ankyrin-Repeat Domain Protein Homeostasis during Dendritic Spine Development. *Neuron*, *105*(3), 506-521.e7. https://doi.org/10.1016/j.neuron.2019.11.003

Yu, X., Zhang, L., Zhao, L., Lyu, Y., Liu, T., & Zhu, D. (2022). Disentangling Spatial-Temporal Functional Brain Networks via Twin-Transformers.

Zaadnoordijk, L., Besold, T. R., & Cusack, R. (2022). Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, *4*(6), 510–520. https://doi.org/10.1038/s42256-022-00488-2

Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., Charman, T., Tillmann, J., Banaschewski, T., Dumas, G., Holt, R., Baron-Cohen, S., Durston, S., Bölte, S., Murphy, D., Ecker, C., Buitelaar, J. K., Beckmann, C. F., & Marquand, A. F. (2019). Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *4*(6), 567–578. https://doi.org/10.1016/j.bpsc.2018.11.013

Zeidan, J., Fombonne, E., Scorah, J., Ibrahim, A., Durkin, M. S., Saxena, S., Yusuf, A., Shih, A., & Elsabbagh, M. (2022). Global prevalence of autism: A systematic review update. *Autism Research*, *15*(5), 778–790. https://doi.org/10.1002/aur.2696

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing*, *26*(7), 3142–3155. https://doi.org/10.1109/TIP.2017.2662206

Zhang, W., Ma, L., Yang, M., Shao, Q., Xu, J., Lu, Z., Zhao, Z., Chen, R., Chai, Y., & Chen, J.-F. (2020). Cerebral organoid and mouse models reveal a RAB39b–PI3K–mTOR pathwaydependent dysregulation of cortical development leading to macrocephaly/autism phenotypes. *Genes & Development*, *34*(7–8), 580–597. https://doi.org/10.1101/gad.332494.119

Zhang, Y., Liu, H., & Hu, Q. (2021). *TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation* (arXiv:2102.08005). arXiv. https://doi.org/10.48550/arXiv.2102.08005

Zhao, L., Wu, Z., Dai, H., Liu, Z., Zhang, T., Zhu, D., & Liu, T. (2022). Embedding Human Brain Function via Transformer. In L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, & S. Li (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (pp. 366–375). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-16431-6_35

Zheng, W., Zhao, Z., Zhang, Z., Liu, T., Zhang, Y., Fan, J., & Wu, D. (2020). Developmental pattern of the cortical topology in high-functioning individuals with autism spectrum disorder. *Human Brain Mapping*, *42*(3), 660–675. https://doi.org/10.1002/hbm.25251

Zheng, W., Zhao, Z., Zhang, Z., Liu, T., Zhang, Y., Fan, J., & Wu, D. (2021). Developmental pattern of the cortical topology in HIGH-FUNCTIONING individuals with autism spectrum disorder. *Human Brain Mapping*, *42*(3), 660–675. https://doi.org/10.1002/hbm.25251

Zielinski, B. A., Prigge, M. B. D., Nielsen, J. A., Froehlich, A. L., Abildskov, T. J., Anderson, J. S., Fletcher, P. T., Zygmunt, K. M., Travers, B. G., Lange, N., Alexander, A. L., Bigler, E. D., & Lainhart, J. E. (2014). Longitudinal changes in cortical thickness in autism and typical development. *Brain*, *137*(6), 1799–1812. https://doi.org/10.1093/brain/awu083

Zuckerman, K., Lindly, O. J., & Chavez, A. E. (2017). Timeliness of Autism Spectrum Disorder Diagnosis and Use of Services Among U.S. Elementary School–Aged Children. *Psychiatric Services*, *68*(1), 33–40. https://doi.org/10.1176/appi.ps.201500549

Zwaigenbaum, L., Thurm, A., Stone, W., Baranek, G., Bryson, S., Iverson, J., Kau, A., Klin, A., Lord, C., Landa, R., Rogers, S., & Sigman, M. (2007). Studying the Emergence of Autism Spectrum Disorders in High-risk Infants: Methodological and Practical Issues. *Journal of Autism and Developmental Disorders*, *37*(3), 466–480. https://doi.org/10.1007/s10803-006-0179-x

APPENDICES

Appendix 1: Supplemental Information for Manuscript: “BrainQCNet: a Deep Learning attention-based model for the automated detection of artefacts in brain structural MRI scans”

8.6.1. A1.1. Comparison of the distribution of probabilities between models

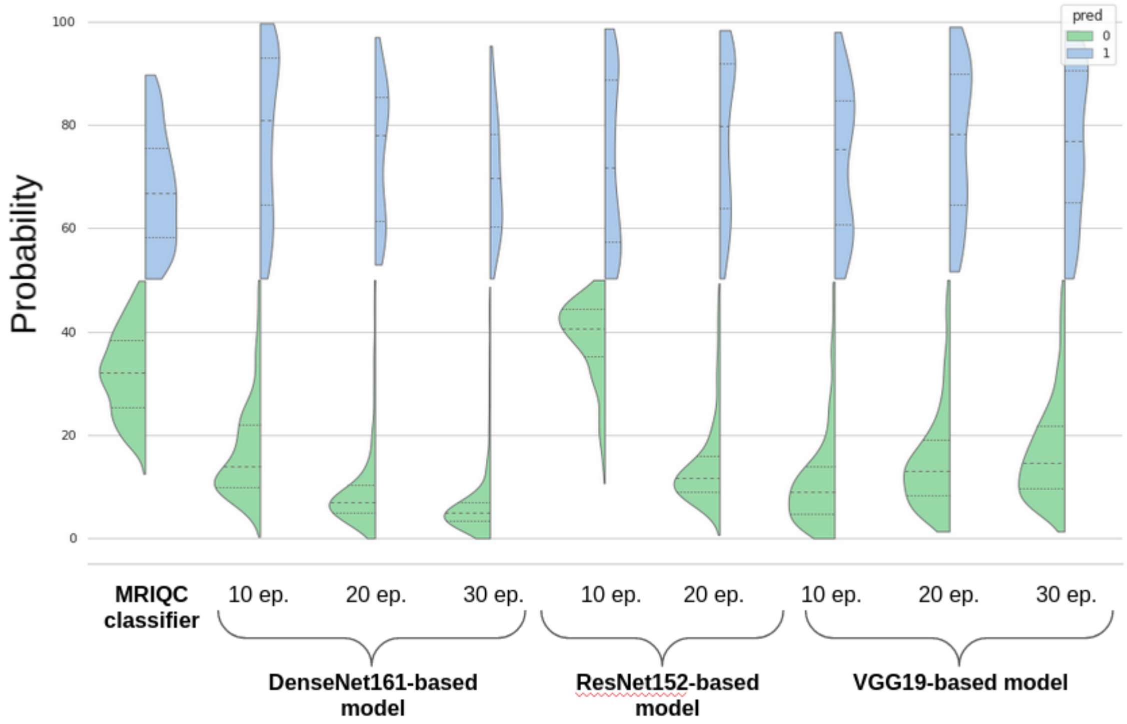


Figure A1.1: Comparison of the distribution of probabilities for the test set (908 scans), colored by predicted class: green for Class 0 (good quality scans), blue for Class 1 (medium/low quality scans).

In **Figure A1.1**, we can see that the distribution of predictions of Class 0 scans (green) looks gaussian for our models. In contrast, the distribution of predictions for Class 1 (blue) looks like a gaussian mixture. This distribution shape is expected since there are some scans that are globally corrupted scans, but others with only local artefact, or less severe global artefact. The proportion of slices classified as Class 1 will therefore be different for the two types.

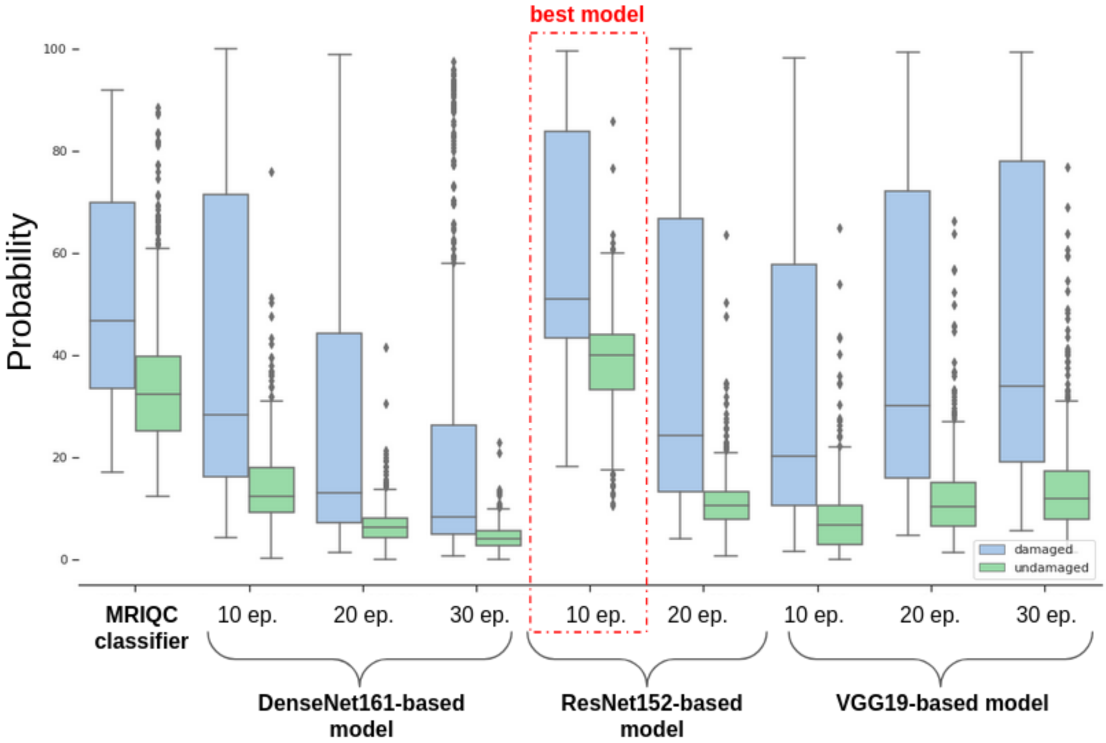


Figure A1.2: The boxplots show the predicted probabilities (% slices predicted to be Class 1, poor quality) for scans manually judged to be free from artefact (Class 0 - good quality; green) vs. those manually judged to be contain some artefact (Class 1 - blue) for all models and for MRIQC, using 980 scans from ABIDE 1. The figure shows that there is some overlap in the global probabilities for Class 0 and Class 1 scans, although this varies by model and by epoch. The greater the overlap, the more False Positives and False Negatives there are. The overlap is least, and optimally located (around 40% probability) for the best-performing model, **proto-R152 (10 epochs).**

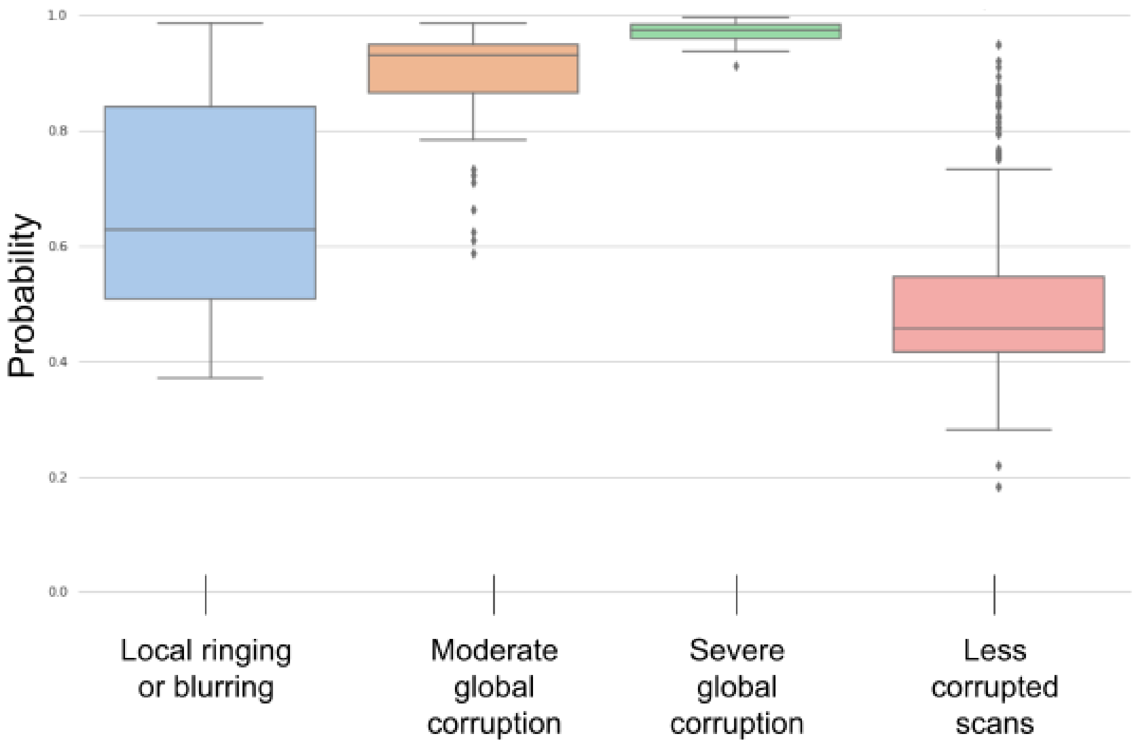


Figure A1.3: Comparison of probabilities for global predictions from proto-ResNet152 trained on 10 epochs for 416 Class 1 (poor quality) scans from ABIDE 1 (30 very low quality scans included in the training set, 6 very low quality scans included in the validation set, 380 less severely poor quality scans included in the test set). 51 scans have local ringing or blurring (blue), 60 are globally corrupted but medium quality (orange), 36 are globally corrupted and very low quality (green; i.e. score=4,4,4,4 and artefacts present on all the 2D slices), 269 are less severely corrupted or exhibit localised artefact only (red).

Figure A1.3 shows that each of the categories of artefact severity are well segregated in terms of their predicted probabilities: globally corrupted scans have probabilities very close to 1, while scans with moderate artefact (e.g., ringing or blurring) have probabilities spread between 0.5-0.8, and other scans with localised or less severe artefact have probabilities around 0.5.

The classes (Local/Severe/Moderate/Less corruption) were defined as follows:

- Severe/Moderate/Less: refers to an artefact or a set of artefacts that was globally evident on the scan. Severe means that the scan was highly corrupted: at least one of the four artefact scores (blurring; ringing; CNR WM/GM; CNR subcortical structures) was 4. Moderate means that at least one of the four artefact scores was 3. Less means that at least one of the four artefact scores was 2.

- Local: refers to an artefact or a set of artefacts (with scores between 2 and 4) that was present only on a demarcated area of the scan, and on for less than half of the slices.

We also evaluated the results on slices for the 66 scans from ABIDE 1 we annotated with local ringing and/or local blurring. We found that in the extremities, the algorithm tends to predict the slices as Class 1, even in the cases it should be Class 0. This means that slices near the edge of the field-of-view containing few brain tend to be identified as corrupted by the algorithm. This might explain why the global distribution of probabilities of the model proto-ResNet152-10ep is higher than the ones of other models (see **Figures A1.1 and A1.2**).

We also found an axis effect at the accuracy level (no effect between the distribution of scores) - while predictions for sagittal images were 89.3% accurate, accuracy for coronal images were 86.4% accurate, and for axial views, 78.8% accurate.

8.6.2. A1.2 Multi-site effect

	good quality - 528 scans	globally medium corrupted - 60 scans	local ringing or blurring - 51 scans	other less corrupted scans - 269 scans
CALTECH	accuracy: 1.0 n scans: 34	na	na	accuracy: 0.0 n scans: 2
CMU	accuracy: 1.0 n scans: 24	na	na	accuracy: 0.3333 n scans: 3
KKI	accuracy: 1.0 n scans: 25	accuracy: 1.0 n scans: 3	na	accuracy: 0.5714 n scans: 14
LEUVEN_1	accuracy: 0.9259 n scans: 27	na	na	accuracy: 0.5 n scans: 2
LEUVEN_2	accuracy: 0.9565 n scans: 23	na	accuracy: 1.0 n scans: 1	accuracy: 0.2 n scans: 10
MAX_MUN	accuracy: 0.9286 n scans: 28	accuracy: 1.0 n scans: 2	accuracy: 1.0 n scans: 1	accuracy: 0.8 n scans: 10
NYU	accuracy: 0.9146 n scans: 82	accuracy: 1.0 n scans: 1	accuracy: 0.5882 n scans: 17	accuracy: 0.2714 n scans: 70
OHSU	accuracy: 0.9091	accuracy: 1.0	na	na

	n scans: 22	n scans: 1		
OLIN	accuracy: 0.75 n scans: 12	na	accuracy: 1.0 n scans: 2	accuracy: 0.4286 n scans: 7
PITT	accuracy: 0.9524 n scans: 21	na	accuracy: 1.0 n scans: 5	accuracy: 0.3913 n scans: 23
SBL	accuracy: 1.0 n scans: 26	na	na	accuracy: 0.0 n scans: 4
SDSU	accuracy: 0.8 n scans: 10	accuracy: 1.0 n scans: 10	na	accuracy: 0.8 n scans: 10
STANFORD	na	accuracy: 1.0 n scans: 10	accuracy: 0.8333 n scans: 12	accuracy: 0.6667 n scans: 6
TRINITY	accuracy: 1.0 n scans: 34	accuracy: 1.0 n scans: 3	accuracy: 1.0 n scans: 1	accuracy: 0.0 n scans: 7
UCLA_1	accuracy: 0.8958 n scans: 48	accuracy: 1.0 n scans: 6	accuracy: 0.6667 n scans: 3	accuracy: 0.8 n scans: 5
UCLA_2	accuracy: 1.0 n scans: 7	accuracy: 1.0 n scans: 3	accuracy: 1.0 n scans: 1	accuracy: 0.4286 n scans: 7
UM_1	accuracy: 1.0	accuracy: 1.0	accuracy: 0.8	accuracy: 0.1471

	n scans: 27	n scans: 7	n scans: 10	n scans: 34
UM_2	accuracy: 1.0 n scans: 13	na	accuracy: 0.6667 n scans: 3	accuracy: 0.25 n scans: 12
USM	accuracy: 1.0 n scans: 60	na	na	accuracy: 0.25 n scans: 4
YALE	accuracy: 1.0 n scans: 5	accuracy: 1.0 n scans: 5	accuracy: 0.5 n scans: 4	accuracy: 0.1795 n scans: 39

Table A1.1: Predictions for each data collection site in the test set (908 scans) for proto-ResNet152 trained on 10 epochs.

Table A1.1 displays the predictions of the model proto-ResNet152 trained on 10 epochs for each data collection site in the first test set of 908 scans from the ABIDE 1 dataset.

For Class 0 (good quality, pass) scans, the model attained 100% accuracy for 10 out of 19 sites, >90% for 16 of 19, and accuracy of at least 75% for the remaining 3 sites. Overall the mean accuracy across sites is 94.9% with a standard deviation of 0.07. There does not appear to be a significant effect on site effect on the prediction of the good quality (Class 0) scans.

For Class 1 (poor quality, fail) scans with moderate levels of artefact, accuracy is 100% across sites. For scans with less severe or localised artefact, there is more variability across sites, but there is also large variation in the number of scans in that category, so it is difficult to quantitatively assess whether there is a significant site effect.

8.6.3. A1.3 Prototypes mostly used by proto-R152

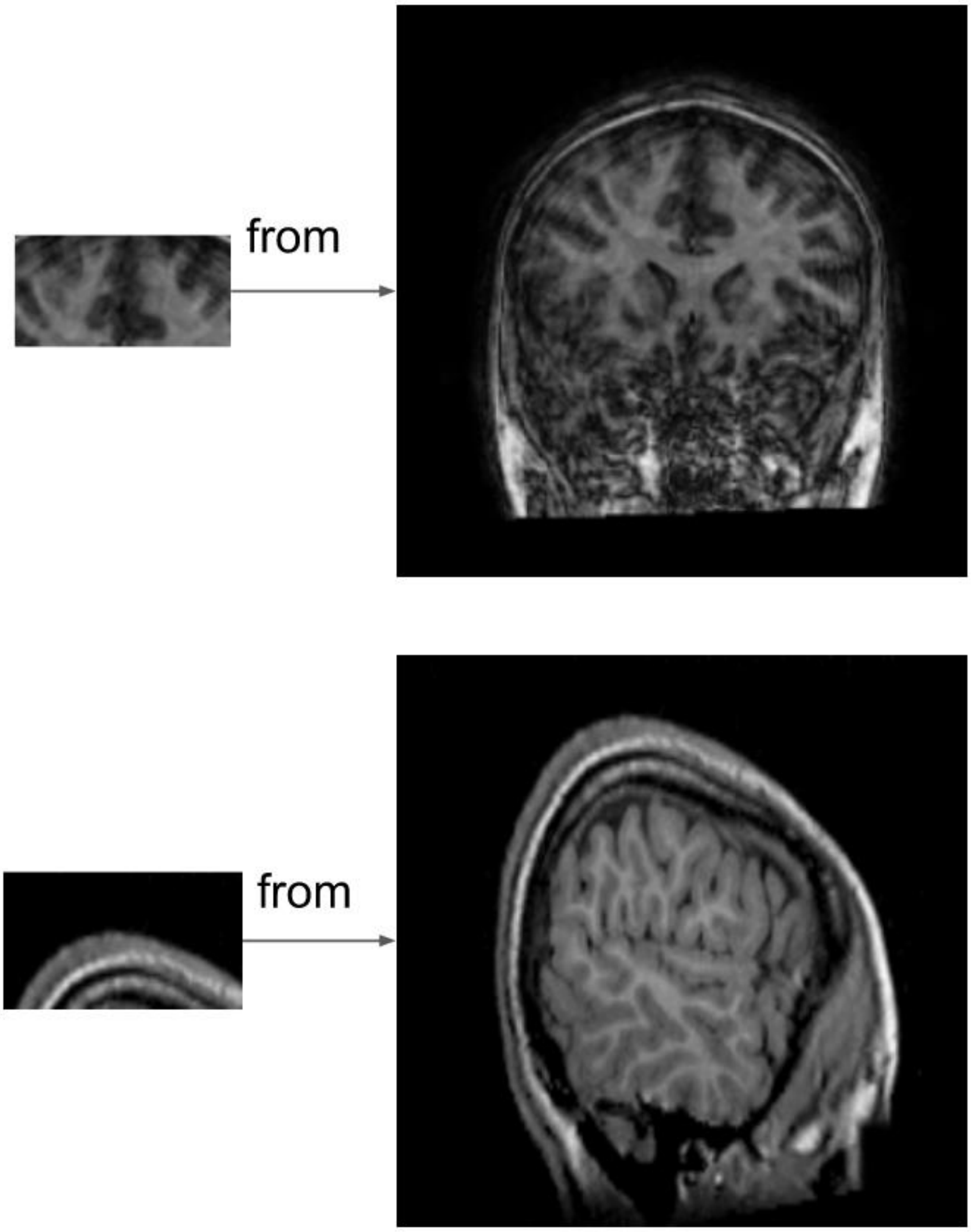


Figure A1.4: Prototypes that commonly appeared in the top-5 prototypes (i.e., those prototypes with the top 5 highest ranking similarity scores to the input 2D slices).

8.6.4. A1.4 Examples of activation maps

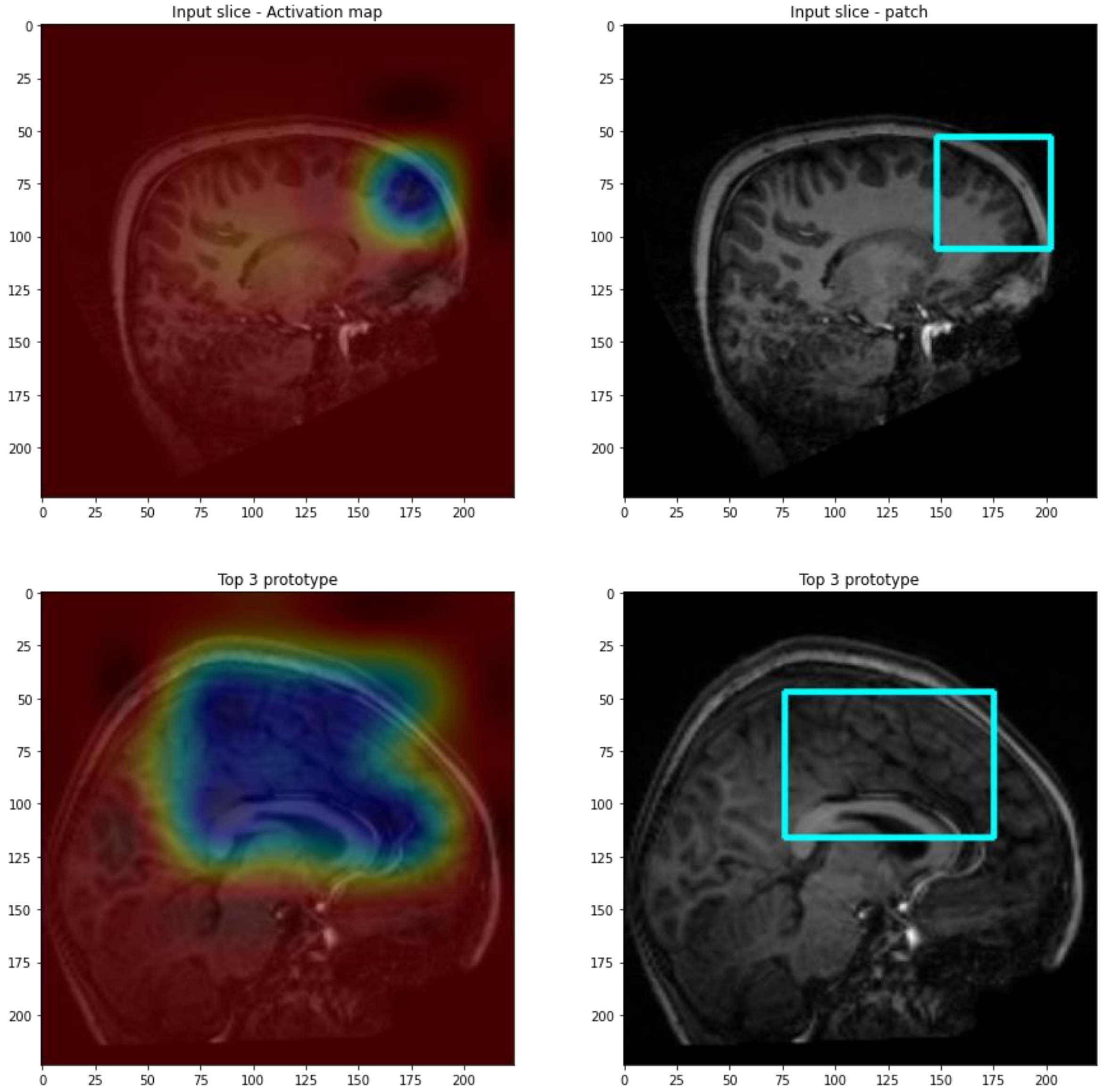


Figure A1.5: Examples of artefact maps and prototypes judged to be meaningful: the upper panel shows the input slice, the lower panel shows the top-3 prototype for the model proto-R152 trained on 10 epochs (i.e., the prototype with the third highest ranking similarity score to the input 2D slices).

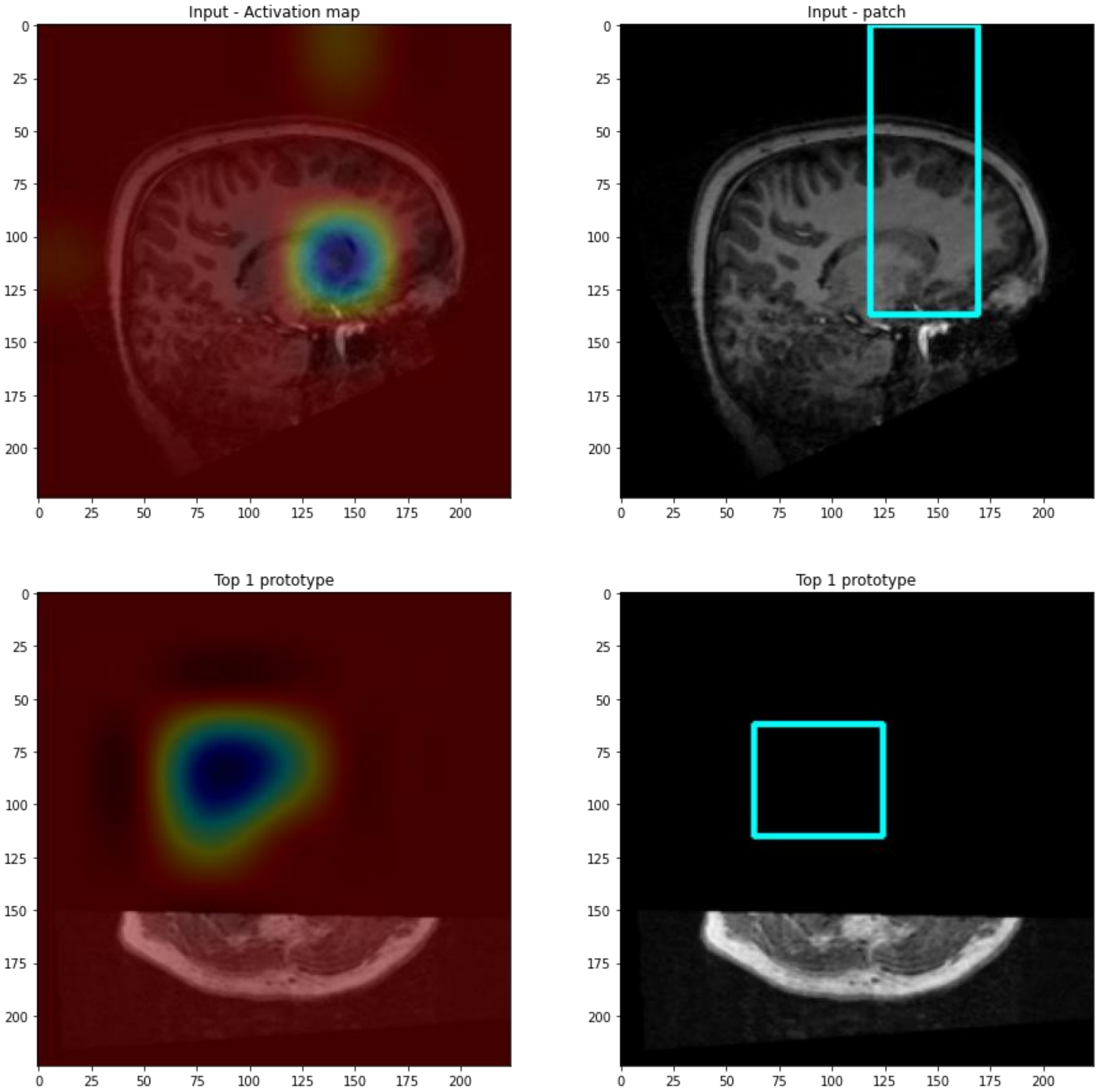


Figure A1.6: Examples of artefact maps and prototypes judged not to be meaningful: the upper panel shows the input slice, the lower panel shows the top-1 prototype for the model proto-V19 trained on 30 epochs (i.e., the prototypes with the highest ranking similarity score to the input 2D slices).

8.6.5. A1.5 Mann-Whitney U-tests between the predicted scores of QC categories of the ABCD data

variable 1	variable 2	U-val	alternative	p-val	RBC	CLES
pass	**fail**	13814,5	two-sided	1,844.10e^{-93}	0,899	0,051
pass	**questionable**	208800,0	two-sided	1,183.10e^{-56}	0,459	0,271
fail	**questionable**	94121,5	two-sided	1,004.10e^{-48}	-0,701	0,850

Table A1.2: Mann-Whitney U-tests between the scores returned by BrainQCNet on each QC category of the ABCD dataset: pass, questionable, fail. It shows that the distribution of scores are well distinct between each other, which is an expected behaviour for our QC algorithm.

Appendix 2: Supplemental Information for Manuscript: “Towards 3D Deep Learning for neuropsychiatry: predicting Autism diagnosis using an interpretable Deep Learning pipeline applied to minimally processed structural MRI data”

8.6.6. A2.1 - Detailed data description

Dataset	Data-collecting site	No comorbidity Train set (1074 sub.)	 Validation set (190 sub.)	 Testing set (65 sub.)	At least one comorbidity Testing set 2 (270 sub.)
ABIDE I	CALTECH	Total: 30 (Autism: 16; No Autism: 14)	Total: 6 (Autism: 2; No Autism: 4)		
	CMU	Total: 25 (Autism: 13; No Autism: 12)	Total: 2 (Autism: 1; No Autism: 1)		
	KKI	Total: 22 (Autism: 6; No Autism: 16)	Total: 3 (Autism: 0; No Autism: 3)		Total: 9 (Autism: 7; No Autism: 2)
	LEUVEN_1			Total: 27 (Autism: 14 ; No Autism: 13)	
	LEUVEN_2	Total: 28 (Autism: 12; No Autism: 16)	Total: 3 (Autism: 1; No Autism: 2)		
	MAX_MUN	Total: 30 (Autism: 10; No Autism: 20)	Total: 3 (Autism: 3; No Autism: 0)		

NYU	Total: 103 (Autism: 27 ; No Autism: 76)	Total: 21 (Autism: 5; No Autism: 16)		Total: 28 (Autism: 28; No Autism: 0)
OHSU	Total: 19 (Autism: 10; No Autism: 9)	Total: 3 (Autism: 1; No Autism: 2)		
OLIN	Total: 12 (Autism: 8; No Autism: 4)	Total: 7 (Autism: 4; No Autism: 3)		
PITT	Total: 31 (Autism: 15; No Autism: 16)	Total: 8 (Autism: 4; No Autism: 4)		
SBL	Total: 26 (Autism: 13; No Autism: 13)	Total: 3 (Autism: 1; No Autism: 2)		
SDSU	Total: 10 (Autism: 1; No Autism: 9)	Total: 2 (Autism: 1; No Autism: 1)		
STANFORD	Total: 5 (Autism: 1; No Autism: 4)	Total: 2 (Autism: 2; No Autism: 0)		
TRINITY	Total: 36 (Autism: 18; No Autism: 18)	Total: 6 (Autism: 2; No Autism: 4)		

	UCLA_1	Total: 45 (Autism: 27; No Autism: 18)	Total: 7 (Autism: 5; No Autism: 2)		
	UCLA_2	Total: 13 (Autism: 6; No Autism: 7)	Total: 2 (Autism: 0; No Autism: 2)		
	UM_1	Total: 59 (Autism: 21; No Autism: 38)	Total: 13 (Autism: 6; No Autism: 7)		
	UM_2	Total: 28 (Autism: 12; No Autism: 16)	Total: 3 (Autism: 0; No Autism: 3)		
	USM	Total: 54 (Autism: 35; No Autism: 19)	Total: 12 (Autism: 8; No Autism: 4)		
	YALE	Total: 46 (Autism: 22; No Autism: 24)	Total: 4 (Autism: 2; No Autism: 2)		
ABIDE II	BNI	Total: 8 (Autism: 7; No Autism: 1)	Total: 1 (Autism: 0; No Autism: 1)		
	EMC			Total: 18 (Autism: 4; No Autism: 14)	Total: 9 (Autism: 9; No Autism: 0)

ETH	Total: 25 (Autism: 7; No Autism: 18)	Total: 5 (Autism: 1; No Autism: 4)		
GU	Total: 56 (Autism: 19; No Autism: 37)	Total: 9 (Autism: 2; No Autism: 7)		
IP	Total: 29 (Autism: 7; No Autism: 22)	Total: 7 (Autism: 4; No Autism: 3)		Total: 8 (Autism: 5; No Autism: 3)
IU	Total: 31 (Autism: 15; No Autism: 16)	Total: 2 (Autism: 1; No Autism: 1)		
KKI	Total: 103 (Autism: 1; No Autism: 102)	Total: 20 (Autism: 0; No Autism: 20)		Total: 36 (Autism: 32; No Autism: 4)
KUL	Total: 12 (Autism: 12; No Autism: 0)	Total: 8 (Autism: 8; No Autism: 0)		Total: 5 (Autism: 5; No Autism: 0)
NYU 1	Total: 36 (Autism: 12; No Autism: 24)	Total: 5 (Autism: 2; No Autism: 3)		Total: 22 (Autism: 22; No Autism: 0)
NYU 2	Total: 5 (Autism: 5; No Autism: 0)	Total: 1 (Autism: 1; No Autism: 0)		Total: 15 (Autism: 15; No Autism: 0)

	OHSU	Total: 47 (Autism: 11; No Autism: 36)	Total: 8 (Autism: 0 ; No Autism: 8)		Total: 34 (Autism: 24; No Autism: 10)
	SDSU	Total: 51 (Autism: 28; No Autism: 23)	Total: 4 (Autism: 3; No Autism: 1)		
	TCD	Total: 29 (Autism: 13; No Autism: 16)	Total: 7 (Autism: 3; No Autism: 4)		
	UCD			Total: 20 (Autism: 8; No Autism: 12)	Total: 5 (Autism: 5; No Autism: 0)
	USM	Total: 20 (Autism: 11; No Autism: 9)	Total: 3 (Autism: 1; No Autism: 2)		
ADHD200	Peking				Total: 23 (Autism: 0; No Autism: 23)
	KKI				Total: 10 (Autism: 0; No Autism: 10)
	NeuroIMAGE				Total: 22 (Autism: 0; No Autism: 22)

NYU	Total: 65 (Autism: 2; No Autism: 63)
OHSU	Total: 20 (Autism: 0; No Autism: 20)

Table A2.1: Partition of ABIDE I, ABIDE II, and ADHD200 into training, validation and testing sets.

	Gender	Age		FIQ	
Train	Males: 853	mean	17.159562	mean	110.290806
		std	8.656338	std	14.888248
		min	5.128000	min	41.000000
		25%	11.005000	25%	101.000000
		50%	14.653000	50%	111.000000
		75%	20.100000	75%	121.000000
		max	64.000000	max	149.000000
	Females: 221	mean	15.026466	mean	111.308458
		std	8.035651	std	14.835831
		min	5.220000	min	66.000000
		25%	9.789041	25%	101.000000
		50%	12.361644	50%	113.000000

		75% 16.800000	75% 122.000000
		max 54.000000	max 146.500000
Validation	Males: 153	mean 17.012265	mean 110.043750
		std 8.623991	std 15.436532
		min 7.150000	min 49.000000
		25% 11.262800	25% 100.750000
		50% 14.800000	50% 112.000000
		75% 20.166667	75% 119.250000
		max 64.000000	max 147.500000
	Females: 37	mean 13.046654	mean 113.972222
		std 5.848126	std 14.624317
		min 5.907000	min 84.000000
		25% 9.665753	25% 105.750000
		50% 10.780822	50% 115.000000
		75% 14.060000	75% 123.000000
		max 32.000000	max 149.000000
Test 1	Males: 57	mean 17.087350	mean 109.976190
(no comorbidity)		std 6.428793	std 12.994348
		min 7.129363	min 83.000000
		25% 10.663929	25% 101.500000
		50% 17.416667	50% 108.500000

		75%	22.000000	75%	118.250000
		max	32.000000	max	146.000000
	Females: 8	mean	12.005540	mean	113.200000
		std	4.022715	std	14.411801
		min	6.395619	min	92.000000
		25%	8.400411	25%	105.000000
		50%	13.500000	50%	120.000000
		75%	14.520833	75%	122.000000
		max	16.500000	max	127.000000
Test 2	Males: 205	mean	11.946206	mean	107.235632
(with comorbidities)		std	5.006252	std	15.966971
		min	5.598000	min	69.000000
		25%	8.646575	25%	97.250000
		50%	10.870000	50%	108.000000
		75%	13.200000	75%	116.000000
		max	35.000000	max	148.000000
	Females: 65	mean	11.973690	mean	107.087719
		std	5.715843	std	12.884488
		min	5.819000	min	74.000000
		25%	9.000000	25%	98.000000
		50%	10.260000	50%	109.000000

75%	12.580000	75%	115.000000
max	38.760000	max	132.000000

Table A2.2: Gender breakdown and distribution of age and FIQ score for each dataset (training, validation, testing, testing 2 sets).

8.6.7. A2.2 - Model architectures

Layers	Output Size	DenseNet121
Convolution	128 x 128 x 128	7 x 7 x 7 conv, stride 2
Pooling	64 x 64 x 64	3 x 3 x 3 pool, stride 2
DenseBlock 1	64 x 64 x 64	1 x 1 x 1 conv → 3 x 3 x 3 conv
		x 6
Transition Layer 1	64 x 64 x 64	1 x 1 x 1 conv
	32 x 32 x 32	2 x 2 x 2 average pool, stride 2
DenseBlock 2	32 x 32 x 32	1 x 1 x 1 conv → 3 x 3 x 3 conv
		x 12

Transition Layer 2	32 x 32 x 32	1 x 1 x 1 conv
	16 x 16 x 16	2 x 2 x 2 average pool, stride 2
DenseBlock 3	16 x 16 x 16	1 x 1 x 1 conv → 3 x 3 x 3 conv x 24
Transition Layer 3	16 x 16 x 16	1 x 1 x 1 conv
	8 x 8 x 8	2 x 2 x 2 average pool, stride 2
DenseBlock 4	8 x 8 x 8	1 x 1 x 1 conv → 3 x 3 x 3 conv x 16
Classification Layer	1 x 1 x 1	8 x 8 x 8 global average pool
		Fully Connected layer, softmax

Table A2.3: Representation of DenseNet121 for our classification task of 3D scans - Input size: 256 x 256 x 256

Layers	Output Size	ResNet50
Convolution	128 x 128 x 128	7 x 7 x 7 conv, stride 2
Max Pooling	64 x 64 x 64	3 x 3 x 3 pool, stride 2
Convolutional Layer (type 1)	64 x 64 x 64	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv -> 1 x 1 x 1 conv
Bottleneck	64 x 64 x 64	
Convolutional Layer (type 2)	64 x 64 x 64	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv
Bottleneck	64 x 64 x 64	
Convolutional Layer (type 2)	64 x 64 x 64	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv
Bottleneck	64 x 64 x 64	
Convolutional Layer (type 1)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv -> 1 x 1 x 1 conv
Bottleneck	32 x 32 x 32	
Convolutional Layer (type 2)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv
Bottleneck	32 x 32 x 32	
Convolutional Layer (type 2)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv

	Bottleneck	32 x 32 x 32	
	Convolutional Layer (type 2)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv
	Bottleneck	32 x 32 x 32	
	Convolutional Layer (type 1)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv -> 1 x 1 x 1 conv
	Bottleneck	32 x 32 x 32	
	Convolutional Layer (type 2)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv
	Bottleneck	32 x 32 x 32	
	Convolutional Layer (type 2)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv
	Bottleneck	32 x 32 x 32	
	Convolutional Layer (type 2)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv
	Bottleneck	32 x 32 x 32	
	Convolutional Layer (type 2)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv
	Bottleneck	32 x 32 x 32	

🟩	Convolutional Layer (type 1)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv -> 1 x 1 x 1 conv
	Bottleneck	32 x 32 x 32	
	Convolutional Layer (type 2)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv
	Bottleneck	32 x 32 x 32	
	Convolutional Layer (type 2)	32 x 32 x 32	1 x 1 x 1 conv -> 3 x 3 x 3 conv -> 1 x 1 x 1 conv
	Bottleneck	32 x 32 x 32	
	Classification Layer	1 x 1 x 1	7 x 7 x 7 global average pool Fully Connected layer, softmax

Table A2.4: Architecture of ResNet50 - in yellow, the layers for which we extracted the parameters from the pre-trained model Med3d; in light green, the layers for which we continued training the parameters to fine-tune the model and adapt it to the task of predicting Autism.

8.6.8. A2.3 - Performance of the models

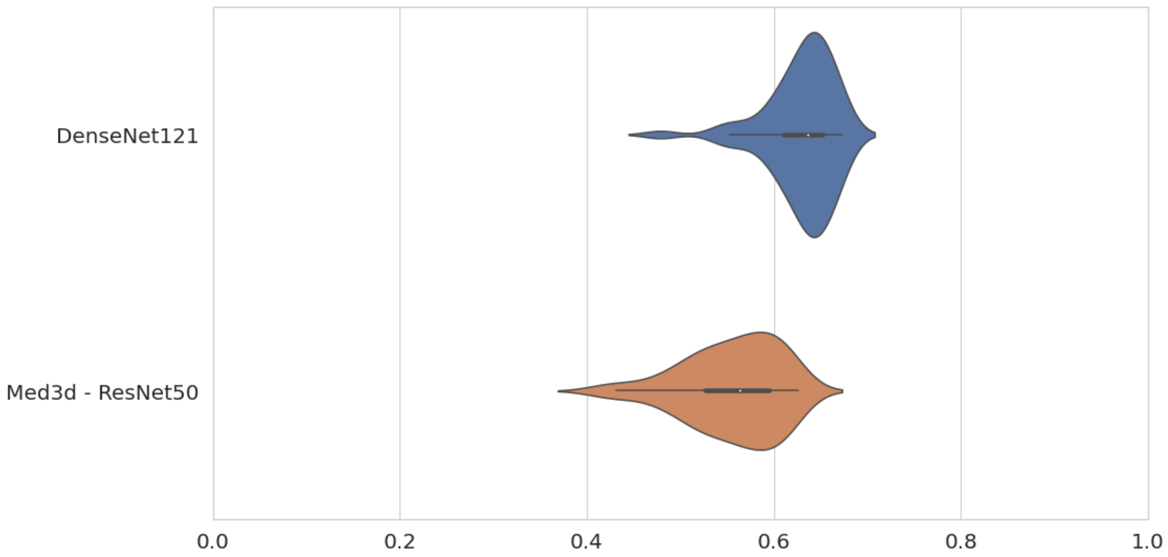


Figure A2.1: Validation set accuracy during training for the two models DenseNet161 and Med3d-ResNet50.

Figure A2.1 compares the distributions of validation set accuracies for each model. DenseNet121 tended to have more sSupplemental Table and higher accuracies on the validation set than Med3d-ResNet50.

	Med3d - ResNet50 - 42 epochs			DenseNet121 - 32 epochs			DenseNet121 - 70 epochs		
Subjects	All	Autism	no-Autism	All	Autism	no-Autism	All	Autism	no-Autism
Training set	Acc.: 94,2 % AUROC: 99,9 %	Acc.: 85,3 %	Acc.: 100 %	Acc.: 65,5 % AUROC: 69,1 %	Acc.: 32,8 %	Acc.: 86,7%	Acc.: 69,7 % AUROC: 77,1 %:	Acc.: 68,2 %	Acc.: 70,8 %
Validation set	Acc.: 62,6 % AUROC: 62,1 %	Acc.: 17,6 %	Acc.: 91,4 %	Acc.: 66,3 % AUROC: 68,8 %	Acc.: 36,5 %	Acc.: 85,3 %	Acc.: 67,4 % AUROC: 68,1 %	Acc.: 66,2 %	Acc.: 68,1 %
Testing set	Acc.: 53,8 % AUROC: 57,3 %	Acc.: 50%	Acc.: 56,4 %	Acc. 55,4 %: AUROC: 60,7 %	Acc.: 84,6 %	Acc.: 35,9 %	Acc.: 40 % AUROC: 38,1 %	Acc.: 69,2 %	Acc.: 20,5 %
All the dataset	Acc.: 87,7 % AUROC: 95,5 %			Acc.: 65,2 % AUROC: 68,4 %			Acc.: 67,9 % AUROC: 74,0 %:		

Table A2.5: Comparison of the performance of the prediction of Autism between the models Med3d - ResNet50 - 42 epochs, DenseNet121 - 32 epochs and DenseNet121 - 70 epochs.

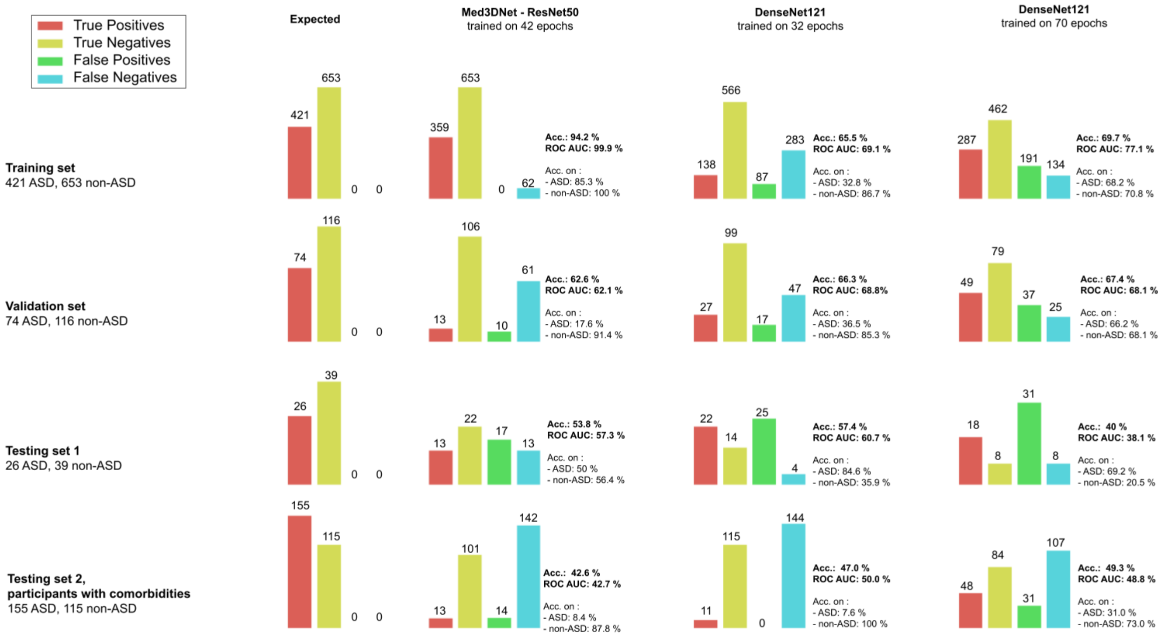


Figure A2.2: True and False Positives and Negatives for each of the three best models - Med3DNet-ResNet50 trained on 42 epochs, DenseNet121 trained on 32 epochs, and DenseNet121 trained on 70 epochs.

Figure A2.2 shows the accuracies (in terms of True/False Positives and Negatives) obtained for each of the three best models for prediction of Autism (Autism vs. non-Autism) and each dataset. We can see that Med3d-ResNet50-42ep overfit the data, because the accuracy and ROC AUC scores were very high on the training set (94.2% and 99.9% respectively), but much lower on the validation (acc = 62.6% and AUC = 62.1%) and testing sets (acc = 53.8% and AUC=57.3%). DenseNet121-32ep appeared to be more sSupplemental Table in terms of its overall performance on the training (acc = 65.5% and AUC = 69.1%), validation (acc =66.3% and AUC = 68.8%) and testing (acc =55.4% and AUC = 60.7%) sets. DenseNet121-70ep had better performance on the training (acc = 69.7% and AUC = 77.1%) and validation (acc = 67.4% and AUC = 68.1%) sets than DenseNet121-32ep, but poorer performance on the testing set (acc = 40% and AUC = 38.1%).

Figure A2.2 shows that DenseNet121-32ep has high specificity on the training and validation sets, while having low sensitivity. Paradoxically, it has high sensitivity but low specificity on the testing set. DenseNet121-70ep behaves similarly on the testing set. Nevertheless, on the training and validation sets, we can see that the sensitivity and specificity are balanced and fairly high. Finally, for Med3d-ResNet50-42ep, we observe that the sensitivity and specificity are very high on the training set, are unbalanced on the validation set with a low sensitivity and very high specificity, and are balanced again on the testing set, but with moderate values.

The lowest panel of **Figure A2.2** shows the accuracies for the second testing set, which included participants with comorbidities. The data show that predicting Autism in the presence of comorbidities is more difficult than predicting Autism when the training and testing sets include only participants without known comorbidities, with a particularly large increase in False Negatives. One potential explanation is that neuroimaging markers become less evident when individuals have another diagnosis involving similar or other neuroimaging markers. Another explanation is that more data are needed to adequately train DL algorithms on the whole spectrum of Autism patients.

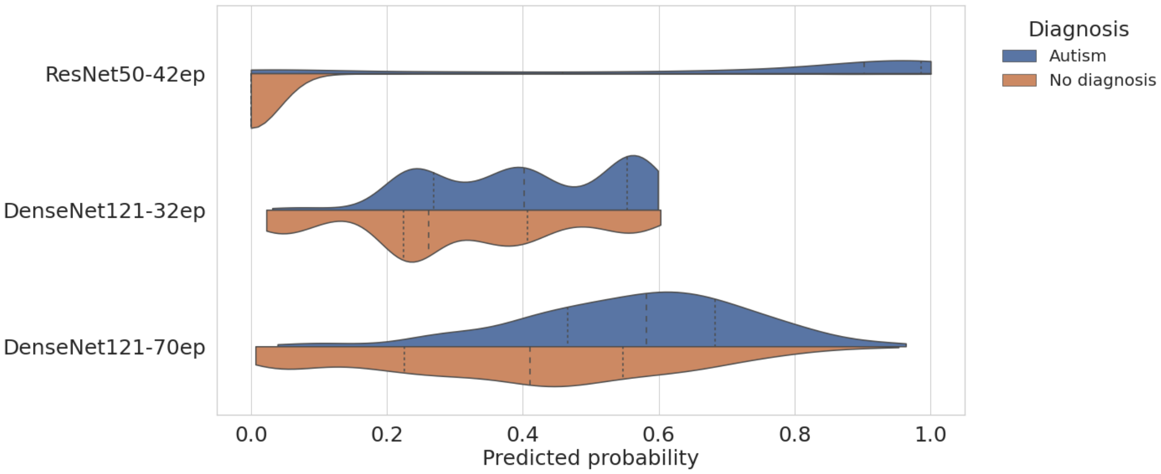


Figure A2.3: Comparison of model predictions across all the datasets without comorbidity (training/validation/test)

Figure A2.3 shows that there is a net difference in the distribution of probabilities for Autistic vs non-Autistic participants for the Med3d-ResNet50-42ep model, in line with the observation of overfitting and the very good performance observed for the training set (1074 subjects). For the two other models, the estimated means are distinct, although the distributions overlap. This observation also reflects the accuracy and ROC AUC scores obtained with these two models.

T-tests indicate no significant difference between the age of patients predicted with Autism and the ones predicted with no Autism (p > .05).

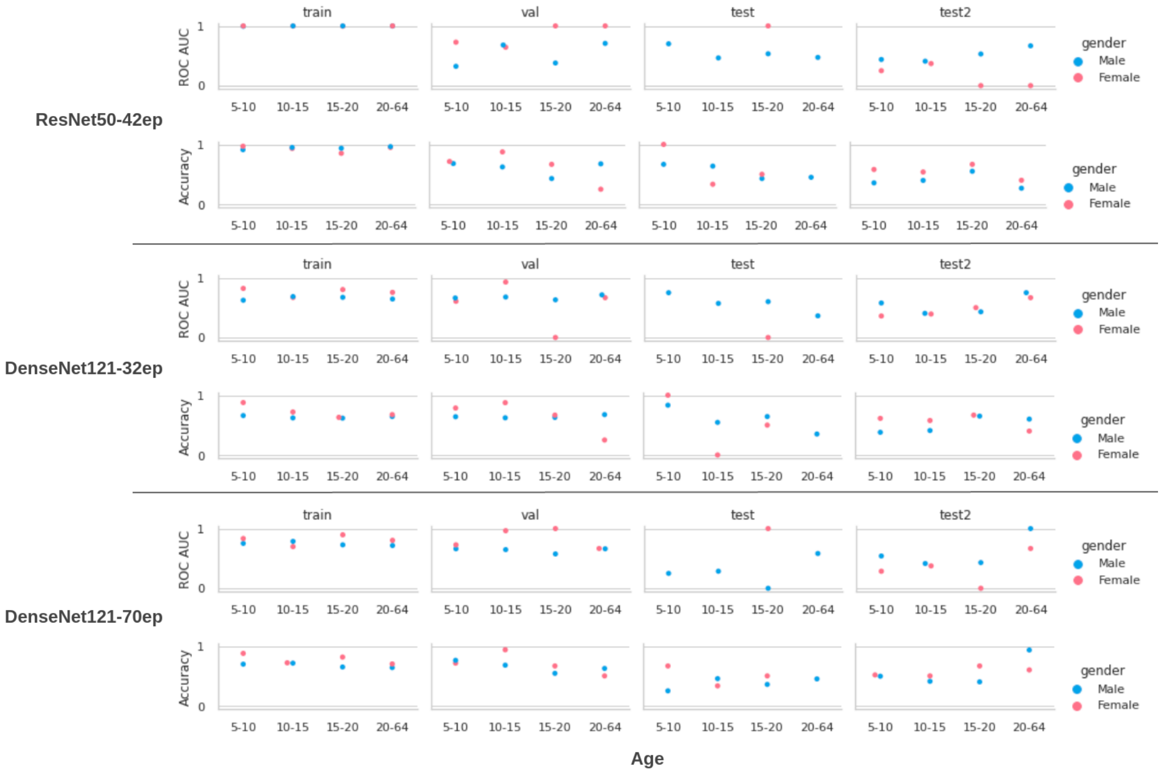


Figure A2.4: ROC AUC and accuracy scores in function of age (between 5 and 10, 10 and 15, 15 and 20, 20 and 64) and gender (male or female) for each model (ResNet50 trained on 42 epochs, DenseNet121 trained on 32 epochs and trained on 70 epochs) for each dataset (training, validation, testing and testing 2 sets).

We observed that the ROC AUC and accuracy scores did not differ between age ranges and between genders in the training set. However, we observed that in the validation set, these scores were variable (see **Figure A2.4**). We also observed this variation in the two testing sets (see **Figure A2.4**). This suggests that we should examine the stability of performance between different age ranges and between males and females.

8.6.9. A2.4 - Analysis of ADI-R and ADOS scores, age, gender and full IQ

To better understand differences between the datasets (training, validation and testing sets) and between the classes (Autism and non-Autism), we performed an analysis incorporating the severity scores from ADI-R and ADOS, the age, the gender and the Full IQ scores.

First, we gathered all the information on the diagnosis available in ABIDE I & II. By combining various questionnaires (ADI-R, ADOS Modules 2, 3 and 4), we obtained scores for (1) social interaction (including the Reciprocal Social Interaction Subscore A for ADI-R, the Social Total Subscore of the classic ADOS, the Social Affect Total Subscore for Gotham Algorithm of ADOS, for (2) verbal communication (including the Abnormalities in Communication Subscore (B) for ADI-R, the Communication Total Subscore of the Classic ADOS), (3) for repetitive, restricted or stereotyped behaviours (including the Restricted, Repetitive, and Stereotyped Patterns of Behaviour Subscore (C) for ADI-R, the Stereotyped Behaviours and Restricted Interests Total Subscore of the Classic ADOS, the Restricted and Repetitive Behaviours Total Subscore for Gotham Algorithm of ADOS) and (4) total scores (including the Abnormality of Development Evident at or before 36 months Subscore (D) Total for ADI-R, the Classic ADOS Score, the Gotham Algorithm of ADOS Score) for 452 subjects. Since all of these questionnaires use different scales, we transformed all the scores into Z-scores to compare individuals.

Second, we gathered the predicted class of each patient from each model, and, from it, we created a variable "prediction type" representing the True Positives, False Negatives, True Negatives and False Positives.

Finally, we compared the distributions of Z-scores across prediction types, to investigate whether there were differences in symptom severity scores between True Positives and False Negatives, and similarly, a difference between True Negatives and False Positives. We visually compared the distribution and performed a T-test

These analyses did not reveal any discernible differences between the predictions of the three models. **Figure A2.5** illustrates an example of this analysis using Med3d - ResNet50 - 42ep.

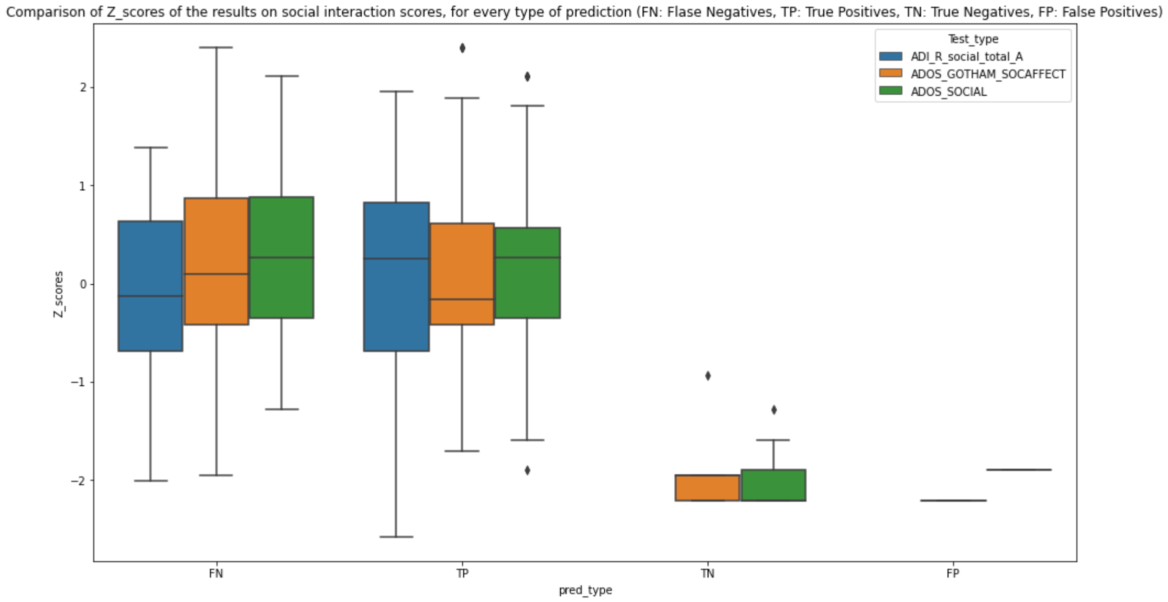


Figure A2.5: Comparison of social interaction Z-scores between False Negatives (FN), True Positives (TP), True Negatives (TN) and False Positives (FP).

We did not find any differences between the three models when we examined the severity scores on the training, validation and testing sets separately. Nor were there differences between males and females in the distribution of probabilities for all the models.

We compared the distribution of age for each prediction type (FN, TP, TN, FP). There was no noticeable difference in age between the samples corresponding respectively to each prediction type for all the models, compared to the distribution of age between the samples corresponding to true labels.

Finally, we also compared the distribution of Full IQ score for each prediction type (FN, TP, TN, FP). There was no noticeable difference between the samples corresponding respectively to each prediction type or all the models, compared to the distribution of FIQ between the samples corresponding to true labels.

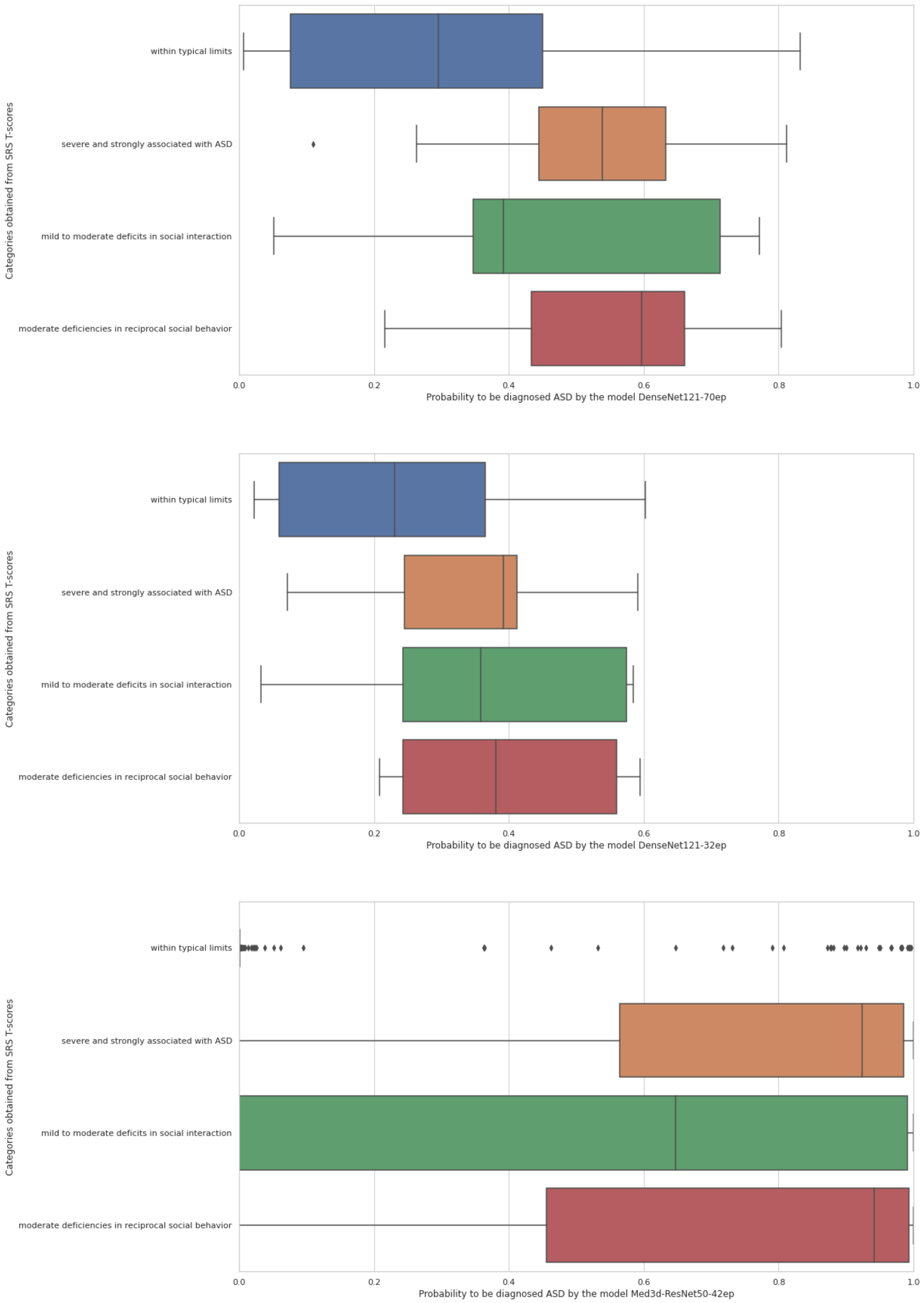


Figure A2.6: Probability scores of each model per category obtained from SRS T-scores in ABIDE 2

In **Figure A2.6**, for every model, the distribution of probability scores is shown for categories created on the basis of the total T-scores of the SRS-2. “Within typical limits” corresponds to a T-score lower than 59, “mild to moderate difficulties in social interaction” corresponds to a T-score between 60 and 65, “moderate difficulties in reciprocal social behaviour” corresponds to a T-score between 66 and 75, and “severe difficulties, strongly associated with Autism” to a T-score greater than 76. We observed that DenseNet121-70ep had a distribution of probability scores that was consistent with these severity scores, with the majority of scores lower than 0.5 for the category “within typical limits”, and the majority of scores greater than 0.5 for the three other categories.

8.6.10. A2.5 - Most important regions for the prediction of True Positives

			R42		D32		D70	
			No comorb	With comorb	No comorb	With comorb	No comorb	With comorb
Frontal lobe	Left	Frontal operculum	2	2	0	0	2	2
Frontal lobe	Left	Middle frontal gyrus	3	4	0	0	0	0
Frontal lobe	Left	Precentral gyrus medial segment	0	0	3	4	1	2
Frontal lobe	Left	Precentral gyrus	1	1	0	0	2	3

Frontal lobe	Left	Triangular part of the inferior frontal gyrus	3	4	0	0	1	1
Limbic system and associated structures	Left	Anterior Cingulate Gyrus	3	4	0	0	0	0
Limbic system and associated structures	Left	Parahippocampal gyrus	1	2	0	0	3	4
Limbic system and associated structures	Left	Subcallosal area	3	4	0	0	0	1
Parietal lobe	Left	Central operculum	2	2	0	0	1	2
Parietal lobe	Left	Parietal operculum	0	1	2	3	2	3
Parietal lobe	Left	Parietal white matter	0	0	2	2	2	3
Parietal lobe	Left	Supplementary motor cortex	0	0	1	2	1	2
Parietal lobe	Left	Supramarginal gyrus	0	0	2	3	2	3
occipital lobe	Left	Posterior orbital gyrus	2	3	0	0	2	2

subcortical structures	Right	Ventral diencephalon	0	1	2	3	1	1
temporal lobe	Left	Planum temporale	1	1	1	2	2	3
temporal lobe	Left	Superior temporal gyrus	2	2	0	0	2	2
temporal lobe	Left	Temporal pole	2	2	0	0	3	3
temporal lobe	Left	Transverse temporal gyrus	1	1	0	1	1	2

Table A2.6: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism): each row is for one region, each column is for one model (R42 for ResNet50 trained on 42 epochs, D32 for DenseNet121 trained on 32 epochs, D70 for DenseNet121 trained on 70 epochs) and one combination of datasets considered (training+validation+testing 1 sets ("no comorb" for no comorbidity), or all these sets + testing set 2 ("with comorb" for containing subjects with comorbidities), each case returns the number of datasets where the region was important for predicting TP for the model considered.

Table A2.6 summarises the most important regions for predicting TP, which identified regions that were common across models. Limiting the results in this way enables us to reduce the bias effect of each algorithm (that leads to regions important only for one model). We used a methodology analogous to a traditional machine learning pipeline here, which identified features on the basis of their importance.

Table A2.6 gives us different types of information:

- The most replicable regions for predicting TP between the models

- The replicability of the regions found between the datasets not containing subjects with comorbidity (thanks to the number in each case in the columns “Training + val. + testing 1 sets”)
- The replicability of the regions found between datasets without comorbidity and dataset with comorbidity (thanks to the number in each case in the columns “Training + val. + testing 1 & 2 sets”). This is also shown by the changes highlighted in light red.

For instance, for the model Med3D-ResNet50 trained on 42 epochs, we found that Right-ACgG-anterior-cingulate-gyrus is an important region for predicting TP on three over the three datasets into the datasets without comorbidity, and on four over the four datasets “Training + val. + testing 1 & 2 sets” that contains subjects with comorbidities in testing set 2. Thus, for the model Med3D-ResNet50 trained on 42 epochs, Right-ACgG-anterior-cingulate-gyrus is important for the prediction of TP, and, by extension, for the detection of Autism, and was robust to comorbidities.

In **Table A2.6**, we observed that several regions were important for the three models, including Left Planum Temporale, Left Parietal Operculum, Right Ventral Diencephalon. However, we saw that for the four regions the replicability is low between the datasets.

We also noticed that, on the one hand, a lot of regions were commonly important for the two DenseNet models but not for ResNet50, including Left Supramarginal Gyrus, Left Parietal White Matter and Left precentral gyrus medial segment. On the other hand, Left subcallosal area, Left Middle Frontal gyrus, Left-MFC-medial-frontal-cortex and Left anterior cingulate gyrus were important for ResNet50 but not for the two DenseNet models, and their importance replicated well over the datasets, including the one with comorbidities.

Further, several regions were important to both ResNet50 and to DenseNet121-70ep, including Left triangular part of the inferior frontal gyrus, Left Temporal Pole , Left Precentral Gyrus, Left posterior orbital gyrus, Left Parahippocampal gyrus. We noticed that for Left triangular part of the inferior frontal gyrus, the replicability over the datasets without comorbidity was higher for the ResNet50 model than for the DenseNet121-70ep model, while we observed the opposite for Left Parahippocampal gyrus. However, we noticed that for the ResNet50 model, the importance of Left Precentral Gyrus did not replicate in the testing set 2 with comorbidities whereas for DenseNet121-70ep it did. The converse was observed for Left posterior orbital gyrus. This disparity is illustrative of the bias introduced by each model, due to the different architectures and levels of optimisation. Even though we set the optimiser parameters similarly between the models, due to the inherent difference in the designs, the models tend to approximate a function that achieves the task of detecting Autism in different ways. This also underlines the importance of considering different types of models in deep learning when possible (machine/funding limitation), analogously to more traditional machine learning pipelines of analysis.

With regard to participants with comorbidities, we see from **Table A2.6** that all the regions important for ResNet50-42ep, but which were not shared with the other models, replicated well in the test set with comorbidities. Globally, the models ResNet50-42ep and DenseNet70-70ep have an equivalent number of areas that were important for predicting TP and which replicated well in the testing set 2, higher than for the model DenseNet121-32ep.

Another interesting point is that certain regions that were not among the most important for predicting TP in the datasets without comorbidities appear important for predicting TP in the dataset with subjects who did have comorbidities. This includes Left subcallosal area for DenseNet121-70ep, and Right Ventral Diencephalon, Left Parietal Operculum for ResNet50-42ep.

Summarising the Supplemental Table, and taking each model separately, the most important regions for predicting Autism across all the models and between datasets are Left triangular part of the inferior frontal gyrus, Left subcallosal area, Left Parahippocampal gyrus, Left precentral gyrus medial segment, Left Middle Frontal gyrus and Left anterior cingulate gyrus.

On the one hand, this result can help us identify neuroimaging markers of Autism, by combining the findings between the models, using deep learning as a way to extract feature importance in a manner similar to Random Forest, for instance. On the other hand, this shows that each model tends to focus on specific parts in the brain, capturing different patterns than the other models, making it difficult to select one model that works best.

8.6.11. A2.6 - Most important regions for the prediction of True Negatives

Overall, after aggregating all the datasets, among the 79 areas most important for predicting TN, 10 areas combining left and right hemispheres, 24 in the left hemisphere and 2 in the right hemisphere were commonly predictive for TP.

Keeping only the areas that replicated the most over the datasets, the areas predictive for TN were largely different from the ones that were important for TP.

			R42		**D32**		**D70**	
			No comorb	**With comorb**	**No comorb**	**With comorb**	**No comorb**	**With comorb**
Frontal lobe	Left	Frontal operculum	3	4	0	0	0	0
Limbic system and associated structures	Left	Posterior cingulate gyrus	1	1	2	3	1	1

Limbic system and associated structures	Right	Posterior cingulate gyrus	0	0	2	3	1	1
Parietal lobe	Left	Precuneus	0	0	2	3	1	2
Parietal lobe	Left	Superior parietal lobule	0	0	2	3	2	3
cerebellum	None	Vermal Lobules VI-VII	0	0	3	4	2	2
cerebellum	Left	Cerebellum exterior	0	0	2	3	2	2
occipital lobe	Left	Angular gyrus	0	0	2	3	2	3
occipital lobe	Left	Calcarine cortex	0	0	0	1	2	3
occipital lobe	Left	Cuneus	0	0	1	1	2	3
occipital lobe	Left	Inferior occipital gyrus	0	0	3	4	2	3

occipital lobe	Left	Lingual gyrus	0	0	2	3	2	3
occipital lobe	Left	Middle occipital gyrus	0	0	2	2	2	3
occipital lobe	Left	Occipital fusiform gyrus	0	0	3	4	2	3
occipital lobe	Left	Occipital White Matter	0	0	2	3	2	3
occipital lobe	Left	Superior occipital gyrus	0	0	1	2	3	4
subcortical structures	Left	Thalamus	3	4	0	0	0	0
subcortical structures	Right	Ventral diencephalon	2	3	0	0	1	1
temporal lobe	Left	Middle temporal gyrus	0	0	2	2	2	2
temporal lobe	Left	Planum polare	3	4	0	0	0	0

Table A2.7: Best regions for predicting True Negatives (TN, i.e. no diagnosis of Autism): each row is for one region, each column is for one model and one combination of datasets considered (training+validation+testing 1 sets (no comorbidity), or all these sets + testing set 2 (containing subjects with comorbidities)), each case returns the number of datasets where the region was important for predicting TN for the model considered.

8.6.12. A2.7- Most replicable regions for False Positives and False Negatives

			R42		D32		D70	
			No comorb	**With comorb**	**No comorb**	**With comorb**	**No comorb**	**With comorb**
Frontal lobe	Left	Frontal operculum	1	1	0	0	3	3
Frontal lobe	Left	Middle frontal gyrus	1	2	0	0	2	3
Frontal lobe	Left	Precentral gyrus medial segment	0	0	2	2	2	3
Frontal lobe	Left	Precentral gyrus	1	2	0	0	3	3
Frontal lobe	Left	Triangular part of the inferior frontal gyrus	1	1	0	0	3	3

Frontal lobe	Right	Precentral gyrus medial segment	0	0	3	3	0	1
Parietal lobe	Left	Parietal operculum	1	1	2	2	1	1
Parietal lobe	Left	Parietal white matter	0	0	2	2	2	2
Parietal lobe	Left	Supplementary motor cortex	0	0	3	3	1	2
Parietal lobe	Left	Supramarginal gyrus	0	1	2	2	2	3
Parietal lobe	Left	Superior parietal lobule	0	0	3	3	1	2
occipital lobe	Left	Angular gyrus	0	0	3	3	2	2
occipital lobe	Left	Posterior orbital gyrus	2	2	0	0	3	3
temporal lobe	Left	Postcentral gyrus	0	0	2	2	2	2
temporal lobe	Left	Temporal pole	1	1	0	0	3	3

Table A2.8: Best regions for predicting False Positives (FP, i.e. prediction of Autism whereas no diagnosis Autism): each row is for one region, each column is for one model and one combination of datasets considered (training+validation+testing 1 sets (no comorbidity), or

all these sets + testing set 2 (containing subjects with comorbidities)), each case returns the number of datasets where the region was important for predicting TN for the model considered.

			R42		D32		D70	
			No comorb	With comorb	No comorb	With comorb	No comorb	With comorb
Frontal lobe	Left	Frontal operculum	3	4	0	0	0	0
Limbic system and associated structures	Left	Posterior cingulate gyrus	0	1	2	3	1	2
Limbic system and associated structures	Right	Posterior cingulate gyrus	0	1	2	3	1	1
Parietal lobe	Left	Precuneus	0	0	3	4	1	2
Parietal lobe	Left	Superior parietal lobule	0	0	1	2	1	2
cerebellum	None	Vermal Lobules VII-X	0	0	1	2	1	2

cerebellum	Left	Cerebellum exterior	0	0	2	3	1	2
occipital lobe	Left	Angular gyrus	0	0	2	3	2	3
occipital lobe	Left	Cuneus	0	0	1	1	2	3
occipital lobe	Left	Inferior occipital gyrus	0	0	3	4	2	3
occipital lobe	Left	Lingual gyrus	0	0	2	3	2	3
occipital lobe	Left	Middle occipital gyrus	0	0	2	3	2	3
occipital lobe	Left	Occipital fusiform gyrus	0	0	3	4	1	2
occipital lobe	Left	Occipital White Matter	0	0	2	3	2	3
occipital lobe	Left	Superior occipital gyrus	0	0	1	1	2	3

temporal lobe	Left	Middle temporal gyrus	0	1	2	2	2	2
temporal lobe	Left	Planum polare	3	4	0	0	0	0
temporal lobe	Left	Superior temporal gyrus	3	4	0	0	0	0

Table A2.9: Best regions for predicting False Negatives (FN, i.e. no prediction of Autism whereas diagnosed Autism): each row is for one region, each column is for one model and one combination of datasets considered (training+validation+testing 1 sets (no comorbidity), or all these sets + testing set 2 (containing subjects with comorbidities)), each case returns the number of datasets where the region was important for predicting TN for the model considered.

8.6.13. A2.8 - True Positives by Gender

			R42		**D32**		**D70**	
			No comorb	**With comorb**	**No comorb**	**With comorb**	**No comorb**	**With comorb**
Frontal lobe	Left	Frontal operculum	2	2	0	0	2	3
Frontal lobe	Left	Middle frontal gyrus	3	4	0	0	0	0

Frontal lobe	Left	Precentral gyrus medial segment	0	0	3	4	1	2
Frontal lobe	Left	Precentral gyrus	1	1	0	0	3	4
Frontal lobe	Left	Triangular part of the inferior frontal gyrus	3	4	0	0	1	1
Limbic system and associated structures	Left	Anterior Cingulate Gyrus	3	4	0	0	0	0
Limbic system and associated structures	Left	Parahippocampal gyrus	1	2	1	1	3	4
Limbic system and associated structures	Left	Posterior insula	2	2	0	0	2	2
Limbic system and associated structures	Left	Subcallosal area	2	3	0	0	1	2
Parietal lobe	Left	Central operculum	2	2	0	0	2	3

Parietal lobe	Left	Parietal operculum	1	2	1	2	2	3
Parietal lobe	Left	Supramarginal gyrus	0	0	2	3	2	3
occipital lobe	Left	Posterior orbital gyrus	2	3	0	0	2	2
subcortical structures	Right	Ventral diencephalon	0	1	2	3	1	1
temporal lobe	Left	Planum temporale	1	1	1	2	2	3
temporal lobe	Left	Superior temporal gyrus	2	2	0	0	2	2
temporal lobe	Left	Temporal pole	2	2	0	0	3	3
temporal lobe	Left	Transverse temporal gyrus	1	1	1	2	1	2

Table A2.10: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism) for Boys

			R42		D32		D70	
			No comorb	**With comorb**	**No comorb**	**With comorb**	**No comorb**	**With comorb**
Frontal lobe	Left	Triangular part of the inferior frontal gyrus	1	2	1	1	1	1
Frontal lobe	Right	Precentral gyrus medial segment	0	0	2	3	0	1
Limbic system and associated structures	Left	Posterior cingulate gyrus	0	0	2	3	2	3
Limbic system and associated structures	Right	Middle cingulate gyrus	0	1	0	1	1	2
Parietal lobe	Left	Parietal operculu m	1	1	1	2	2	3

Parietal lobe	Left	Parietal white matter	0	0	0	1	2	3
Parietal lobe	Left	Supramarginal gyrus	0	0	1	2	2	3
Parietal lobe	Left	Superior parietal lobule	0	0	2	2	3	3
Parietal lobe	Right	Supplementary motor cortex	1	1	0	1	1	2
occipital lobe	Left	Angular gyrus	0	0	2	2	2	2
occipital lobe	Left	Occipital pole	0	0	1	2	1	2
temporal lobe	Left	Postcentral gyrus medial segment	0	1	2	2	2	2
temporal lobe	Left	Postcentral gyrus	0	0	0	1	2	3
temporal lobe	Right	Postcentral gyrus	0	0	2	2	2	2

		medial segment						

Table A2.11: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism) for Girls.

8.6.14. A2.9 - True Positives by Gender and Age

			R42		**D32**		**D70**	
			No comorb	**With comorb**	**No comorb**	**With comorb**	**No comorb**	**With comorb**
Frontal lobe	Left	Middle frontal gyrus	2	2	0	0	2	2
Frontal lobe	Left	Precentral gyrus	2	2	0	0	3	4
Frontal lobe	Left	Triangular part of the inferior frontal gyrus	1	2	0	0	2	3
Limbic system and associated	Left	Hippocampus	2	2	0	0	1	2

structures								
Parietal lobe	Left	Central operculum	2	2	0	0	3	4
Parietal lobe	Left	Supramarginal gyrus	1	2	0	0	1	2
occipital lobe	Left	Posterior orbital gyrus	1	1	0	0	3	4
temporal lobe	Left	Temporal pole	2	2	0	0	2	3

Table A2.12: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism) for Boys aged 5 to 10.

			R42		**D32**		**D70**	
			No comorb	**With comorb**	**No comorb**	**With comorb**	**No comorb**	**With comorb**
Frontal lobe	Left	Frontal White Matter	1	1	0	0	2	3
Frontal lobe	Left	Triangular part of the inferior	3	4	0	0	2	2

		frontal gyrus						
Frontal lobe	Right	Precentral gyrus medial segment	0	0	3	3	1	1
Limbic system and associated structures	Left	Parahippocampal gyrus	0	1	1	2	2	3
Limbic system and associated structures	Left	Subcallosal area	3	4	0	0	0	0
Parietal lobe	Left	Parietal operculum	1	2	0	1	1	2
Parietal lobe	Left	Supramarginal gyrus	0	0	3	4	1	2
Parietal lobe	Left	Superior parietal lobule	0	0	3	3	1	1

occipital lobe	Left	Posterior orbital gyrus	2	3	0	0	3	3
tempora l lobe	Left	Planum temporale	1	1	0	1	1	2
tempora l lobe	Left	Postcentral gyrus	0	0	1	2	1	2
tempora l lobe	Left	Temporal pole	2	3	0	0	3	3
tempora l lobe	Left	Transverse temporal gyrus	2	2	1	2	2	3

Table A2.13: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism) for Boys aged 10 to 15.

			R42		**D32**		**D70**	
			No comorb	**With comorb**	**No comorb**	**With comorb**	**No comorb**	**With comorb**
Frontal lobe	Left	Precentral gyrus medial segment	0	0	2	3	2	3

Limbic system and associated structures	Left	Posterior cingulate gyrus	0	0	2	2	2	2
Limbic system and associated structures	Left	Parahippocampal gyrus	1	2	1	1	1	1
Limbic system and associated structures	Right	Cingulate White Matter	1	2	0	1	2	2
Limbic system and associated structures	Right	Middle cingulate gyrus	0	0	1	2	1	2
Parietal lobe	Left	Parietal operculum	1	2	2	3	1	1
Parietal lobe	Left	Parietal white matter	0	0	2	2	2	2
Parietal lobe	Left	Supplementary motor cortex	0	0	2	3	1	2

Parietal lobe	Left	Supramarginal gyrus	0	0	2	3	2	2
Parietal lobe	Right	Supplementary motor cortex	0	0	1	2	2	3
occipital lobe	Left	Angular gyrus	0	0	2	2	2	2
occipital lobe	Left	Posterior orbital gyrus	3	3	0	0	1	1
subcortical structures	Left	Putamen	3	3	0	0	1	1
subcortical structures	Right	Ventral diencephalon	0	1	2	3	1	2
temporal lobe	Left	Postcentral gyrus	0	0	0	1	2	3
temporal lobe	Left	Superior temporal gyrus	2	3	0	0	1	1
temporal lobe	Left	Temporal pole	2	2	0	0	1	2

Table A2.14: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism) for Boys aged 15 to 20.

			R42		**D32**		**D70**	
			No comorb	**With comorb**	**No comorb**	**With comorb**	**No comorb**	**With comorb**
Frontal lobe	Left	Precentral gyrus medial segment	0	1	2	2	0	1
Frontal lobe	Left	Opercular part of the inferior frontal gyrus	2	3	0	0	1	2
Frontal lobe	Left	Precentral gyrus	2	2	0	0	1	2
Limbic system and associated structures	Left	Parahippocampal gyrus	0	0	1	1	3	3
Limbic system	Left	Posterior insula	2	2	0	0	2	2

and associated structures								
Limbic system and associated structures	Left	Subcallosal area	2	2	0	0	1	2
Parenchyma	None	3rd Ventricle	0	1	1	1	1	2
Parietal lobe	Left	Parietal operculum	0	1	2	2	2	2
Parietal lobe	Left	Supplementary motor cortex	0	1	1	1	1	2
subcortical structures	Right	Ventral diencephalon	0	0	3	3	1	1
temporal lobe	Left	Planum temporale	1	2	1	1	2	2

temporal lobe	Left	Superior temporal gyrus	2	2	0	0	3	3
temporal lobe	Left	Temporal pole	2	2	0	0	3	3

Table A2.15: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism) for Boys aged 20 to 64.

			R42		**D32**		**D70**	
			No comorb	**With comorb**	**No comorb**	**With comorb**	**No comorb**	**With comorb**
Limbic system and associated structures	Left	Parahippocampal gyrus	1	1	1	1	1	1
Parietal lobe	Left	Parietal white matter	0	0	1	1	2	2
Parietal lobe	Left	Supramarginal gyrus	0	0	1	1	2	2
Parietal lobe	Left	Superior parietal lobule	0	0	1	1	2	2
occipital lobe	Left	Angular gyrus	0	0	1	1	2	2

subcortical structures	Left	Thalamus	1	1	1	1	1	1

Table A2.16: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism) for Girls aged 5 to 10.

			R42		**D32**		**D70**	
			No comorb	**With comorb**	**No comorb**	**With comorb**	**No comorb**	**With comorb**
Frontal lobe	Left	Precentral gyrus medial segment	0	0	1	2	1	2
Frontal lobe	Right	Precentral gyrus medial segment	0	0	1	2	1	2
Limbic system and associated structures	Left	Posterior cingulate gyrus	0	0	1	2	1	2
Parietal lobe	Left	Supramarginal gyrus	0	0	1	2	1	2

Parietal lobe	Right	Supplementary motor cortex	0	0	1	2	2	3

Table A2.17: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism) for Girls aged 10 to 15.

			R42		D32		D70	
			No comorb	With comorb	No comorb	With comorb	No comorb	With comorb
temporal lobe	Right	Postcentral gyrus medial segment	0	0	2	2	2	2

Table A2.18: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism) for Boys aged 15 to 20.

			R42		D32		D70	
			No comorb	With comorb	No comorb	With comorb	No comorb	With comorb
Parietal lobe	Left	Parietal operculum	1	1	1	1	2	2

Table A2.19: Best regions for predicting True Positives (TP, i.e. true diagnosis of Autism) for Girls aged 20 to 64.

8.6.15. A2.10 - Multi-site effect

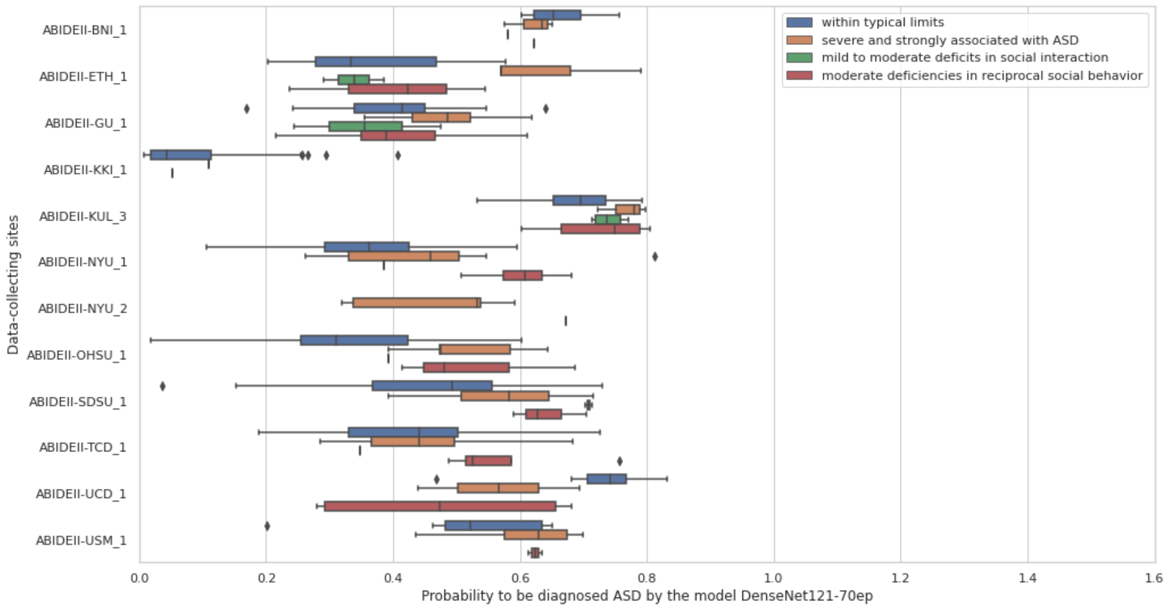


Figure A2.7: Comparison of probabilities of Med3d-ResNet50-42ep and categories obtained from SRS T-scores for different sites

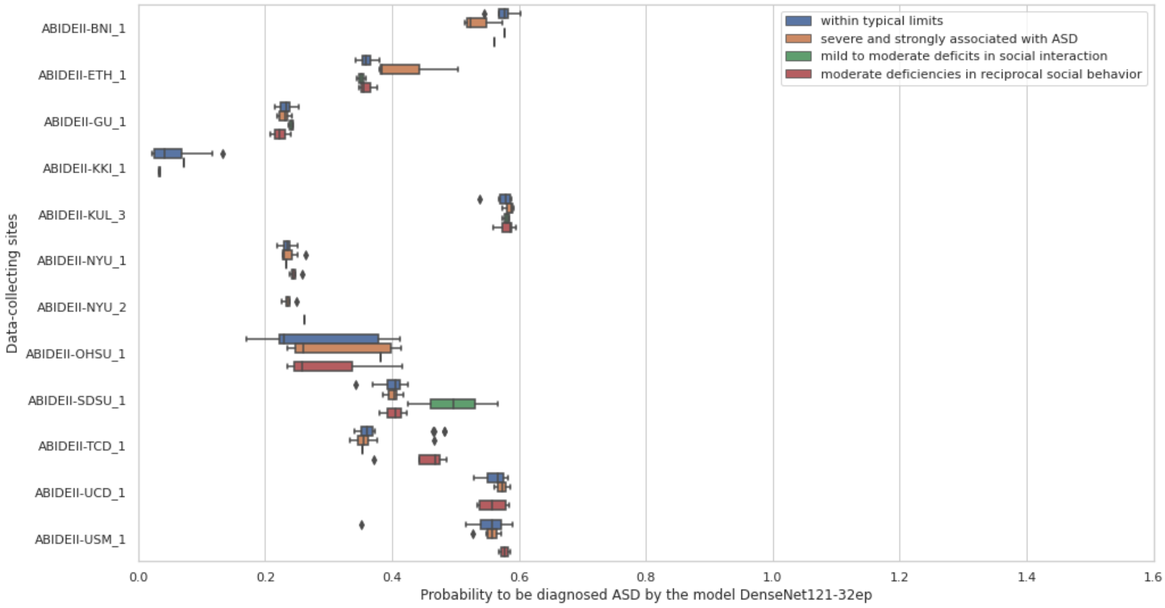


Figure A2.8: Comparison of probabilities of DenseNet121-32ep and categories obtained from SRS T-scores for different sites

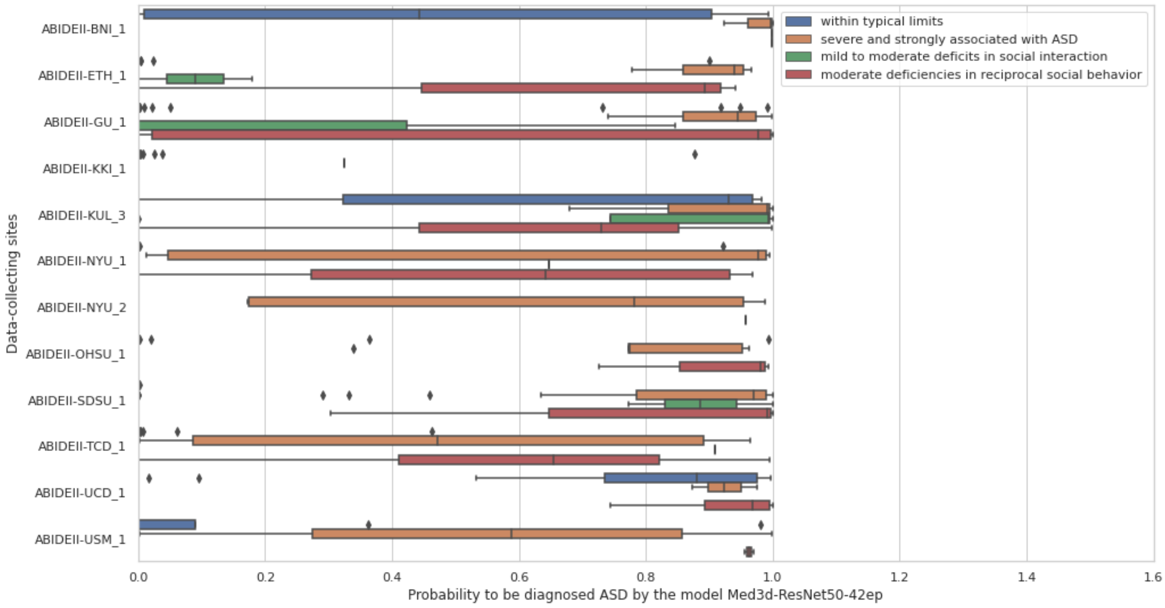


Figure A2.9: Comparison of probabilities of DenseNet121-70ep and categories obtained from SRS T-scores for different sites

In **Figure A2.7**, **Figure A2.8**, and **Figure A2.9**, we observe an inhomogeneous consistency of the distributions of probability scores between the different sites. Results in **Table A2.20**, which displays the accuracy scores for every site in the whole dataset (training+validation+testing sets), confirm the multi-site effect already observed in **Figure A2.7**, **Figure A2.8**, and **Figure A2.9**.

			Med3d-ResNet50-42ep			DenseNet-32ep			DenseNet-70ep		
	site	n	Acc	Acc Autism	Acc TD	Acc	Acc Autism	Acc TD	Acc	Acc Autism	Acc TD
Training or validation set	ABIDEII BNI_1	9	100.0	100.0	100.0	77.8	100.0	0.0	77.8	100.0	0.0
	ABIDEII ETH_1	30	96.7	87.5	100.0	76.7	12.5	100.0	83.3	62.5	90.9
	ABIDEII GU_1	65	98.5	95.2	100.0	67.7	0.0	100.0	72.3	33.3	90.9
	ABIDEII IP_1	36	86.1	63.6	96.0	69.4	0.0	100.0	69.4	54.5	76.0

ABIDEII IU_1	33	93.9	93.8	94.1	54.5	50.0	58.8	57.6	87.5	29.4
ABIDEII KKI_1	123	98.4	0.0	99.2	99.2	0.0	100.0	99.2	0.0	100.0
ABIDEII KUL_3	20	75.0	75.0	NaN	100.0	100.0	NaN	100.0	100.0	NaN
ABIDEII NYU_1	41	85.4	57.1	100.0	65.9	0.0	100.0	75.6	42.9	92.6
ABIDEII NYU_2	6	66.7	66.7	NaN	0.0	0.0	NaN	66.7	66.7	NaN
ABIDEII OHSU_1	55	98.2	90.9	100.0	80.0	0.0	100.0	80.0	45.5	88.6
ABIDEII SDSU_1	55	89.1	80.6	100.0	45.5	3.2	100.0	69.1	80.6	54.2

ABIDEII TCD_1	36	80.6	56.2	100.0	55.6	0.0	100.0	52.8	31.2	70.0
ABIDEII USM_1	23	91.3	83.3	100.0	56.5	100.0	9.1	73.9	91.7	54.5
CALTECH	36	94.4	94.4	94.4	50.0	0.0	100.0	52.8	94.4	11.1
CMU	27	88.9	78.6	100.0	48.1	0.0	100.0	66.7	71.4	61.5
KKI	25	84.0	33.3	100.0	76.0	0.0	100.0	76.0	0.0	100.0
LEUVEN_2	31	83.9	61.5	100.0	41.9	100.0	0.0	67.7	84.6	55.6
MAX_MUN	33	87.9	69.2	100.0	60.6	0.0	100.0	51.5	23.1	70.0
NYU	124	91.9	68.8	100.0	74.2	0.0	100.0	69.4	31.2	82.6
OHSU	22	77.3	63.6	90.9	50.0	0.0	100.0	59.1	90.9	27.3

OLIN	19	63.2	41.7	100.0	63.2	100.0	0.0	63.2	100.0	0.0
PITT	39	84.6	73.7	95.0	51.3	0.0	100.0	51.3	84.2	20.0
SBL	29	75.9	50.0	100.0	58.6	85.7	33.3	48.3	92.9	6.7
SDSU	12	100.0	100.0	100.0	83.3	0.0	100.0	75.0	100.0	70.0
STANFORD	7	57.1	0.0	100.0	57.1	0.0	100.0	57.1	0.0	100.0
TRINITY	42	85.7	70.0	100.0	52.4	0.0	100.0	50.0	40.0	59.1
UCLA_1	52	86.5	81.2	95.0	61.5	100.0	0.0	65.4	87.5	30.0
UCLA_2	15	93.3	100.0	88.9	46.7	100.0	11.1	53.3	100.0	22.2
UM_1	72	90.3	74.1	100.0	62.5	0.0	100.0	66.7	22.2	93.3
UM_2	31	93.5	83.3	100.0	61.3	0.0	100.0	77.4	66.7	84.2
USM	66	87.9	83.7	95.7	63.6	95.3	4.3	62.1	93.0	4.3

Testing set	YALE	50	88.0	79.2	96.2	52.0	0.0	100.0	62.0	87.5	38.5
	LEUVEN_1	27	44.4	21.4	69.2	51.9	100.0	0.0	48.1	64.3	30.8
	ABIDEII EMC_1	18	72.2	50.0	78.6	77.8	0.0	100.0	38.9	100.0	21.4
	ABIDEII UCD_1	20	50.0	100.0	16.7	40.0	100.0	0.0	30.0	62.5	8.3

Table A2.20: Comparing accuracy scores between data collection sites

For Med3d-ResNet50-42ep, the overall accuracy scores are between 44,4% - 100%, with 75% of the data-collecting sites having an accuracy higher than 78,9%, and an overall median accuracy of 87,9%. The sensitivity is between 0% - 100%, with a median of 74%. The specificity is between 16,7% - 100%, with a median of 100%.

For DenseNet121-32ep, the overall accuracy scores are between 0% - 100%, with 75% of the data-collecting sites having an accuracy higher than 51,6%, and an overall median accuracy of 60,6%. The sensitivity is between 0% - 100%, with a median of 0%. The specificity is between 0% - 100%, with a median of 100%.

For DenseNet121-70ep, the overall accuracy scores are between 30% - 100%, with 75% of the data-collecting sites having an accuracy higher than 53,1%, and an overall median accuracy of 66,7%. The sensitivity is between 0% - 100%, with a median of 71,4%. The specificity is between 0% - 100%, with a median of 55,6%.

Appendix 3: Supplemental information on Transformer and multi-tasking to detect ASD using rs-fMRI

8.6.16. A3.1. Shapiro-Wilk tests of normality

diff_Models	**W**	**p_val**	**normal**
diff_M2_M1	0,988	0,500	True
diff_M3_M1	0,987	0,409	True
diff_M3_M2	0,991	0,708	True
diff_M4_M1	0,974	0,047	False
diff_M5_M1	0,970	0,023	False

Table A3.1: Shapiro-Wilk tests of normality of the differences between the models

Appendix 4: Review of the manuscript of (Horien et al., 2022) for the journal Biological Psychiatry

Reviewer 4: Comments to authors:

The review paper by Horien et. al. discusses what can be gained from fMRI-based predictive models of ASD and outlines a number of considerations that are required to make these models clinically and biologically useful. These considerations include issues related to the nature of the diagnosis - ASD - and the nature of the data - functional MRI. The review begins by briefly considering ASD in terms of symptoms, genetics, and neuroimaging findings, and the suggestion that the condition likely comprises a number of subtypes. Next, the authors discuss a number of considerations that might present specific problems for predictive models of ASD using fMRI data, including sample size, data decay, site effects, motion, tolerability of the scanning environment, and sex imbalance. Next, the authors detail three predictive modelling frameworks - case-control classification, dimensional prediction, and subtyping. They discuss each framework in terms of biological insight and clinical utility. They also highlight some important limitations of each framework, including data reliability and the heterogeneity of autism symptomatology for case-control classification, the reliability of behavioural and symptom scores for dimensional prediction, and replicability for subtyping applications. Finally, the review concludes by discussing the potential for "broad" and "deep: data to support the discovery of cross-modal markers and brain trajectories of autism. The ethical considerations described in the Appendix are a welcome part of the paper.

The paper is very well-written and structured. It covers a large number of studies (more than 160), which are very well synthesised. The considerations discussed are very important to consider when building predictive models and systematically outlining these will be beneficial to the field. Nevertheless, it is surprising that there is no mention of the issue of substantial rates of comorbidities amongst Autistic participants, as well as the considerable phenotypic overlap with other psychiatric conditions. Discussing how these challenging factors can be addressed in predictive models would be important (for instance, a link could be made with dimensional prediction and subtyping applications).

The categorisation of predictive frameworks is very clear, and the concepts and frameworks are very well illustrated and accessible thanks to many concrete examples drawn from previous studies. It would also be beneficial to explain the types of machine learning algorithms each of the predictive frameworks are typically associated with. For instance, most case-control classification studies have used classification algorithms, while studies on dimensional models mostly uses regression models, and subtyping applications are often based on clustering or classification algorithms. Making these links would help the reader to organise the different approaches from a machine learning perspective. In addition, it may help readers to better select an algorithm that is most appropriate for their particular study and predictive framework. It might also be helpful to include a section explaining the potential biases associated with the different types of models.

Minor comments:

- It is curious that "machine learning" is not explained or even referred to often in the text. Perhaps the relationship between machine learning and predictive modelling could be clarified

- The section on "Balancing large sample sizes, concerns about data decay, and site-effects" mentions data decay, but what this means is not well explained.

- "When a growing number of investigators analyse the same sample, the false positive rate increases." Why not mention Specificity vs Sensitivity - these are important notions typically considered for predictive models.

- "Nevertheless, implementing GSR is not without controversy; see (40) for a full discussion)": could you describe in one sentence the main points from this discussion?

- "Further, the activity of compensatory circuits may be heightened during the stressful experience of being scanned" - is there any evidence to support this statement?

Appendix 5: Writing steps of the thesis

As a non-native English language speaker, I experienced significant progress in written and oral communication in this language throughout the course of the PhD.

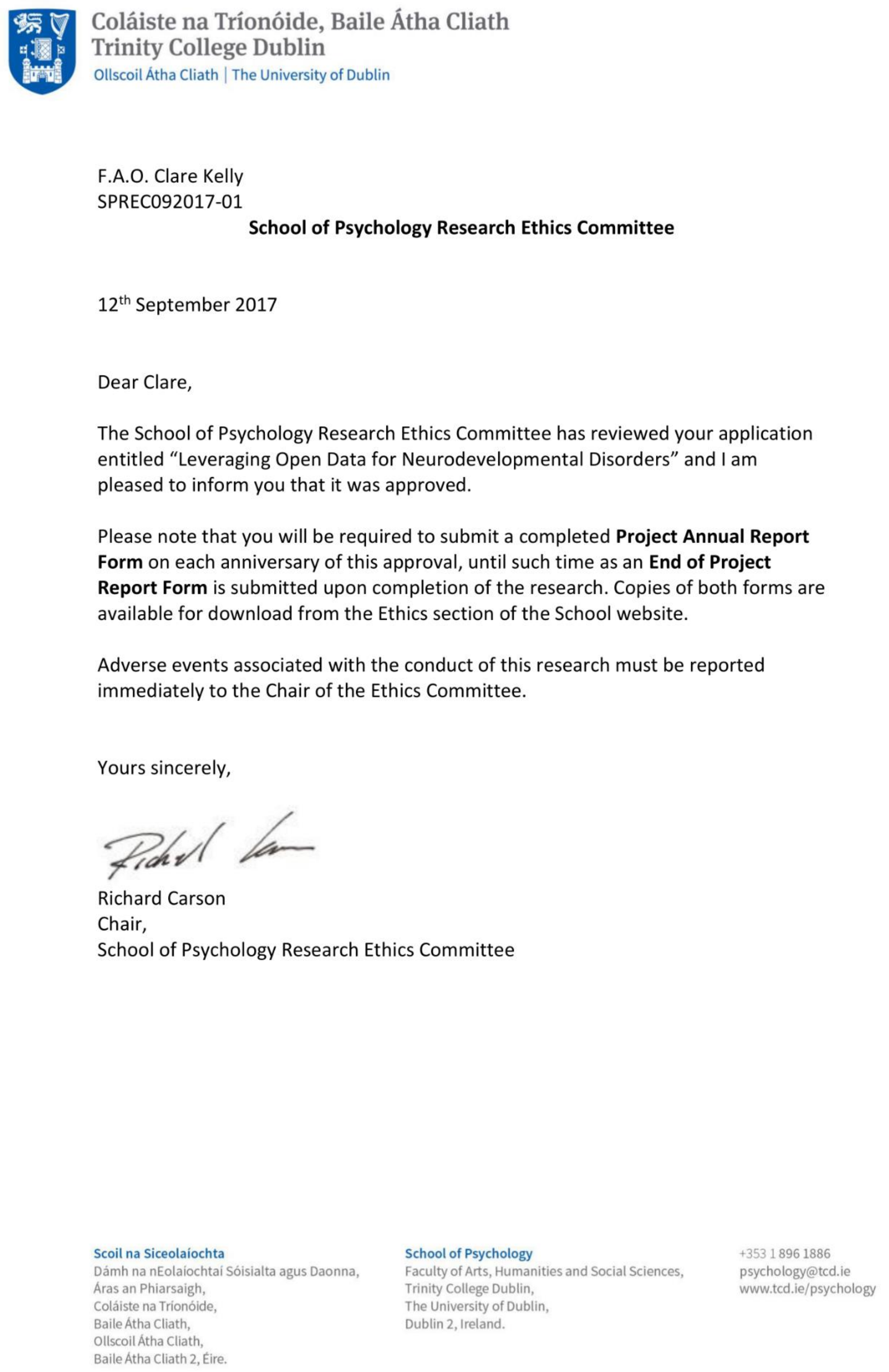
Exchanging with my PhD supervisor, my colleagues and living in Ireland had a particular positive influence on this progression.

In order to increase the quality of my written communication in several parts of this manuscript, I used an AI tool (a prompt interface with a pretrained model of Llama 2 (https://huggingface.co/blog/llama2) that I fine-tuned on a corpus of open scientific articles). Knowing that the use of AI to assist researchers in writing is currently under debate (Berdejo-Espinola & Amano, 2023), being transparent about the writing steps of this thesis appeared essential to make clear how, when and to what extent my written work was changed by an AI tool.

I created an open GitHub repository (https://github.com/garciaml/PhD_thesis) where the old versions of the Chapter 1 (Introduction), Chapter 5 (Third empirical study) and Chapter 8 (Discussion) can be found, as well as the corresponding AI-modified versions (that are not the final chapter versions because these were polished again after).

AI was a helpful tool to polish my communication skills without removing the core work of the thesis.

Appendix 6: Approval from the School of Psychology Research Ethics Committee



Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

F.A.O. Clare Kelly
SPREC092017-01

School of Psychology Research Ethics Committee

12th September 2017

Dear Clare,

The School of Psychology Research Ethics Committee has reviewed your application entitled "Leveraging Open Data for Neurodevelopmental Disorders" and I am pleased to inform you that it was approved.

Please note that you will be required to submit a completed **Project Annual Report Form** on each anniversary of this approval, until such time as an **End of Project Report Form** is submitted upon completion of the research. Copies of both forms are available for download from the Ethics section of the School website.

Adverse events associated with the conduct of this research must be reported immediately to the Chair of the Ethics Committee.

Yours sincerely,

Richard Carson
Chair,
School of Psychology Research Ethics Committee

Scoil na Siceolaíochta
Dámh na nEolaíochtaí Sóisialta agus Daonna,
Áras an Phiarsaigh,
Coláiste na Tríonóide,
Baile Átha Cliath,
Ollscoil Átha Cliath,
Baile Átha Cliath 2, Éire.

School of Psychology
Faculty of Arts, Humanities and Social Sciences,
Trinity College Dublin,
The University of Dublin,
Dublin 2, Ireland.

+353 1 896 1886
psychology@tcd.ie
www.tcd.ie/psychology

Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

SCHOOL OF PSYCHOLOGY
Arás an Phiarsaigh
Trinity College
Dublin 2

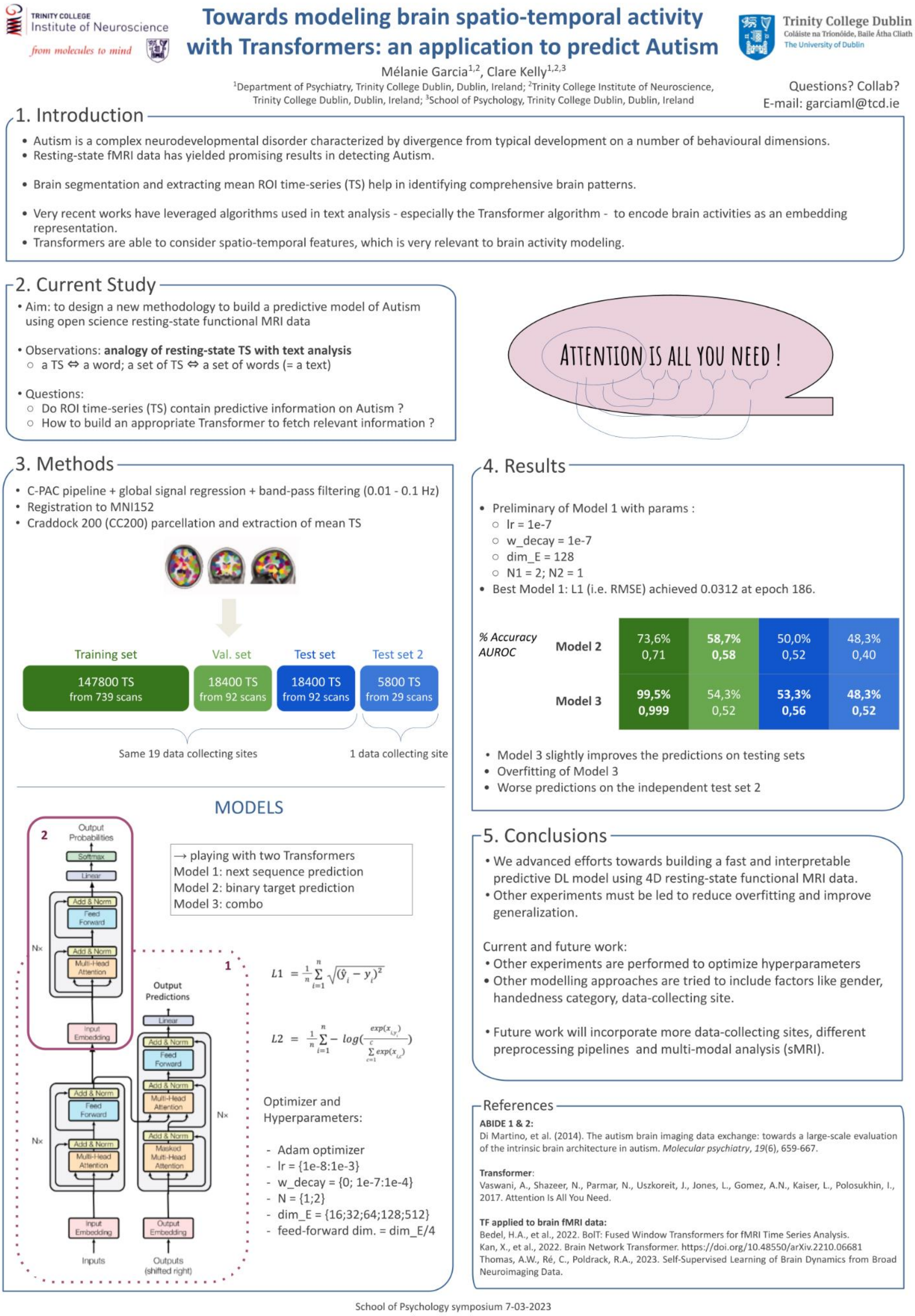
Scoil na Siceolaíochta
Dámh na nEolaíochtaí Sóisialta agus Daonna,
Áras an Phiarsaigh,
Coláiste na Tríonóide,
Baile Átha Cliath,
Ollscoil Átha Cliath,
Baile Átha Cliath 2, Éire.

School of Psychology
Faculty of Arts, Humanities and Social Sciences,
Trinity College Dublin,
The University of Dublin,
Dublin 2, Ireland.

+353 1 896 1886
psychology@tcd.ie
www.tcd.ie/psychology

2

Appendix 7: Poster presented at the PsySchool Symposium at Trinity College Dublin, 2023.



TRINITY COLLEGE Institute of Neuroscience *from molecules to mind*

Towards modeling brain spatio-temporal activity with Transformers: an application to predict Autism

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Mélanie Garcia[1,2], Clare Kelly[1,2,3]

[1]Department of Psychiatry, Trinity College Dublin, Dublin, Ireland; [2]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland; [3]School of Psychology, Trinity College Dublin, Dublin, Ireland

Questions? Collab?
E-mail: garciaml@tcd.ie

1. Introduction

- Autism is a complex neurodevelopmental disorder characterized by divergence from typical development on a number of behavioural dimensions.
- Resting-state fMRI data has yielded promising results in detecting Autism.
- Brain segmentation and extracting mean ROI time-series (TS) help in identifying comprehensive brain patterns.
- Very recent works have leveraged algorithms used in text analysis - especially the Transformer algorithm - to encode brain activities as an embedding representation.
- Transformers are able to consider spatio-temporal features, which is very relevant to brain activity modeling.

2. Current Study

- Aim: to design a new methodology to build a predictive model of Autism using open science resting-state functional MRI data
- Observations: **analogy of resting-state TS with text analysis**
  - a TS ⇔ a word; a set of TS ⇔ a set of words (= a text)
- Questions:
  - Do ROI time-series (TS) contain predictive information on Autism ?
  - How to build an appropriate Transformer to fetch relevant information ?

3. Methods

- C-PAC pipeline + global signal regression + band-pass filtering (0.01 - 0.1 Hz)
- Registration to MNI152
- Craddock 200 (CC200) parcellation and extraction of mean TS

Training set	Val. set	Test set	Test set 2
147800 TS from 739 scans	18400 TS from 92 scans	18400 TS from 92 scans	5800 TS from 29 scans

Same 19 data collecting sites — 1 data collecting site

MODELS

→ playing with two Transformers
Model 1: next sequence prediction
Model 2: binary target prediction
Model 3: combo

$$L1 = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(\hat{y}_i - y_i)^2}$$

$$L2 = \frac{1}{n}\sum_{i=1}^{n} - log(\frac{exp(x_{i,y_i})}{\sum_{c=1}^{c} exp(x_{i,c})})$$

Optimizer and Hyperparameters:

- Adam optimizer
- lr = {1e-8:1e-3}
- w_decay = {0; 1e-7:1e-4}
- N = {1;2}
- dim_E = {16;32;64;128;512}
- feed-forward dim. = dim_E/4

4. Results

- Preliminary of Model 1 with params :
  - lr = 1e-7
  - w_decay = 1e-7
  - dim_E = 128
  - N1 = 2; N2 = 1
- Best Model 1: L1 (i.e. RMSE) achieved 0.0312 at epoch 186.

% Accuracy *AUROC*	**Model 2**	73,6% 0,71	**58,7%** **0,58**	50,0% 0,52	48,3% 0,40
	Model 3	**99,5%** **0,999**	54,3% 0,52	**53,3%** **0,56**	**48,3%** **0,52**

- Model 3 slightly improves the predictions on testing sets
- Overfitting of Model 3
- Worse predictions on the independent test set 2

5. Conclusions

- We advanced efforts towards building a fast and interpretable predictive DL model using 4D resting-state functional MRI data.
- Other experiments must be led to reduce overfitting and improve generalization.

Current and future work:

- Other experiments are performed to optimize hyperparameters
- Other modelling approaches are tried to include factors like gender, handedness category, data-collecting site.
- Future work will incorporate more data-collecting sites, different preprocessing pipelines and multi-modal analysis (sMRI).

References

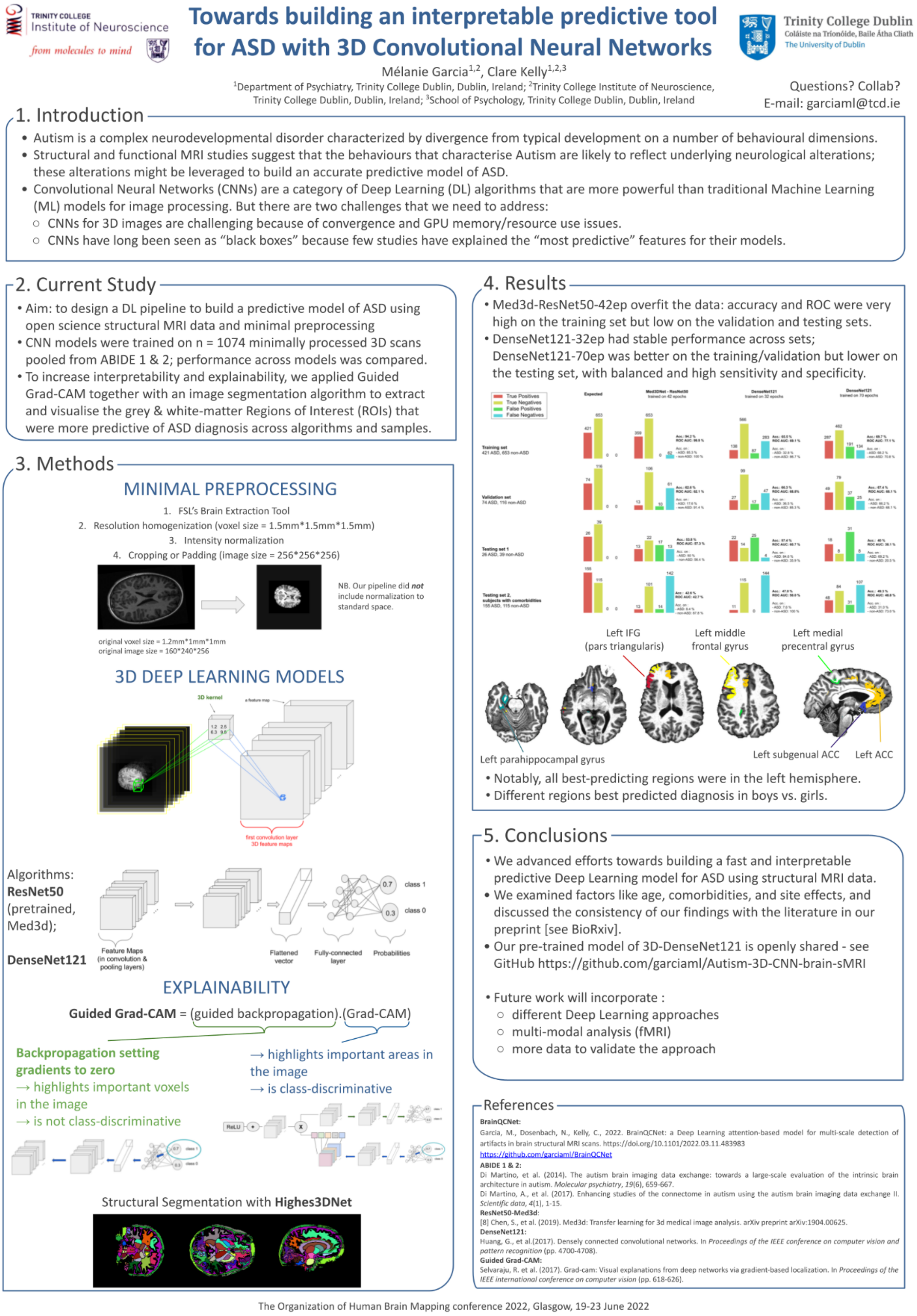
ABIDE 1 & 2:
Di Martino, et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, *19*(6), 659-667.

Transformer:
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need.

TF applied to brain fMRI data:
Bedel, H.A., et al., 2022. BolT: Fused Window Transformers for fMRI Time Series Analysis.
Kan, X., et al., 2022. Brain Network Transformer. https://doi.org/10.48550/arXiv.2210.06681
Thomas, A.W., Ré, C., Poldrack, R.A., 2023. Self-Supervised Learning of Brain Dynamics from Broad Neuroimaging Data.

School of Psychology symposium 7-03-2023

Appendix 8: Poster presented at the conference OHBM 2022.



TRINITY COLLEGE
Institute of Neuroscience
from molecules to mind
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Towards building an interpretable predictive tool for ASD with 3D Convolutional Neural Networks

Mélanie Garcia[1,2], Clare Kelly[1,2,3]

[1]Department of Psychiatry, Trinity College Dublin, Dublin, Ireland; [2]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland; [3]School of Psychology, Trinity College Dublin, Dublin, Ireland

Questions? Collab?
E-mail: garciaml@tcd.ie

1. Introduction

- Autism is a complex neurodevelopmental disorder characterized by divergence from typical development on a number of behavioural dimensions.
- Structural and functional MRI studies suggest that the behaviours that characterise Autism are likely to reflect underlying neurological alterations; these alterations might be leveraged to build an accurate predictive model of ASD.
- Convolutional Neural Networks (CNNs) are a category of Deep Learning (DL) algorithms that are more powerful than traditional Machine Learning (ML) models for image processing. But there are two challenges that we need to address:
  - ○ CNNs for 3D images are challenging because of convergence and GPU memory/resource use issues.
  - ○ CNNs have long been seen as "black boxes" because few studies have explained the "most predictive" features for their models.

2. Current Study

- Aim: to design a DL pipeline to build a predictive model of ASD using open science structural MRI data and minimal preprocessing
- CNN models were trained on n = 1074 minimally processed 3D scans pooled from ABIDE 1 & 2; performance across models was compared.
- To increase interpretability and explainability, we applied Guided Grad-CAM together with an image segmentation algorithm to extract and visualise the grey & white-matter Regions of Interest (ROIs) that were more predictive of ASD diagnosis across algorithms and samples.

3. Methods

MINIMAL PREPROCESSING

1. FSL's Brain Extraction Tool
2. Resolution homogenization (voxel size = 1.5mm*1.5mm*1.5mm)
3. Intensity normalization
4. Cropping or Padding (image size = 256*256*256)

NB. Our pipeline did *not* include normalization to standard space.

original voxel size = 1.2mm*1mm*1mm
original image size = 160*240*256

3D DEEP LEARNING MODELS

Algorithms:
ResNet50 (pretrained, Med3d);

DenseNet121

EXPLAINABILITY

Guided Grad-CAM = $\underbrace{\text{(guided backpropagation)}}.\underbrace{\text{(Grad-CAM)}}$

Backpropagation setting gradients to zero
→ highlights important voxels in the image
→ is not class-discriminative

→ highlights important areas in the image
→ is class-discriminative

Structural Segmentation with **Highes3DNet**

4. Results

- Med3d-ResNet50-42ep overfit the data: accuracy and ROC were very high on the training set but low on the validation and testing sets.
- DenseNet121-32ep had stable performance across sets; DenseNet121-70ep was better on the training/validation but lower on the testing set, with balanced and high sensitivity and specificity.

- Notably, all best-predicting regions were in the left hemisphere.
- Different regions best predicted diagnosis in boys vs. girls.

5. Conclusions

- We advanced efforts towards building a fast and interpretable predictive Deep Learning model for ASD using structural MRI data.
- We examined factors like age, comorbidities, and site effects, and discussed the consistency of our findings with the literature in our preprint [see BioRxiv].
- Our pre-trained model of 3D-DenseNet121 is openly shared - see GitHub https://github.com/garciaml/Autism-3D-CNN-brain-sMRI

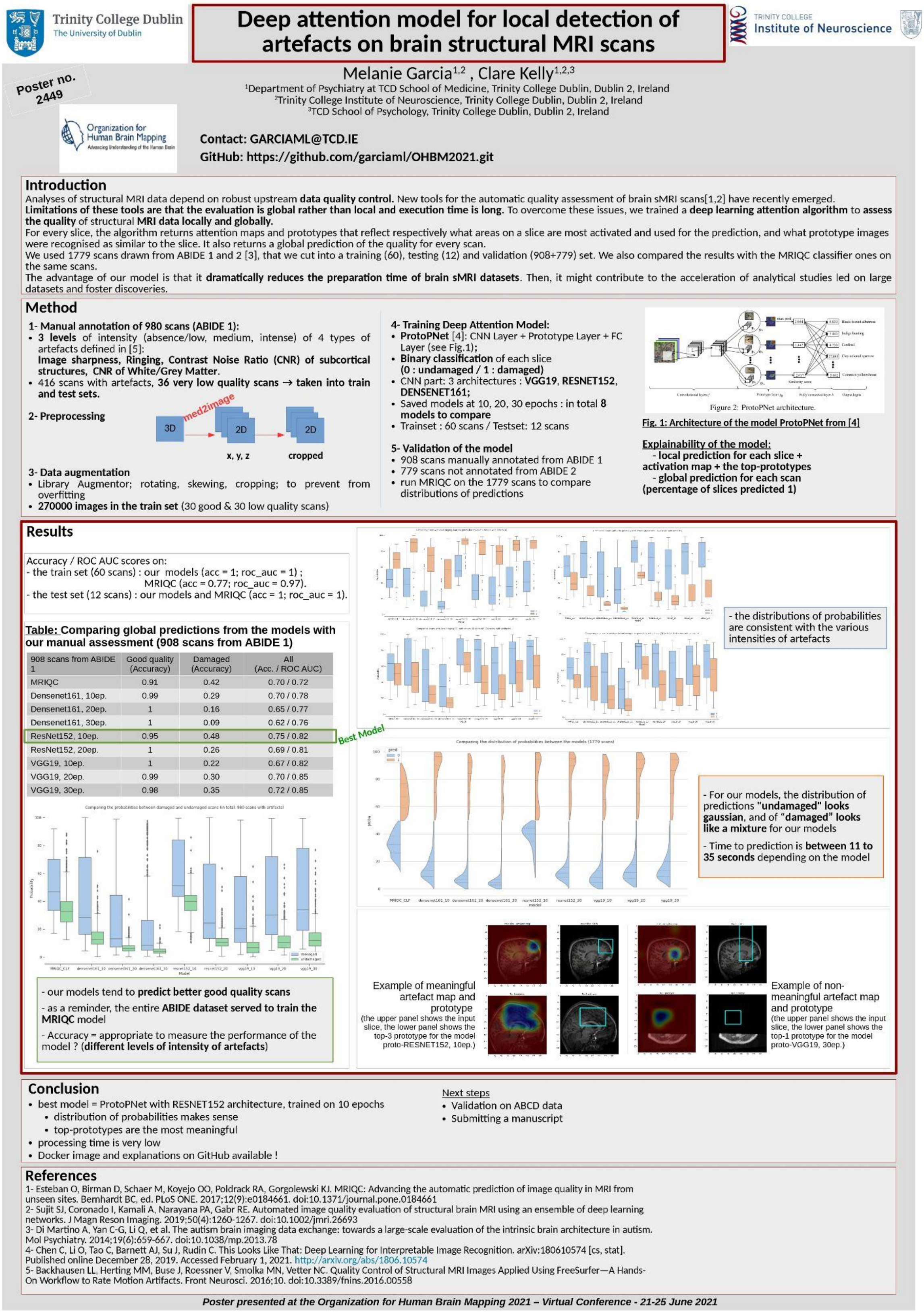
- Future work will incorporate :
  - ○ different Deep Learning approaches
  - ○ multi-modal analysis (fMRI)
  - ○ more data to validate the approach

References

BrainQCNet:
Garcia, M., Dosenbach, N., Kelly, C., 2022. BrainQCNet: a Deep Learning attention-based model for multi-scale detection of artifacts in brain structural MRI scans. https://doi.org/10.1101/2022.03.11.483983
https://github.com/garciaml/BrainQCNet
ABIDE 1 & 2:
Di Martino, et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, *19*(6), 659-667.
Di Martino, A., et al. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific data*, *4*(1), 1-15.
ResNet50-Med3d:
[8] Chen, S., et al. (2019). Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625.
DenseNet121:
Huang, G., et al.(2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).
Guided Grad-CAM:
Selvaraju, R. et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

The Organization of Human Brain Mapping conference 2022, Glasgow, 19-23 June 2022

Appendix 9: Poster presented at the conference OHBM 2021.



Deep attention model for local detection of artefacts on brain structural MRI scans

Melanie Garcia[1,2] , Clare Kelly[1,2,3]

[1]Department of Psychiatry at TCD School of Medicine, Trinity College Dublin, Dublin 2, Ireland
[2]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland
[3]TCD School of Psychology, Trinity College Dublin, Dublin 2, Ireland

Contact: GARCIAML@TCD.IE
GitHub: https://github.com/garciaml/OHBM2021.git

Poster no. 2449

Introduction

Analyses of structural MRI data depend on robust upstream **data quality control.** New tools for the automatic quality assessment of brain sMRI scans[1,2] have recently emerged.
Limitations of these tools are that the evaluation is global rather than local and execution time is long. To overcome these issues, we trained a **deep learning attention algorithm** to **assess the quality** of structural **MRI data locally and globally.**
For every slice, the algorithm returns attention maps and prototypes that reflect respectively what areas on a slice are most activated and used for the prediction, and what prototype images were recognised as similar to the slice. It also returns a global prediction of the quality for every scan.
We used 1779 scans drawn from ABIDE 1 and 2 [3], that we cut into a training (60), testing (12) and validation (908+779) set. We also compared the results with the MRIQC classifier ones on the same scans.
The advantage of our model is that it **dramatically reduces the preparation time of brain sMRI datasets**. Then, it might contribute to the acceleration of analytical studies led on large datasets and foster discoveries.

Method

1- Manual annotation of 980 scans (ABIDE 1):
- **3 levels** of intensity (absence/low, medium, intense) of 4 types of artefacts defined in [5]:
  Image sharpness, Ringing, Contrast Noise Ratio (CNR) of subcortical structures, CNR of White/Grey Matter.
- 416 scans with artefacts, **36 very low quality scans → taken into train and test sets.**

2- Preprocessing

3- Data augmentation
- Library Augmentor; rotating, skewing, cropping; to prevent from overfitting
- **270000 images in the train set** (30 good & 30 low quality scans)

4- Training Deep Attention Model:
- **ProtoPNet** [4]: CNN Layer + Prototype Layer + FC Layer (see Fig.1);
- **Binary classification** of each slice **(0 : undamaged / 1 : damaged)**
- CNN part: 3 architectures : **VGG19**, **RESNET152**, **DENSENET161;**
- Saved models at 10, 20, 30 epochs : in total **8 models to compare**
- Trainset : 60 scans / Testset: 12 scans

5- Validation of the model
- 908 scans manually annotated from ABIDE 1
- 779 scans not annotated from ABIDE 2
- run MRIQC on the 1779 scans to compare distributions of predictions

Fig. 1: Architecture of the model ProtoPNet from [4]

Explainability of the model:
- **local prediction for each slice + activation map + the top-prototypes**
- **global prediction for each scan (percentage of slices predicted 1)**

Results

Accuracy / ROC AUC scores on:
- the train set (60 scans) : our models (acc = 1; roc_auc = 1) ;
  MRIQC (acc = 0.77; roc_auc = 0.97).
- the test set (12 scans) : our models and MRIQC (acc = 1; roc_auc = 1).

Table: Comparing global predictions from the models with our manual assessment (908 scans from ABIDE 1)

908 scans from ABIDE 1	Good quality (Accuracy)	Damaged (Accuracy)	All (Acc. / ROC AUC)
MRIQC	0.91	0.42	0.70 / 0.72
Densenet161, 10ep.	0.99	0.29	0.70 / 0.78
Densenet161, 20ep.	1	0.16	0.65 / 0.77
Densenet161, 30ep.	1	0.09	0.62 / 0.76
ResNet152, 10ep.	0.95	0.48	0.75 / 0.82
ResNet152, 20ep.	1	0.26	0.69 / 0.81
VGG19, 10ep.	1	0.22	0.67 / 0.82
VGG19, 20ep.	0.99	0.30	0.70 / 0.85
VGG19, 30ep.	0.98	0.35	0.72 / 0.85

- our models tend to **predict better good quality scans**
- as a reminder, the entire **ABIDE dataset served to train the MRIQC** model
- Accuracy = appropriate to measure the performance of the model ? (**different levels of intensity of artefacts**)

- the distributions of probabilities are consistent with the various intensities of artefacts

- For our models, the distribution of predictions **"undamaged" looks gaussian**, and of **"damaged" looks like a mixture** for our models
- Time to prediction is **between 11 to 35 seconds** depending on the model

Example of meaningful artefact map and prototype
(the upper panel shows the input slice, the lower panel shows the top-3 prototype for the model proto-RESNET152, 10ep.)

Example of non-meaningful artefact map and prototype
(the upper panel shows the input slice, the lower panel shows the top-1 prototype for the model proto-VGG19, 30ep.)

Conclusion

- best model = ProtoPNet with RESNET152 architecture, trained on 10 epochs
  - distribution of probabilities makes sense
  - top-prototypes are the most meaningful
- processing time is very low
- Docker image and explanations on GitHub available !

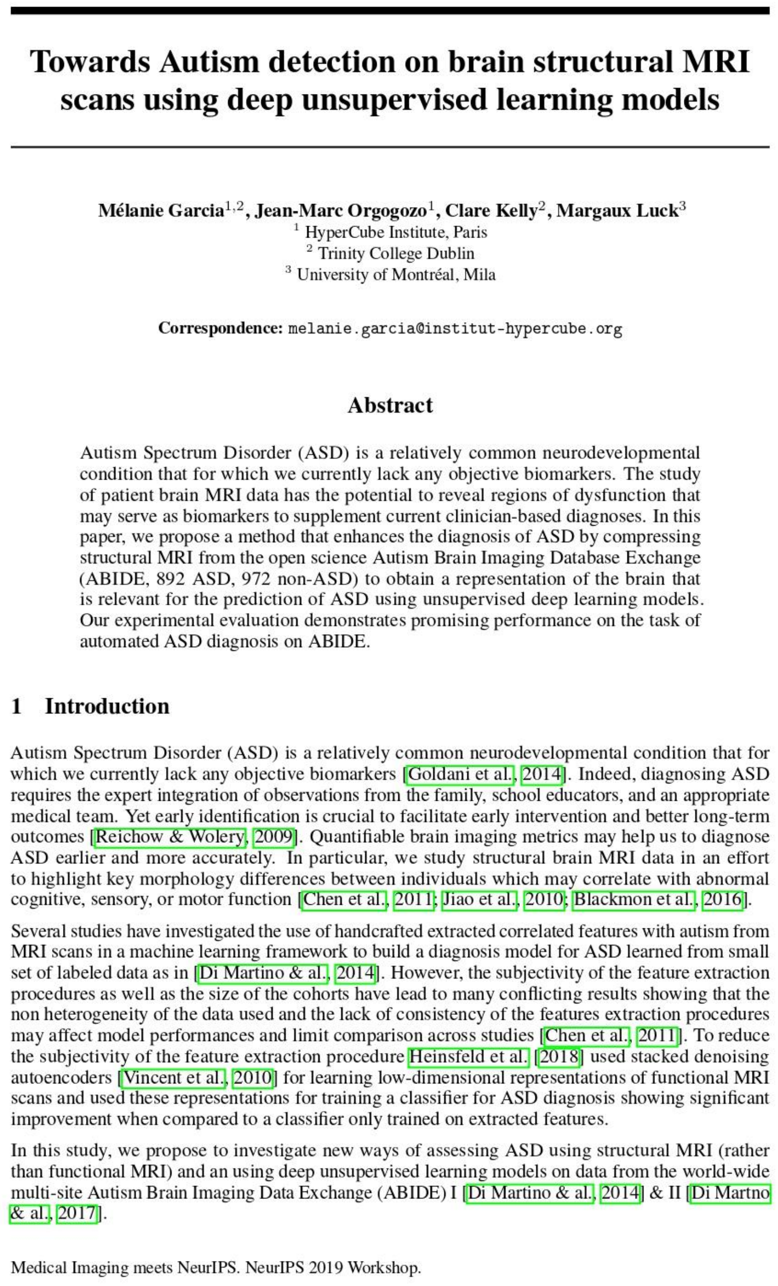
Next steps
- Validation on ABCD data
- Submitting a manuscript

References

1- Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, Gorgolewski KJ. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. Bernhardt BC, ed. PLoS ONE. 2017;12(9):e0184661. doi:10.1371/journal.pone.0184661
2- Sujit SJ, Coronado I, Kamali A, Narayana PA, Gabr RE. Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. J Magn Reson Imaging. 2019;50(4):1260-1267. doi:10.1002/jmri.26693
3- Di Martino A, Yan C-G, Li Q, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry. 2014;19(6):659-667. doi:10.1038/mp.2013.78
4- Chen C, Li O, Tao C, Barnett AJ, Su J, Rudin C. This Looks Like That: Deep Learning for Interpretable Image Recognition. arXiv:180610574 [cs, stat]. Published online December 28, 2019. Accessed February 1, 2021. http://arxiv.org/abs/1806.10574
5- Backhausen LL, Herting MM, Buse J, Roessner V, Smolka MN, Vetter NC. Quality Control of Structural MRI Images Applied Using FreeSurfer—A Hands-On Workflow to Rate Motion Artifacts. Front Neurosci. 2016;10. doi:10.3389/fnins.2016.00558

Poster presented at the Organization for Human Brain Mapping 2021 – Virtual Conference - 21-25 June 2021

Appendix 10: Abstract and Poster presented at the workshop Medical Imaging meets NeurIPS 2019.



Towards Autism detection on brain structural MRI scans using deep unsupervised learning models

Mélanie Garcia[1,2], **Jean-Marc Orgogozo**[1], **Clare Kelly**[2], **Margaux Luck**[3]
[1] HyperCube Institute, Paris
[2] Trinity College Dublin
[3] University of Montréal, Mila

Correspondence: `melanie.garcia@institut-hypercube.org`

Abstract

Autism Spectrum Disorder (ASD) is a relatively common neurodevelopmental condition that for which we currently lack any objective biomarkers. The study of patient brain MRI data has the potential to reveal regions of dysfunction that may serve as biomarkers to supplement current clinician-based diagnoses. In this paper, we propose a method that enhances the diagnosis of ASD by compressing structural MRI from the open science Autism Brain Imaging Database Exchange (ABIDE, 892 ASD, 972 non-ASD) to obtain a representation of the brain that is relevant for the prediction of ASD using unsupervised deep learning models. Our experimental evaluation demonstrates promising performance on the task of automated ASD diagnosis on ABIDE.

1 Introduction

Autism Spectrum Disorder (ASD) is a relatively common neurodevelopmental condition that for which we currently lack any objective biomarkers [Goldani et al., 2014]. Indeed, diagnosing ASD requires the expert integration of observations from the family, school educators, and an appropriate medical team. Yet early identification is crucial to facilitate early intervention and better long-term outcomes [Reichow & Wolery, 2009]. Quantifiable brain imaging metrics may help us to diagnose ASD earlier and more accurately. In particular, we study structural brain MRI data in an effort to highlight key morphology differences between individuals which may correlate with abnormal cognitive, sensory, or motor function [Chen et al., 2011; Jiao et al., 2010; Blackmon et al., 2016].

Several studies have investigated the use of handcrafted extracted correlated features with autism from MRI scans in a machine learning framework to build a diagnosis model for ASD learned from small set of labeled data as in [Di Martino & al., 2014]. However, the subjectivity of the feature extraction procedures as well as the size of the cohorts have lead to many conflicting results showing that the non heterogeneity of the data used and the lack of consistency of the features extraction procedures may affect model performances and limit comparison across studies [Chen et al., 2011]. To reduce the subjectivity of the feature extraction procedure Heinsfeld et al. [2018] used stacked denoising autoencoders [Vincent et al., 2010] for learning low-dimensional representations of functional MRI scans and used these representations for training a classifier for ASD diagnosis showing significant improvement when compared to a classifier only trained on extracted features.

In this study, we propose to investigate new ways of assessing ASD using structural MRI (rather than functional MRI) and an using deep unsupervised learning models on data from the world-wide multi-site Autism Brain Imaging Data Exchange (ABIDE) I [Di Martino & al., 2014] & II [Di Martno & al., 2017].

Medical Imaging meets NeurIPS. NeurIPS 2019 Workshop.

To our knowledge this is the first attempt to use such deep learning models on structural MRI scans on ABIDE. Our experimental evaluation demonstrates promising performance on the task of automated ASD diagnosis.

2 Experiments/Methods

2.1 Data

The data used in this study are structural MRIs and phenotypic data from the Autism Brain Data Exchange I [Di Martino & al., 2014] and II [Di Martno & al., 2017]. This initiative has aggregated functional and structural brain imaging data collected from laboratories around the world to investigate the neural basis of Autism. We used the Configurable Pipeline for the Analysis of Connectomes (C-PAC) [Craddock et al., 2013] as a tool for the preprocessing step to make MRI comparable between each other: we transformed structural MRIs with deobliquing, reorienting, skull-stripping, intensity normalization and registration with the MNI 152 template [Fonov et al., 2011, 2009].

We used 1607 (86% of the dataset) MRI scans to train our unsupervised deep learning models and 275 (14%) for the holdout test set. In both sets there were the same proportion of data from ABIDE I and ABIDE II, and the same proportion of autism cases. The training and test sets were contributed by different data collection sites. Images were grid sampled with a window size of (32, 32, 32), each volume sample corresponded to a location on the brain and was input of our models. There were 36 sampled locations per image.

2.2 Method and architecture

Our model is designed such that learning is split into two parts: a feature extractor and a classifier. The feature extractor is a Variational Autoencoder (VAE) [Kingma & Welling, 2013; Rezende et al., 2014] or with Adversarial Learned Inference (ALI) [Dumoulin et al., 2016]. We used a common architecture for the two networks, inspired from the one used in the maxout network study [Goodfellow et al.], which was also tested in the ALI study [Dumoulin et al., 2016]. Specifically, the discriminator consists of convolution and max pooling blocks, followed by maxout layers.

The classifier is the interpretation with respect to ASD prediction. For each brain location, we trained different classifiers on latent data with a binary target autistic / non-autistic and compared them: elascticnet penalized logistic regression with a L1 ratio of 0.5 or a one-layer perceptron with 10 neurons. This resulted in a probability and the prediction of a class for each location which was used to train a second classifier. The second classifiers trained on the first predictions were a simple majority rule classifier, an elasticnet regression with a L1 ratio of 0.5, or a one-layer perceptron with 5 neurons.

3 Results

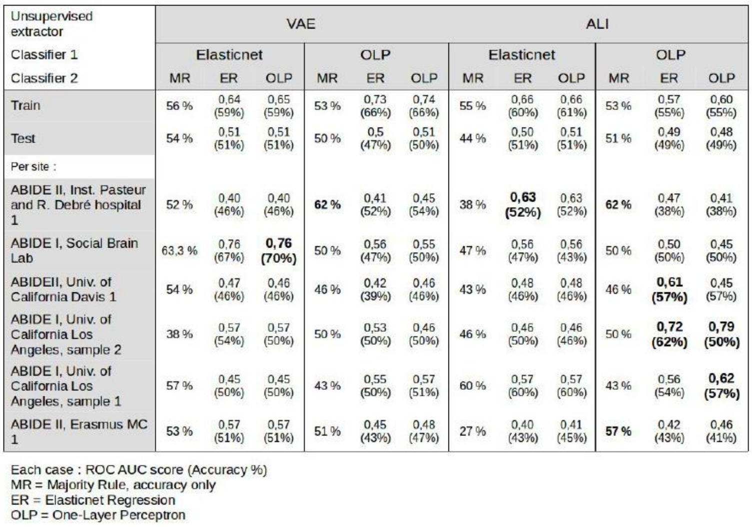
We trained ALI and VAE on five epochs on the train set, with a validation step on 10% of the set every 10 iterations. According to the training loss values, the two algorithms reached an asymptote. We used the ROC AUC score and accuracy as metrics to evaluate our models.

The first classifier on each cerebral location enabled us to find locations responsible for the prediction of ASD. We noticed that the highlighted brain regions vary according to the pipeline of feature extraction used, which includes the unsupervised deep learning algorithm and the first linear or non-linear classifier. With VAE as MRI compressor, we noticed that the locations involved in ASD predictions corresponded to the frontal lobe known to be linked with planning and attention, and to the occipital lobe known to be linked with vision recognition. With ALI as MRI compressor, the occipital lobe was again a salient region for predicting ASD, but also the parietal lobe known to be linked with touch. To obtain more accurate results, it would probably be necessary to review the grid sampling step and use multi-scale algorithms.

For the second classifier, particularly for the logistic regression and for the one-layer perceptron, we reduced the training set to the 10 brain sampled locations inferring the highest standard-deviated predictions, according to each combination of extractor and classifier. The threshold for the majority rule was more than half of the locations predicting ASD. The logistic regression was trained with

4-folds cross-validation to optimize the strength of regularization, and a final refit was made on the whole train set to get the coefficients. We inferred predictions on the whole test set and detailed the scores on subsets corresponding to each data provider in the test set.

The resulting scores of performance are shown in figure 1 and we can observe significant variations between pipelines and the providing site concerned. The best accuracy and ROC AUC scores for every site provide a state-of-the-art for detection of Autism with with brain structural MRI on multi-site study using ABIDE [Di Martino & al., 2014; Di Martno & al., 2017]. Our results showed that data collected at different centers yielded widely varying results. Further analysis is needed to understand and control this variance.



Unsupervised extractor	VAE						ALI					
Classifier 1	Elasticnet			OLP			Elasticnet			OLP		
Classifier 2	MR	ER	OLP	MR	ER	OLP	MR	ER	OLP	MR	ER	OLP
Train	56 %	0,64 (59%)	0,65 (59%)	53 %	0,73 (66%)	0,74 (66%)	55 %	0,66 (60%)	0,66 (61%)	53 %	0,57 (55%)	0,60 (55%)
Test	54 %	0,51 (51%)	0,51 (51%)	50 %	0,5 (47%)	0,51 (50%)	44 %	0,50 (51%)	0,51 (51%)	51 %	0,49 (49%)	0,48 (49%)
Per site :												
ABIDE II, Inst. Pasteur and R. Debré hospital 1	52 %	0,40 (46%)	0,40 (46%)	**62 %**	0,41 (52%)	0,45 (54%)	38 %	**0,63 (52%)**	0,63 (52%)	**62 %**	0,47 (38%)	0,41 (38%)
ABIDE I, Social Brain Lab	63,3 %	0,76 (67%)	**0,76 (70%)**	50 %	0,56 (47%)	0,55 (50%)	47 %	0,56 (47%)	0,56 (43%)	50 %	0,50 (50%)	0,45 (50%)
ABIDEII, Univ. of California Davis 1	54 %	0,47 (46%)	0,46 (46%)	46 %	0,42 (39%)	0,46 (46%)	43 %	0,48 (46%)	0,48 (46%)	46 %	**0,61 (57%)**	0,45 (57%)
ABIDE I, Univ. of California Los Angeles, sample 2	38 %	0,57 (54%)	0,57 (50%)	50 %	0,53 (50%)	0,46 (50%)	46 %	0,46 (50%)	0,46 (46%)	50 %	**0,72 (62%)**	**0,79 (50%)**
ABIDE I, Univ. of California Los Angeles, sample 1	57 %	0,45 (50%)	0,45 (50%)	43 %	0,55 (50%)	0,57 (51%)	60 %	0,57 (60%)	0,57 (60%)	43 %	0,56 (54%)	**0,62 (57%)**
ABIDE II, Erasmus MC 1	53 %	0,57 (51%)	0,57 (51%)	51 %	0,45 (43%)	0,48 (47%)	27 %	0,40 (43%)	0,41 (45%)	**57 %**	0,42 (43%)	0,46 (41%)

Each case : ROC AUC score (Accuracy %)
MR = Majority Rule, accuracy only
ER = Elasticnet Regression
OLP = One-Layer Perceptron

Figure 1: Resulting scores from the different pipelines of brain MRI scan compression and binary classification on the presence of ASD; ROC AUC score (accuracy score%) are given for the whole train and test set, and detailed for each site in the test set.

4 Limitations/Conclusion

This work is a proof of concept that deep learning could be useful for ASD diagnosis, and could raise new paths of research on finding Autism neurological markers. This pipeline on brain structural MRI to detect Autism give promising results, providing weak locations of the part of the brain involved. Nevertheless, these findings are very preliminary and should be compared with more classic models as baselines. The approach could also be improved by taking into account the data collection sites and other modalities like functional MRI during training.

5 Acknowledgement

We would like to thank Martin Weiss our colleague at Mila for reviewing and editing this manuscript.

References

Blackmon, K., Ben-Avi, E., Wang, X., Pardoe, H. R., Di Martino, A., Halgren, E., Devinsky, O., Thesen, T., and Kuzniecky, R. Periventricular white matter abnormalities and restricted repetitive behavior in autism spectrum disorder. *NeuroImage: Clinical*, 10:36–45, 2016.

Chen, R., Jiao, Y., and Herskovits, E. H. Structural mri in autism spectrum disorder. *Pediatr Res*, 69 (5 Pt 2):63R–8R, 2011.

Craddock, C., Sikka, S., Cheung, B., S. Ghosh, S., Khanuja, R., Yan, C., Li, Q., Lurie, D., Vogelstein, J., Burns, R., Colcombe, S., Mennes, M., Kelly, C., Di Martino, A., Castellanos, F. X., and Milham, M. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). *Neuroinformatics*, 2013.

Di Martino, A. and al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 2014.

Di Martno, A. and al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific Data*, 2017.

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L., Group, B. D. C., et al. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*, 54(1):313–327, 2011.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C., and Collins, D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, (47):S102, 2009.

Goldani, A. A., Downs, S. R., Widjaja, F., Lawton, B., and Hendren, R. L. Biomarkers in autism. *Frontiers in psychiatry*, 5:100, 2014.

Goodfellow, I. J., Warde-Farley, D., Courville, A., and Bengio, Y. Maxout networks.

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, 17: 16–23, 2018.

Jiao, Y., Chen, R., Ke, X., Chu, K., Lu, Z., and Herskovits, E. H. Predictive models of autism spectrum disorder based on brain regional cortical thickness. *Neuroimage*, 50(2):589–599, 2010.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Reichow, B. and Wolery, M. Comprehensive synthesis of early intensive behavioral interventions for young children with autism based on the ucla young autism project model. *Journal of Autism and Developmental Disorders*, 39(1), 2009.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

Towards Autism detection on brain structural MRI scans using deep unsupervised learning methods

Mélanie Garcia, Hypercube Institute Paris, Trinity College Dublin, melanie.garcia@institut-hypercube.org
Jean-Marc Orgogozo, Hypercube Institute Paris, jean-marc.orgogozo@institut-hypercube.org
Clare Kelly, Trinity College Dublin, IMMALab, kellyc58@tcd.ie
Margaux Luck, University of Montréal, Mila, margaux.luck@gmail.com

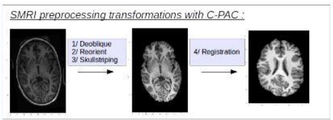
Abstract

Autism Spectrum Disorder (ASD) is a relatively common neurodevelopmental condition that for which we currently lack any objective biomarkers. The study of patient brain MRI data has the potential to reveal regions of dysfunction that may serve as biomarkers to supplement current clinician-based diagnoses. In this paper, we propose **a method that enhances the diagnosis of ASD by compressing structural MRI from the open science Autism Brain Imaging Database Exchange (ABIDE, 892 ASD, 972 non-ASD) to obtain a representation of the brain that is relevant for the prediction of ASD using unsupervised deep learning models.** Our experimental evaluation demonstrates promising performance on the task of automated ASD diagnosis on ABIDE.

Data

- Autism Brain Imaging Database Exchange I and II [1]
- Structural MRI
- Preprocessing with C-PAC*
- 1607 and 257 patients (86%-14%) in the train and the test set
- 50% with ASD

*Configurable Pipeline for the Analysis of Connectomes


SMRI preprocessing transformations with C-PAC :

Grid Sampling

- scan shape after preprocessing : **91x109x91**
- window size **32x32x32**

MRI compression, features extractor

- Variational AutoEncoder (**VAE**)
- Adversarially Learned Inference (**ALI**) [2]

Classifiers

- Majority Rule
- Elasticnet Logistic Regression (L1 ratio : 0,5)
- One Layer Perceptron

Framework

- **Niftynet** 0.6.0 [3], TensorFlow
- code on **https://github.com/blackPacha/MedNeurIPS_2019**
- VAE and ALI trained on 10 epochs, converged in 5 epochs

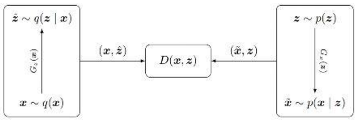
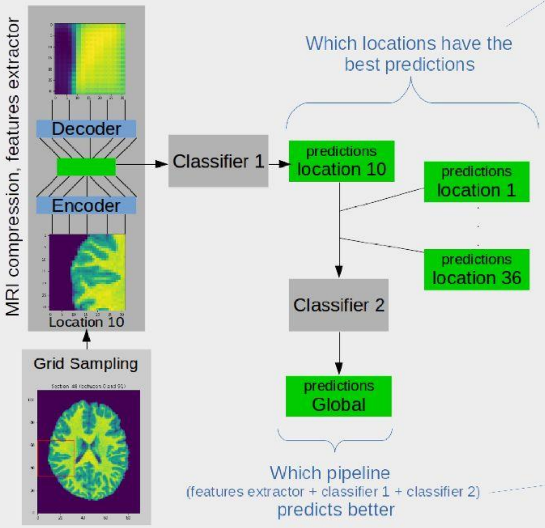


Figure 1: The adversarially learned inference (ALI) game.


MRI compression, features extractor

Decoder

Encoder

Location 10

Grid Sampling

Classifier 1

predictions location 10

predictions location 1

predictions location 36

Classifier 2

predictions Global

Which locations have the best predictions

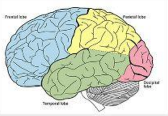
Which pipeline (features extractor + classifier 1 + classifier 2) predicts better

Results 1

grid sampling ~ brain lobe sizes

- VAE : occipital, frontal lobes
- ALI : occipital, parietal lobes
- ER : occipital, frontal, temporal, parietal lobes and cerebellum
- MLP : temporal lobe and cerebellum

Not influencing the prediction : several zones in the frontal, parietal lobes, and in the cerebellum



⇒ Different locations according to pipelines

Results 2 : ROC AUC and accuracy for each pipeline

Unsupervised extractor	VAE						ALI					
Classifier 1	Elasticnet			OLP			Elasticnet			OLP		
Classifier 2	MR	ER	OLP	MR	ER	OLP	MR	ER	OLP	MR	ER	OLP
Train	56 %	0,64 (59%)	0,65 (59%)	53 %	0,73 (66%)	0,74 (66%)	55 %	0,66 (60%)	0,66 (61%)	53 %	0,57 (55%)	0,60 (55%)
Test	54 %	0,51 (51%)	0,51 (51%)	50 %	0,5 (47%)	0,51 (50%)	44 %	0,50 (51%)	0,51 (51%)	51 %	0,49 (49%)	0,48 (49%)
Per site :												
ABIDE II, Inst. Pasteur and R. Debré hospital 1	52 %	0,40 (46%)	0,40 (46%)	**62 %**	0,41 (52%)	0,45 (54%)	38 %	**0,63 (52%)**	0,63 (52%)	**62 %**	0,47 (38%)	0,41 (38%)
ABIDE I, Social Brain Lab	63,3 %	0,76 (67%)	**0,76 (70%)**	50 %	0,56 (47%)	0,55 (50%)	47 %	0,56 (47%)	0,58 (43%)	50 %	0,50 (50%)	0,45 (50%)
ABIDEII, Univ. of California Davis 1	54 %	0,47 (46%)	0,46 (46%)	46 %	0,42 (39%)	0,46 (46%)	43 %	0,48 (46%)	0,48 (46%)	46 %	**0,61 (57%)**	0,45 (57%)
ABIDE I, Univ. of California Los Angeles, sample 2	38 %	0,57 (54%)	0,57 (50%)	50 %	0,53 (50%)	0,46 (50%)	46 %	0,46 (50%)	0,46 (46%)	50 %	**0,72 (62%)**	**0,79 (50%)**
ABIDE I, Univ. of California Los Angeles, sample 1	57 %	0,45 (50%)	0,45 (50%)	43 %	0,55 (50%)	0,57 (51%)	60 %	0,57 (60%)	0,57 (60%)	43 %	0,56 (54%)	**0,62 (57%)**
ABIDE II, Erasmus MC 1	53 %	0,57 (51%)	0,57 (51%)	51 %	0,45 (43%)	0,48 (47%)	27 %	0,40 (43%)	0,41 (45%)	**57 %**	0,42 (43%)	0,46 (41%)

Each case : ROC AUC score (Accuracy %)
MR = Majority Rule, accuracy only
ER = Elasticnet Regression
OLP = One-Layer Perceptron

⇒ Differences between data collection sites
⇒ Differences between pipelines
⇒ best ROC AUC scores range [0,6 – 0,8] comparable or better than in the litterature

Conclusion

PoC : deep learning could be useful for ASD diagnosis
⇒ could raise new paths of research on finding Autism neurological markers
Pipeline on brain sMRI providing weak locations of the brain areas involved in good prediction of ASD

Limitations & Future work

- findings very preliminary
- should be compared with more classic models as baselines on ABIDE I and II
- should take into account the data collection sites and other modalities like functional MRI during training

References
[1] ABIDE data collection :
« The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. » Di Martino et al., Molecular Psychiatry 2014
« Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. » Di Martino et al., Scientific data 2017
[2] ALI : *« Adversarially Learned Infeence »* Dumoulin, Belghazi et al., ICLR 2017
[3] NiftyNet : *« NiftyNet : a deep-learning platform for medical imaging. »* Gibson, Li et al., Computer Methods and Programs in Biomedecine, volume 158, May 2018, pages 113-122

Workshop Medical Imaging meets NeurIPS 2019 – 14th December 2019