10. Discussion

This PhD thesis aimed to develop interpretable Deep Learning models for identifying 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 tested specific hypotheses about brain imaging patterns in ASD. Additional open science works were made by developing standards and tutorials. While revealing limitations, this work provided initial methods and justification for biologically-grounded AI to elucidate neural organisation of Autism. The identified limitations led to recommendations to address challenges related to model interpretation, biases, and optimization. Overall, this research developed an explainable imaging analysis framework to quantify and discover the heterogeneous neurobiological underpinnings of ASD in a clinically meaningful way.

10.1. Summary

Three main projects have been done around this topic in particular, and other side projects on developing reproducible and ethical neuroimaging scientific practices have been done in parallel all along the PhD years.

10.1.1. Project 1:

Manual quality control of structural MRI data is essential but time-consuming. To address this, we developed an interpretable Deep Learning 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 datasets - ABCD (2141 scans), ADHD200, and ABIDE II - demonstrating excellent performance with 91.4% sensitivity for detecting artefacts compared to human raters of ABCD. Critically, BrainQCNet showed higher sensitivity than previous methods while requiring no intensive scan preprocessing. However, some detected patterns required further examination to make the tool clinically relevant. In particular, at a local level, it was not clear if all the patterns detected by the model were relevant or not for the prediction. To support open adoption, several BIDS apps implementing BrainQCNet on GPU and CPU systems were developed. All the code was publicly released on GitHub under an open licence. Overall, this project showed that DL can fastly automate and enhance sMRI quality control to improve the reliability of downstream analyses.

10.1.2. Project 2:

Standard neuroimaging pipelines rely on intensive preprocessing like spatial normalisation that may hide small 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 were compared, DenseNet121 and ResNet50, trained and tested across multiple datasets including ABIDE 1 and 2, and ADHD200. Testing over these various datasets provided more robust results. The models achieved 50-70% prediction accuracy for Autism, lower with comorbid conditions. Using guided grad-CAM for 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. A granular analysis also revealed some differences in predictive regions depending on gender and age. Critically, models did not rely on non-brain background. By avoiding potentially biased preprocessing while revealing interpretable neuroimaging patterns, this work provides DL imaging biomarkers for Autism. The multimodal integration and validation across datasets bolsters generalizability. Globally, the project advances the development of ML for Autism diagnosis, taking into account confounds in the interpretation, while moderating the risks of standard processing pipelines

10.1.3. 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 [X] subjects to classify Autism and get 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/optimisations were tested using 100-fold cross-validation, including multitask models, that were an original methodological approach from our study, that also predicted gender and age. Cross-validation constrained model exploration but improved evaluation. All the models achieved approximately [Y]% accuracy but overfitted the training data, likely from limited data and model complexity. To interpret learned representations, chord diagrams were plotted. I also implemented LIME, a local explanation method, to explain individual predictions. However, LIME explanation was limited by the high dimensionality of fMRI data, preventing to perform a quantitative analysis across subjects. Further hyperparameter optimization and regularization may reduce overfitting and improve generalizability. Though predictive performance was modest, this novel application of Transformers with multitask learning to fMRI data demonstrated a potential to discover “human invisible” imaging biomarkers of Autism informed by neuroscience priors.

10.1.4. Side projects and trainings

Early in the PhD, it was recognized that reproducibility is in crisis in neuroimaging. Action was taken through initiatives like co-leading the first Reproducibility Journal Club in Ireland. Tools like BIDS for standardised data organisation were championed by contributing BIDS tutorials and developing multiple BIDS apps to enhance reusability of the projects. Educational materials on leveraging GitHub for transparent, collaborative coding were also created. 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.

In addition, 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 upcoming NeuroHackademy to expand Data Science skills in Neuroscience. These training programs helped me to gain strong scientific engagement and commitment to lifelong knowledge growth.

Early in the PhD, it was recognized that there is a reproducibility crisis in neuroimaging. I took action like co-leading the first Reproducibility Journal Club in Ireland. Tools like BIDS for standardised data organisation were valorised by contributing BIDS tutorials and developing multiple BIDS apps to enhance reusability of the projects. Educational materials on leveraging GitHub for transparent, collaborative coding were also created. These various efforts to promote open science led me to be invited to author book chapters providing practical guidance on BIDS and GitHub for the wider neuroscience community.

In addition, I actively kept on developing professionally 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 skills; the ECNP neuropsychiatry program to learn about an emergent research area; the ARAPI, an Autism research conference, to connect with diverse stakeholders; and NeuroHackademy to expand Data Science skills in Neuroscience. These training programs helped me to gain strong scientific engagement and commitment to knowledge growth.

10.2. Interpretations

Several key hypotheses motivated the work on developing predictive models for ASD using neuroimaging data and DL. 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 DL. A fourth hypothesis was that while autistic individuals show different severities of symptoms, they share common characteristic patterns in the brain that can be captured by models. Finally, the hypothesis was that current neuroimaging data variability is sufficient to train a robust ASD prediction algorithm generalizable 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 or not, suggesting future additional experiments to further evaluate the potential for brain-based ASD prediction.

The first study showed that deep learning can rapidly automate MRI quality control - a crucial preprocessing step where manual classification is repetitive and time-consuming. Our attention-based BrainQCNet model achieved excellent global performance detecting artefacts, underscoring how deep learning can augment human annotators for simple but tedious neuroimaging tasks. However, interpretation remains challenging; while overall accuracy was high, local model behaviour showed both realistic and unrealistic patterns, highlighting the difficulty of explaining complex artificial neural networks. Additional optimization and experiments are needed to improve local-level accuracy and model understanding. Nevertheless, initial results were promising - the large training data yielded a rather robust global classifier, though broader artefact diversity could further enhance generalizability and reduce False Negatives. On the whole, this study established deep learning, especially attention-based architectures, as a trustable approach to automating certain MRI preprocessing steps.

This argument is reinforced by the fact that other studies showed deep learning as a good tool for automating preprocessing pipelines [1-7]. For instance, [1] showcases the application of deep learning for noise reduction, which is a critical preprocessing step in MRI. [2] proposed the use of CNNs for super-resolution in diffusion MRI (dMRI), automating the enhancement of MRI resolution. [7] detailed a deep learning approach for automated brain extraction, which is a fundamental step in many MRI-based studies.

All together, these references provide a strong foundation to support the idea that deep learning is not only a viable approach to automate MRI preprocessing steps but is also achieving state-of-the-art results in various tasks. Moving forward, pairing strong global performance with refined model interpretability remains a key challenge as we scale up deep learning for improved neuroimaging workflows.

While showing promise, the predictive modelling studies in this thesis revealed current limitations in using structural and functional MRI alone for robust Autism detection. Despite finding some consistent regional patterns aligned with Autism neuroscience, overall performance was modest across multiple algorithms, with classification accuracy ranging from 50-70% and high variability based on factors like age, gender, and site. Several studies showed similar accuracies [8-12], but more recent ones showed higher accuracy range from 75-85% Those represent studies that have achieved moderate to high success but still highlight the challenges inherent in MRI-based classification. In particular, it is crucial to understand that the accuracy may vary based on the methodologies, algorithms, and datasets used in the research.

The results in this thesis and in the literature underscore the diversity of neural phenotypes in Autism and suggest current neuroimaging biomarkers lack the specificity and precision to generalize broadly, especially when there are comorbidities. Other research work also faced these challenges. In their review, [18] emphasized the heterogeneity in neuroimaging findings in Autism and the need for a more nuanced understanding that get the diversity of neural phenotypes. [20] illustrated the considerable variability in anatomical findings across different MRI studies of autism, raising questions about the reliability and specificity of potential biomarkers. [24] pointed out the high prevalence of psychiatric comorbidity in autism, suggesting that these conditions could confound neuroimaging findings.

However, the results in this thesis also show that refined analysis of important features highlighted potential shared neural signatures within the heterogeneity. This motivates to build larger datasets with greater representation.

Critically, each model architecture impacted the results, indicating the importance of multi-model convergence to mitigate individual algorithm biases. When isolated, a single model biases can dominate, but combining diverse architectures can reveal more robust generalizable biomarkers. [26] a foundational book on DL, talks about the benefits of ensemble methods in various sections. And longer before, [29] discussed the advantages of using ensembles of ANN to improve generalization performance. [33] focused on situations when individual neural networks provide conflicting outputs and how ensemble methods can help in such cases. By leveraging the combined strengths and mitigating the individual disadvantages of multiple models, ensemble methods can indeed provide more robust and accurate predictions in DL.

10.3. Implications of the PhD project

In parallel to the core modelling projects, this thesis work greatly incorporates open science practices in neuroimaging to maximise research impact. Broad self-training in programming, information technology tools, and artificial intelligence increased my expertise that was used across studies. Participating in AI and Health summer schools also enabled mastering techniques in responsible and interpretable deep learning.

Significant effort went into developing BIDS-apps implementing models on GPU systems using CUDA/CuDNN, overcoming technical hurdles to share reproducible code. All code was openly available on GitHub when studies were published to facilitate adoption. In addition, tutorials created on BIDS and on Git and GitHub lower entry barriers so more neuroscientists can leverage these platforms.

Overall, this thesis demonstrated the feasibility of interpreting deep learning models and building ethical, responsible AI algorithms aligned with community needs. The integration of open science principles follows FAIR data instructions, enhancing discovery and collaboration. This multi-faceted approach combining methodological advances with openness and ethics shows how to construct neuroimaging AI to benefit the Autism community. The long-term impact is in the novel techniques introduced, as well as my commitments to openness, outreach, and conscientious application.

10.4. Limitations of the PhD project

Our work established a novel methodological foundation but highlights significant challenges remaining in Autism prediction from MRI.

10.4.1. On interpreting and explaining DL models

While study 2 developed an initial pipeline for interpreting predictive brain regions, our work revealed significant lingering challenges in evaluating and quantifying uncertainty within deep learning models. The high complexity of modern artificial neural networks often renders their inner workings opaque and decision-making inscrutable, even for developers. This "black box" nature makes quantitative analysis of algorithm behavior 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” ANNs (Ribeiro et al., 2016; Lundberg & Lee, 2017; Sundararajan et al., 2017), but currently lack scalability and standardization.

Equally crucial is quantifying uncertainty - conveying when predictions may be unreliable. Study 3’s extensive cross-validation enabled better accuracy estimation and confidence intervals around model performance. [42] shows the importance of uncertainty in DL models, especially in domains like healthcare where a wrong decision can have dire consequences. 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 to not just advance predictive performance of DL in neuroscience but crucially, also boost model transparency, accountability, and 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.

10.4.2. On preprocessing pipelines

Our work reveals the complex double impacts of neuroimaging preprocessing on downstream DL analysis. Study 1 demonstrated the value of rigorous quality control by developing a model to accelerate critical MRI artefact detection. However, common preprocessing steps like spatial normalisation, smoothing, and registration make assumptions about typical anatomy that risk distorting or hiding small morphological features associated with ASD. In Study 2, the use of raw structural MRI data as input avoided such risks that could constrain detection of atypical patterns. However, this raw data approach then limited biological interpretability of the learned features and biomarkers important for model predictions, since no anatomical context was provided.

This show the inherent trade-off between keeping natural, unbiased brain signatures in the data and getting specific anatomical meaning needed to relate findings to clinical traits and neuroscience knowledge. While Deep Learning is good at extracting signals directly from minimally processed data, relating the discovered patterns to tangible biological insights and clinical use is still essential for practical adoption. [37] talks about the convergence of AI and human intelligence in Medicine and highlights the importance of interpretability and clinical relevance.

Our focus on only MRI data alone also limited full understanding of the predictive features and biomarkers detected by models. Multimodal integration of neuroimaging with genetics, cognitive tests, and clinical assessments looks critical for grounding Deep Learning models in biological mechanisms relevant to heterogeneous conditions like Autism. Other studies pointed out the importance of multimodal data like [45] that show need to integrate genomics with functional data to elucidate the pathways and circuits implicated in autism, and the review [43] that emphasised the importance of multimodal data fusion, especially in capturing more complex, high-dimensional representations of the brain.

Overall, our 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 human-driven neuroscience.

10.4.3. On dataset biases

Our work show big risks of bias amplification and skewed representation in current neuroimaging datasets that Deep Learning algorithms could potentially make worse. Crossing MRI data with genetic information and splitting analyses by genetic clusters is essential for building fair models tuned to diverse populations rather than just overrepresented subgroups. However, most datasets like ABIDE have limited genetic data. The available samples also have concerning demographic representations—for example, in ABIDE datasets, too many younger male subjects risks models tuned only to this group. While multi-site data pooling has enabled larger samples, variability in scanning protocols across scanners can further confuse analysis.

More worrying is the limited description of clinical, behavioural, and phenotypic traits with neural data. Details on medications, co-occurring conditions, symptom profiles, and more are often not reported, yet crucial for relating brain patterns to real life functioning. Such issues likely partly result 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, life-long condition, studying static data slices alone also risks overlooking important developmental trajectories.

Ultimately, creating more balanced, representative datasets will require active involvement of autistic people and advocates in protocol co-design. This could usefully 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 Deep Learning to complement today’s small homogeneous samples with fair insights benefitting the whole community.

10.4.4. On Deep Learning

Deep learning's explosive grow in model complexity introduce new challenge in robust training and generalisation. While our datasets were large compare to many studies on Autism, several models still show overfitting - hinting insufficient diversity to capture heterogeneous neurological conditions. Aligning with our work, multiple recent studies estimate sample sizes well into tens or hundreds thousands are necessary for reliable train deep neural network 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 vast tuning space. However, exhaustive tuning risk simply overfit to quirks of limited datasets rather than learn generalizable and replicable patterns that transfer robust to new out-of-sample cases.

Our work thus underscore 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. 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. 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 conditions like Autism.

Overall, our studies reveal current limitations but outline research program for progress through bigger, more varied dataset creation, integrated predictive modelling, and grounding in behavioural dimensions.

Though significant challenges remain, this PhD work develop core justifications and methods to pursue biologically-grounded DL for elucidating Autism complex neural correlates in clinically meaningful way.

10.5. Recommendations

The limitations shown in the studies point to several recommendations for advancing biologically-grounded AI modelling of ASD using neuroimaging data:

- Integrate multimodal data beyond just MRI, including genetics, cognition, and clinical assessments, to enhance biological interpretation. Fusing neuroimaging with broader biological and phenotypic data can help to learn patterns having clinical meanings.
- Contextualize studies with more specific inclusion criteria if dataset size is limited. Focusing on targeted demographic or behavioural factors can reduce heterogeneity and improve characterization of neural correlates within defined ASD context.
- Test diverse DL model architectures for any predictive modelling task. Varying approaches mitigates individual algorithm biases and enables convergence on most robust generalizable patterns.
- Employ statistical methods such as N-fold cross-validation frameworks to rigorously evaluate model performance and uncertainty. However, balance model exploration time with 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 related 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 advancing fastly, requiring lifelong learning to apply them effectively in multidisciplinary research.
- Carefully evaluate the ethical implications of AI techniques before application in Autism research or care. Ensure models are transparent, fair, and designed to safely complement clinicians rather than replace them.

Following these recommendations can promote development of more reliable, interpretable, and clinically useful AI models of ASD using brain imaging and related data. By considering ethical concerns in parallel to methodological advances, research should lead to more responsible AI to benefit the Autism community.

Conclusion

This thesis presents an original exploration in 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, initial methods were developed and tested against specific hypotheses related to viability of brain imaging for study Autism heterogeneity.

While falling in robust prediction and revealing significant limitations, these projects highlight future pathways 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 establishes justification and initial methodology for biologically-grounded AI modelling to quantify and interpret complex neural patterns of Autism traits. The limitations identified motivate specific recommendations to overcome current challenges in explainable and equitable neuroimaging analysis. By depicting this groundwork and direction for the field, this thesis is a springboard for future efforts to refine data-driven imaging biomarkers that can translate to enhanced clinical insights and precision care for autistic individuals.