10. Discussion

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 tested specific hypotheses about brain imaging patterns in ASD. Additional open science contributions were made through developing standards and educating peers. While revealing limitations, this work provided initial methods and justification for biologically-grounded AI to elucidate neural correlates of Autism. The identified limitations motivated recommendations to address challenges related to model interpretation, biases, and optimization. Overall, this research developed an explainable imaging analysis framework to quantify and elucidate 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 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 datasets - ABCD (2141 scans), ADHD200, and ABIDE II - demonstrating excellent performance with 91.4% sensitivity for detecting artefacts compared to human raters on ABCD. Critically, BrainQCNet showed higher sensitivity than previous methods while requiring no intensive scan preprocessing. However, some detected patterns required further examination for clinical relevance. 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 code was publicly released on GitHub under an open licence. Overall, this project showed DL can rapidly automate and enhance sMRI quality control to improve the reliability of downstream analysis.

10.1.2. Project 2:

Standard neuroimaging pipelines rely on intensive preprocessing like spatial normalisation that 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 were compared, DenseNet121 and ResNet50, trained and tested 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 with comorbid conditions. 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. By avoiding potentially biassed preprocessing while revealing interpretable neuroimaging patterns, this work provides clinically-grounded DL biomarkers for Autism. The multimodal integration and validation across datasets bolsters generalizability. Overall, the project advances biologically-informed ML for Autism diagnosis while mitigating 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 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 [Y]% accuracy but overfitted training data, likely from limited data and model complexity. To interpret learned representations, chord diagrams showed models partially captured functional connections aligned with Autism neuroscience. 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 generalizability. 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.

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.

10.2. Interpretations

Several key hypotheses motivated the work on developing predictive models for ASD using neuroimaging data and DL. First, the hypothesis was that structural MRI data alone contains sufficient precision to build a predictive model of ASD. Second, the hypothesis was that functional MRI data alone also holds adequate specificity for building an ASD prediction model. Third, the hypothesis was that the brain is a relevant variable for studying ASD, with diagnostic neuroimaging biomarkers detectable through ML. Fourth, the hypothesis was that while autistic individuals express 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 refuted, driving 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 where manual classification is repetitive and time-consuming. Our attention-based BrainQCNet model achieved excellent global performance detecting artefacts, underscoring how DL 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

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 generalizability and reduce False Negatives. On the whole, this study established DL, especially attention-based architectures, as a viable approach to automating 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 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.
These references collectively provide a strong foundation to support the idea that deep learning is not only a viable approach for automating MRI preprocessing steps but is also achieving state-of-the-art results in various tasks. Moving forward, pairing strong global performance with granular model interpretability remains a key challenge as we scale up DL for enhanced 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 generalise broadly, especially amidst comorbid conditions. Other research work also faced these challenges. In their review, [18] emphasised the heterogeneity in neuroimaging findings in Autism and the need for a more nuanced understanding that captures 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 granular analysis of important features hinted at potential shared neural signatures within the heterogeneity, motivating larger datasets with greater representation.

Critically, each model architecture impacted 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 generalizable biomarkers. [26] a foundational book on DL, touched upon the benefits of ensemble methods in various sections. And long before, [29] discussed the benefits of using ensembles of ANN to improve generalisation performance, and [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 weaknesses of multiple models, ensemble methods can indeed provide more robust and accurate predictions in DL.

10.3. Implications of the PhD project

Alongside the core modelling projects, this thesis work strongly embraced open science practices in neuroimaging 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 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 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.

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 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 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-edged 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 obscuring subtle morphological features associated with ASD. Study 2's use of raw structural MRI data as input avoided such pitfalls that could constrain detection of atypical patterns. However, this raw data approach then limited biological interpretability of the learned features and biomarkers driving model predictions, since no anatomical context was provided.

This illustrates the inherent trade-offs between preserving naturalistic, unbiased brain signatures in the data and gaining specific anatomical meaning needed to relate findings to clinical traits and neuroscientific knowledge. While DL thrives on extracting signals directly from minimally processed data, relating the discovered patterns to tangible biological insights and clinical utility remains essential for practical adoption. [37] discusses the convergence of AI and human intelligence in Medicine and highlights the importance of interpretability and clinical relevance.

Our monomodal focus on MRI data in isolation also constrained 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 conditions like Autism. Other studies pointed out the importance of multimodal data like [45] that exemplified the 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, rather than diverges from, human-driven neuroscience.

10.4.3. On dataset biases

Our 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 for building equitable models tuned to diverse populations rather than just overrepresented subgroups.

However, most datasets like ABIDE contain limited genetic data. The available samples also have concerning demographic representations—for example, in ABIDE datasets, heavy preponderance of younger male subjects risks models tuned only to this group. While multi-site data pooling has enabled larger samples, variability in acquisition protocols across scanners can further confound analysis.

More concerning is the limited characterization of clinical, behavioural, and phenotypic traits alongside neural data. Details on medications, comorbid conditions, symptom profiles, and more 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 condition, 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. 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, our work strongly motivates interweaving cutting-edge modelling with participatory data improvements to ensure neuroimaging AI meaningfully serves diverse populations.

10.4.4. On Deep Learning

Deep learning's explosive growth in model complexity introduces new challenges in robust training and generalisation. While our datasets were large compared to many studies led on Autism, several models still demonstrated overfitting - hinting insufficient diversity to capture heterogeneous neurological conditions. Aligning with our work, multiple recent studies have estimated 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 generalizable and replicable patterns that transfer robustly to new out-of-sample cases.

Our 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. 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 revealed current limitations but outlined a 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 developed core justifications and methods to pursue biologically-grounded DL for elucidating Autism's complex neural correlates in a clinically meaningful way.

10.5. Recommendations

The limitations revealed 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 ground learned patterns in tangible clinical meanings.
- Contextualise 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 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 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 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.
- Carefully evaluate the ethical implications of AI techniques prior to application in Autism research or care. Ensure models are transparent, fair, 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.

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, initial methods were developed and tested against specific hypotheses related to the viability of brain imaging for studying Autism heterogeneity.

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