Explainable and Generalized Deep Learning Framework: Study on Atypical Brain Network Development

A Thesis submitted by
Km Bhavna

in partial fulfillment of the requirements for the award of the degree of
Doctor of Philosophy



Indian Institute of Technology Jodhpur
Department of Computer Science and Engineering
April 2025

8
Conclusions

8.1 CONCLUSION

This thesis primarily focuses on developing an end-to-end explainable and generalized AI framework to analyze social brain network development from early childhood to adulthood in neuro-typicals as well as autistics. It also aims to identify the brain fingerprints feature set as a biomarker for prediction and classification. Deep learning has several benefits in the analysis of brain network connectivity. Deep neural networks were first derived from the neurological system and intended to replicate the brain's information processing. Therefore, it may enhance our comprehension of the human brain. Furthermore, it is driven by specific objectives and trained using a substantial volume of data, reducing human mistakes considerably. This thesis primarily focuses on examining social brain networks using functional connectivity. "functional connectivity" often refers to the dependency of different brain regions' activity. The social brain network may be a valuable biomarker for distinguishing between atypical and healthy brain network development. Furthermore, it can be treated as data for further examination in clinical decision-making. The contribution of the thesis is as follows:

Chapter 3 aims to identify temporal stability of ToM network from early childhood (i.e., 3 yrs) using movie-watching stimuli, and if the stimuli contain activation of two separate networks, then identifying the age at which ToM network is getting considerable temporal stability. To validate our hypothesis, here we used a dataset of 155 subjects (age range from 3yrs to adults) who were watching a naturalistic stimulus movie that was validated to activate ToM and pain sensory networks. After fMRI signal acquisition, ToM-related false-belief was also performed, and according to performance, participants were categorized into pass, inconsistent, and fail groups. To identify the change in temporal stability of ToM and pain networks over time or early childhood to adulthood, we calculated Angular and Mahalanobis distance between dominant dynamic functional connectivity subspaces and validated results using entropy and other measures. Our results suggested that ToM was getting segregated from other networks at early as 3-yrs, and getting stabilized from 5-yrs, and that continued till adulthood. We hypothesized that temporal stability could be a predictive measure for performance in false-belief tasks. Our results support the proposed hypothesis. We observed a change in false-belief task performance with respect to angular and Mahalanobis distance matrices.

Chapter 4 shows the AI framework to decode how social cognition tasks have been processed from early childhood to adulthood. The brain-computer interface (BCI) provides a platform for decoding brain tasks. It aims to identify brain states using neural activity, further helping understand behavior crucial for cognitive neuroscience and neural engineering. We proposed an Explainable spatiotemporal connectivity-based Graph Convolutional Neural Network (Ex-stGCNN) model for decoding of brain tasks. Here, we consider same dataset that was used in chapter 3. We trained our proposed model using Static Functional Connectivity (SFC) and Inter-Subject Functional Correlations (ISFC) matrices separately. We observed that the stimulus-driven feature set (ISFC) could capture ToM and Pain brain states more accurately with average accuracy

of 94%, whereas achieved 85% accuracy using SFC matrices. We also validated our results using five-fold cross-validation and achieved an average accuracy of 92%. We applied SHAP approach to identify neuro-biological brain fingerprints that contributed the most to predictions. We hypothesized that ToM network brain connectivity could predict individual performance on false-belief tasks. We proposed an Explainable Convolutional Variational AutoEncoder model using functional connectivity to predict individual performance on false-belief tasks. We achieved 90% accuracy. This framework not only decoded brain states from developmental age groups and highly imbalanced dataset with high accuracy from short-time course data but also predicted individual performance in false-belief tasks and categorized participants into pass, fail, and inconsistent groups.

Chapter 5 shows an explainable and generalized AI framework to predict how social networks are developing in autistics and neuro-typicals, as well as identified sources of variability between autistics and neuro-typicals. Theory-of-Mind (ToM) is an evolving ability that significantly impacts human learning and cognition. Early development of ToM ability allow one to comprehend other people's aims and ambitions, as well as thinking that differs from one's own. Autism Spectrum Disorder (ASD) is the prevalent pervasive neurodevelopmental disorder in which participants' brains appeared to be marked by diffuse variations throughout large-scale brain systems made up of functionally connected but physically separated brain areas that got abnormalities in willed action, self-monitoring and monitoring the intents of others, often known as ToM. Although functional neuroimaging techniques have been widely used to establish the neural correlates implicated in ToM, the specific mechanisms still need to be clarified. The availability of current Big data and Artificial Intelligence (AI) frameworks paves the way for systematically identifying Autistics from typically developing by identifying neural correlates and connectome-based features to generate accurate classifications and predictions of socio-cognitive impairment. We develop an Ex-AI model that quantifies the common sources of variability in ToM brain regions between typically developing and ASD individuals. Our results identify a feature set on which the classification model can be trained to learn characteristics differences and classify ASD and TD ToM development more distinctly. This approach can also estimate heterogeneity within ASD ToM subtypes and their association with the symptom severity scores based on socio-cognitive impairments. Based on our proposed framework, we obtain an average accuracy of more than 90 % using Explainable ML (Ex-Ml) models and an average of 96 % classification accuracy using Explainable Deep Neural Network (Ex-DNN) models. Our findings identify three important sub-groups within ASD samples based on the key differences and heterogeneity in resting state ToM regions' functional connectivity patterns and predictive of mild to severe atypical social cognition and communication deficits through early developmental stages.

Chapter 6 shows an explainable and generalized AI framework to classify autistics from neuro-typicals using meta-connectivity as a feature set and to identify hidden sub-groups present within the autistic samples and their symptom severity scores. Autism spectrum disorder (ASD) is a pervasive neurodevelopmental disorder associated with difficulties in social communication, interaction, and restricted/repetitive behaviors or activities that emerge in early childhood and endure into adulthood. The current era, marked by the availability of big data and Artificial Intelligence (AI) frameworks, provides an unprecedented opportunity to leverage connectome-based features associated with ASD. However, despite recent neuroscience research emphasizing machine learning (ML) and deep neural network (DNN) frameworks to classify ASD from neuro-typicals (TD), the specific mechanisms still need to be clarified for non-Euclidean space. The existing ML/DNN approaches focus on node-centric features; as a result, performance is moderate on unseen data due to ignorance of edge-centric features. In the current study, we proposed a lightweight, explainable, generalized DNN framework, in which we used Autism Brain Imaging Data Exchange (ABIDE I and ABIDE II (Total no. of sample 1500)) dataset and considered Theory-of-Mind, Default-mode Network, Central executive, and Salience networks from which calculated meta-connectivity as edge-centric feature set which measured the correlation between

edges associated with mentioned brain networks. Then, to classify ASD samples from TD samples, we trained an Explainable and Generalized node-edge connectivity-based graph attention neural network (Ex-NEGAT) on ABIDE I dataset (No. of samples =840) and performed testing on ABIDE II (no. of samples =660). We achieved 88% accuracy with F1-score of 0.89 on large unseen dataset. We also identified symptom severity scores for each individual using meta-connectivity using Connectome-based Prediction Modelling (CPMM) and identified brain networks that contributed positively and negatively to the prediction. Our approach accurately predicted ADOS-Total, ADOS-Social, ADOS-Communication, ADOS-Module, ADOS-STEREO, and FIQ scores.

Chapter 7 shows identification of sub-groups within ASD samples using Rough Fuzzy C-Means Clustering approach. Autism spectrum disorder (ASD) is a neurological condition characterized by deficits in social cognition, interaction, communication, restricted behaviors, and sensory abnormalities. Using fMRI, it has been found that Theory-of-Mind (ToM), Default-Mode Network (DMN), Central Executive Network (CEN), and Salience Network (SN) brain networks are involved in the case of autism. ASD's diverse clinical presentation complicates its diagnosis and treatment, highlighting the need to understand its heterogeneity better. This study aims to identify a novel feature set from resting-state fMRI data to enhance understanding of ASD's heterogeneity. By focusing on edge connectivity (Meta connectivity) as higher-order correlations among ToM, DMN, CEN, and SN brain networks, we use the Rough Fuzzy C-Means (RFCM) algorithm to identify biomarkers for subgrouping ASD individuals. The study leverages the Autism Brain Imaging Data Exchange (ABIDE) datasets, analyzing dynamic functional connectivity (dFC) and extracting meta-connectivity (MC) matrices for a detailed examination of brain topology. Using the proposed approach, we identified three sub-groups within the samples and their association with each other. We found that 1st and 3rd sub-groups contained unique samples with less overlapping, whereas the 2nd sub-group was more associated with 1st sub-groups with fewer unique samples. We also identified an association between identified clusters and behavioral scores. To the best of our knowledge, the proposed work is unique and novel as there is no work that identified sub-groups using higher-order correlation as a feature set for the RFCM algorithm.

8.2 FUTURE WORK

Our current methods have not yet been applied to cognitive tasks involving Theory-of-Mind (ToM) data from both ASD and typically developing (TD) individuals. However, ongoing work in our lab is focused on this, and results will be shared in future studies. The identification of robust, individualized functional brain connectivity biomarkers in key ToM regions of ASD individuals has the potential to significantly enhance our understanding of the causes, diagnosis, and connectivity patterns in this neurodevelopmental disorder. Additionally, while our framework was tested using resting-state meta-connectivity data, future work will explore its application during cognitive tasks. There is also a need to extend our analysis to longitudinal datasets, allowing us to track brain connectivity changes over time. This will be a key focus in our upcoming research.