Abstract

Social cognition refers to the ability to understand, process, and respond to social interactions and the behaviors of others. The Theory-of-Mind (ToM) brain network, associated with social cognition, is responsible for understanding another person's intentions and ideas, comprising regions such as the medial prefrontal cortex, temporoparietal junction, and superior temporal sulcus. These areas are crucial for recognizing and interpreting others' thoughts, intentions, and emotions. Recent studies highlight significant interactions between the Fronto-Parietal Network (FPN) and the Temporo-Parietal Junction (TPJ), which are critical for processing social information and predicting social evaluations. Using functional magnetic resonance imaging (fMRI) data, recent research has focused on the early development of ToM in children, extending into middle childhood and adolescence. By age five, children typically develop the ability to understand others' aims and predict actions based on false-belief paradigms, marking a critical stage in ToM development. Deficits in ToM are evident in neurodevelopmental disorders like autism spectrum disorder (ASD), where individuals show significant impairments in social cognition and communication. Despite extensive research, the heterogeneity within the ASD population and incomplete understanding of its neurobiology pose challenges.

Our research focuses on quantifying the temporal stability of ToM and Pain brain networks from early childhood (3 years) to adulthood, using fMRI-based dynamic functional connectivity analyses. We investigate whether temporal stability patterns are associated with performance in false-belief reasoning tasks, particularly in children aged 3–12 years. To decode cognitive states during a naturalistic movie-watching task, we developed an explainable spatiotemporal connectivity-based graph convolutional neural network (Ex-stGCNN). We further employ an explainable convolutional variational autoencoder (Ex-Convolutional VAE) to predict individual false-belief task performance, identifying key brain regions contributing to these predictions as subject-specific neural fingerprints.

In the context of ASD, which affects social behavior and communication ability, we utilize large-scale fMRI datasets to develop an explainable deep learning framework aimed at identifying both shared and ASD-specific variability in ToM brain networks. Our models classify typically developing (TD) and ASD individuals using functional connectivity and meta-connectivity features derived from ToM, Default-Mode Network (DMN), Central Executive Network (CEN), and Salience Network (SN). Furthermore, we use these features to predict clinical scores related to ASD symptoms, including ADOS-Total, ADOS-Social, DSM-IV, and Full-Scale IQ (FIQ), through connectome-based predictive modeling (CPMM). The model evaluation demonstrated robust classification performance (AUC 0.86) and accurate symptom severity prediction with minimal error across test sets. Finally, we also identify ASD sub-groups through edge-centric meta-connectivity features using a Rough Fuzzy C-means approach. Overall, this work demonstrates the power of explainable deep learning models in bridging brain network dynamics and symptom heterogeneity in ASD, contributing to more individualized insights into atypical brain development.