7

Identification of Sub-Groups within ASD samples using Rough Fuzzy C-Means Clustering Approach

7.1 INTRODUCTION

Autism spectrum disorder (ASD) is a pervasive neurological disorder marked by deficiencies in social communication abilities and interactions, as well as limited, repetitive behavior and sensory abnormalities. According to the previous studies Baron-Cohen *et al.* [1985]; Leekam and Perner [1991]; Bhavna *et al.* [2023a]; Müller and Fishman [2018] using fMRI, it has been found that autism mainly affects the abilities of social cognition, interaction, and communication that are associated with the specific brain networks such as Theory-of-Mind (ToM) network, Default-Mode Network (DMN), Central Executive network (CEN), and Salience network (SN). Over the past few decades, the diagnosis of autism has transformed from being a narrowly defined and rare disorder that affects young children to a more inclusive spectrum disorder that presents lifelong challenges. Currently, more than 1% of children are diagnosed with autism, and it is characterized by a highly diverse nature, meaning that no two individuals with ASD have identical phenotypic profiles Lord *et al.* [2018]. Individuals with ASD exhibit multi-level heterogeneity Jeste and Geschwind [2014] in biological and behavioral characteristics, including genetics Geschwind *et al.* [2015]; Baker and Jeste [2015], neural systems Hahamy *et al.* [2015]; Hong *et al.* [2018], and clinical phenotypes. Individuals with ASD exhibit variations in their clinical characteristics, which encompass the extent of specific behavioral impairments in several dimensions, such as social skills, communication, and language proficiency, age of symptom onset, intellectual functioning Feczko *et al.* [2018], gender disparities Halladay *et al.* [2015]; Bedford *et al.* [2020]; van't Westeinde *et al.* [2020]; Lai *et al.* [2015], executive functioning Geurts *et al.* [2014], and developmental history Lord *et al.* [2015]; Pickles *et al.* [2014]. Therefore, ASD is often understood as a collection of several spectrums rather than a singular phenotype. Many clinicians and researchers hypothesize that it consists of many subgroups or multidimensional nodes or subdivisions that are believed to have potential utility in understanding the causes and management of the condition Hong *et al.* [2018]; Lombardo *et al.* [2019]; Tang *et al.* [2020]. However, these subgroups have not yet been properly recognized and defined. The significant diversity across several aspects at multiple levels is indicative of a fundamental difficulty in the neurobehavioral modeling of ASD or its subcategories. Over the past few decades, researchers have made efforts to identify subcategories of ASD by analyzing certain behavioral or cognitive traits Fountain *et al.* [2012]; Munson *et al.* [2008]; Zheng *et al.* [2020]. Nevertheless, distinct neurobiological processes may influence subgroups with consistent clinical characteristics. Hence, the discovery of neuroimaging-based neurophysiological subgroups offers a valuable understanding of how various neurophysiological parameters contribute to the diversity within the overall clinical phenotype Clementz *et al.* [2016].

The heterogeneity inherent in ASD is a significant obstacle in the prevention, diagnosis, and treatment process Lenroot and Yeung [2013]; Collins and Varmus [2015]. Several studies have indicated that utilizing a biomarker approach to categorize the variability within the ASD samples into subgroups with neurobehavioral significance could enhance the accuracy of identifying causes, early detection, diagnosis, and targeted treatment instead of using a generic approach for

all individuals Lombardo *et al.* [2019]; Tang *et al.* [2020]; Lenroot and Yeung [2013]; Collins and Varmus [2015]. The previous study Hong *et al.* [2018] categorized Autism Spectrum Disorder (ASD) into three unique anatomical subtypes using MRI markers that encompass the whole cortex. These categories are: ASD-I, characterized by cortical thickness, increased surface area, and tissue blurring; ASD-II, characterized by cortical thinning and decreased distance; and ASD-III, characterized by increased distance. Both of them are currently concentrating on creating new subcategories of Autism Spectrum Disorder (ASD) by categorizing neuroimaging characteristics into distinct clusters Hong *et al.* [2020b]. The previous study Sadiq *et al.* [2022] reported three main conditions within ASD: Autistic disorder (ATD), Asperger's disorder (APD), and Pervasive developmental disorder not otherwise specified (PDD). ATD is characterized by significant impairments in social interaction, communication, stereotyped behaviors, and interests, typically manifesting before age three and often accompanied by language delays and intellectual disabilities. APD involves challenges in social interaction, nonverbal communication, and repetitive behaviors, but typically without delays in language or cognitive development. PDD is diagnosed when an individual shows some but not all symptoms of ATD or APD or has milder symptoms not fully meeting the criteria for either condition. Neuroimaging studies have shown that individuals with the Asperger's subtype had reduced gray matter volume (GMV) in the prefrontal gyrus and limbic striatal regions relative to healthy individuals Faridi and Khosrowabadi [2017]. A recent research suggested that instead of using categorical categories, ASD can be better understood by considering a continuous variance across individuals. This variation was associated with three distinct elements that affect the functional connectivity of the brain's default mode network, resulting in both hypo and hyper resting state functional connectivity Tang *et al.* [2020].

In the scenarios mentioned above, the previous study focused on identifying heterogeneity within the ASD samples, either using a structured feature set or functional connectivity as a feature set or clinically defined sub-types based on behavioral scores. There is a need for a unique feature set that understands complex brain topology within the ASD samples. To overcome the mentioned challenges, the current study focused on the following research questions:

1. **Identification of Novel Feature-set:** How can we identify a unique feature set for resting-state fMRI data that can improve the identification of heterogeneity within the ASD samples?

2. **Model Architecture:** How can we propose a better explainable and generalized model architecture that could understand the representation of a higher-order correlation feature set (meta-connectivity)?

In this study, we hypothesized that edge connectivity, representing higher-order correlations among the ToM, DMN, CEN, and SN brain networks, could serve as a feature set for the Rough Fuzzy C-means algorithm, potentially acting as a biomarker to identify sub-groups within the ASD samples without the need for ad hoc feature engineering (refer to Figure 7.1). Large-scale collaborative databases, such as the Autism Brain Imaging Data Exchange (ABIDE), exchange terabytes of brain fMRI data compiled from labs worldwide Di Martino *et al.* [2014]. This helps to facilitate a better opportunity to analyze autistic data. In this study, we considered the ABIDE I and ABIDE II datasets, which included 711 autistic samples (after preprocessing) (refer to Table 7.1 and Table 7.2). The key contribution of correct study is as follows:

1. We have included ToM, DMN, CEN, and SN brain networks that are involved in the case of autism Chen *et al.* [2022b]; Supekar *et al.* [2022]. To extract the node-edge connectivity-based feature set as a higher-order correlation feature set, we have calculated meta-connectivity (MC) matrices for each individual in which we extracted time-series signals from networks as mentioned above and calculated dynamic functional connectivity (dFC) stream using a sliding

window approach Arbabyazd *et al.* [2020]. We consider each functional connectivity (FC) link as a dynamic variable that links two regions. Finally, we extracted N(N-1)-times-N(N-1) matrix of covariance between the time-courses of different dynamic variable links. We called such type of inter-link covariance matrix as meta-connectivity

2. We identified ASD sub-groups or heterogeneity within ASD samples using the Rough fuzzy c-means algorithm. It used fuzzy membership as an index to determine which cluster the sample should belong to. Objects were divided into two approximate sets: the lower approximate set of clusters with the highest fuzzy membership degree and the upper approximate set of numerous clusters with lesser degrees.

3. We also analyzed the association between identified sub-groups and behavioral scores.

7.2 METHODS AND MATERIALS

7.2.1 fMRI Dataset and Preprocessing

The study's fMRI anatomical and resting-state functional scans were acquired from the Autism Brain Imaging Data Exchange (ABIDE I and ABIDE II) datasets. The functional scan parameters listed per site were utilized for further preprocessing. ABIDE I contains preprocessed images that were used for the study. The data acquired from ABIDE II was further preprocessed, keeping in line with the process utilized to preprocess the ABIDE I dataset. The data was preprocessed using the DPARSF (Data Processing Assistant for Resting-State fMRI) toolbox, which is compatible with MATLAB. The first 4 time points were deleted to account for the time taken by the subject to settle down. The reference slice for slice timing correction was set as the slice acquired in the middle of the slice order. The Friston-24 model was employed for head motion correction, and timeframes with head motion above 0.2 mm, along with one previous and two following timeframes, were added as regressors. Nuisance regression was done using SPM apriori masks, and global signal regression was not employed as it is a contentious preprocessing measure. A bandpass filter of 0.01 to 0.1 Hz was applied, and the data was normalized using DARTEL. Smoothing was completed with a FWHM kernel size of 6 mm. Scrubbing was done for all timeframes with a head motion above 0.5 mm and one previous and two following timeframes. The 4D preprocessed image thus obtained was further used to extract time series signals for 25 ROIs, and these time series signals obtained for both ABIDE I and ABIDE II datasets were used for all further processes (After preprocessing: total no. of subjects = 1500, comprised of 711 Autistic and 789 neurotypical samples) (Refer to Tables 7.1 and 7.2).

The ABIDE dataset included (fMRI) data from several sites, leading to heterogeneity caused by engineering and biological sampling biases. We used the Combat harmonization approach to resolve this problem, assuming that the error caused by the imaging features may be corrected by adjusting the position (means) and scale (variances) across the batches. Harmonization was conducted using multivariate linear mixed-effects regression using the following formula Maikusa *et al.* [2021]:

$$y(i,j,v) = \alpha(v) + X^T(i,j)\beta(v) + \gamma(i,v) + \delta(i,v)\varepsilon(i,j,v) \tag{7.1}$$

Where α = average functional volume, $\beta(v) = p*1$ vector of coefficients associated with phenotypic parameters, i.e., Age, Gender, ScannerID, X(I,j) = design matrix, p = no. of biological covariates, $\varepsilon(i,j,v)$ = error calculation, $\gamma(i,v)$ and $\delta(i,v)$ = additive and multiplicative of site effect. An

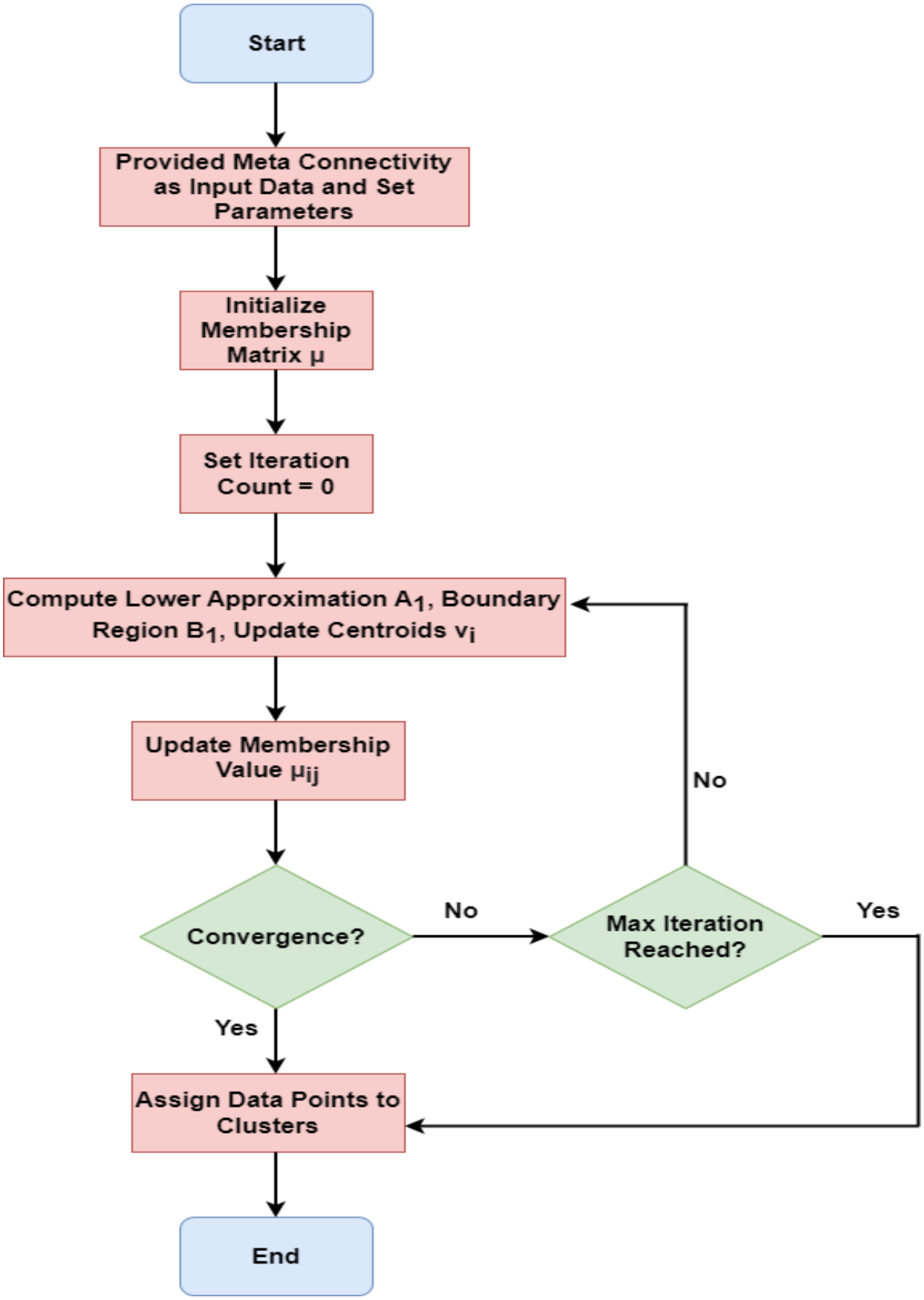

Figure 7.1: Figure illustrates the workflow of proposed Rough Fuzzy C-Means algorithm, in which meta-connectivity between ToM, DMN, CEN, SN brain network for each individual subjected as input.

Sr. No.	Age (In Years)	Samples	IQ	ADI-R Social	ADI-R Verbal
ASD- group	Mean:13.9±6.6, S.D:6.6(range: 6 to 42)	711	105±17	18.8±5.5	15.4±4 .5
TD-Group	Mean:14.95±5.9, S.D:9(range: 6 to 40)	789	112±14	-	-

Table 7.1: Demographic information of ASD and TD groups from ABIDE dataset

Sr. No.	Site	No. of ASD	No. of TD	No. of Male	No. of Female
1	Caltech	19	24	29	14
2	CMU	21	16	31	6
3	KKI	30	35	46	19
4	Leuven	41	44	65	20
5	MaxMun	36	36	64	8
6	NYU	122	190	240	72
7	OHSU	18	21	39	0
8	OLIN	27	21	38	10
9	PITT	39	37	63	13
10	SBL	22	18	40	0
11	SDSU	21	24	34	11
12	Stanford	24	23	35	12
13	Trinity	31	40	71	0
14	UCLA	74	64	113	25
15	UM	86	125	173	38
16	USM	62	32	94	0
17	Yale	38	39	55	22

Table 7.2: Demographic information of participants after pre-processing including ABIDE I and ABIDE II datasets

empirical Bayesian framework is utilized in ComBat harmonization to compute $\gamma^*(i,v)$ and $\delta^*(i,v)$. The final values that ComBat harmonized are represented as Maikusa *et al.* [2021]:

$$y^{combat}_{(i,j,v)} = \frac{y_{(i,j,v)} - \hat{\alpha}_v - X_{(i,j)}\hat{\beta}_v - \gamma^*_{(i,v)}}{\delta^*_{(i,v)}} + \hat{\alpha}_v + X_{(i,j)}\hat{\beta}_v \tag{7.2}$$

This method used empirical Bayes to improve the model estimation Ingalhalikar *et al.* [2021]; Yamashita *et al.* [2019]; Xie *et al.* [2022]; Saponaro *et al.* [2022]. It removed unwanted variation associated with the site and maintained biological association in the dataset.

Networks	**Region-of-Interest**	**MNI-Coordinates (X,Y,Z)**
Salience Network, SN	Anterior Cingulate Cortex (ACC)	0, 22, 35
	Left Anterior Insula (LAINS)	-44, 13, 1
	Right Anterior Insula (RAINS)	47, 14, 0
Default-Mode Network	Medial prefrontal Cortex (MPFC)	0, 48, -18
	Left Lateral Parietal (LLP)	-39, -77, 33
	Right Lateral Parietal (RLP)	47, -67, 29
	Precuneus Cortex	0, -49, 40
Central Executive Network, CEN	Left Lateral Prefrontal Cortex, LLPFC	-43, 33, 28
	Right Lateral Prefrontal Cortex, RLPFC	41, 38, 30
	Left Posterior Parietal Cortex, LPPC	-46, -58, 49
	Right Posterior Parietal Cortex, RPPC	52, -52, 45
ToM Network	Posterior Cingulate Cortex, PCC	0, -52, 22
	Left superior temporal sulcus (LSTS)	-58, -48, 7
	Right superior temporal sulcus (RSTS)	58, -44, 8
	Left Temporoparietal Junction (LTPJ)	-58, -58, 22
	Right Temporoparietal Junction (RTPJ)	58, -52, 20
	Left Inferior Frontal Gyrus (LIFG)	-48, 14, 26
	Right Inferior Frontal Gyrus (RIFG)	48, 14, 26
	Left Angular Gyrus (LANG)	-46, -72, 31
	Right Angular Gyrus (RANG)	48, -69, 31
	Left Cerebellum (LCEREB)	-20, -72, -36
	Right Cerebellum (RCEREB)	20, -72, -36
	Left Inferior Parietal Cortex (LIPC)	-45, -46, 53
	Right Inferior Parietal Cortex (RIPC)	52, -42, 50
	Ventrolateral Prefrontal cortex (VLPFC)	42, 46, 0

Table 7.3: Brain region with MNI coordinates that are associated with Theory-of-Mind, Default-mode-Network, Central Executive, and Salience Networks

7.2.2 Calculation of Meta-Connectivity Feature-set

Calculation of Functional Connectivity and Dynamic Functional Connectivity Stream

We extracted time-series signals from ToM, DMN, CEN, and SN brain networks to calculate meta-connectivity (Refer to Table 7.3). Based on previous literature, we considered 25 Region-of-Interest (ROIs) Baetens *et al.* [2014]; Mazziotta *et al.* [1995]; Kana *et al.* [2015]; Chand and Dhamala [2016]. Using MNI coordinates, we created a spherical binary mask with a 10 mm radius for all selected ROIs and extracted time-series signals. We computed static functional connectivity for each individual using eq (3) Arbabyazd *et al.* [2020]:

$$\overline{\mathrm{FC}}_{ij} = Corr[TS_i(t), TS_j(t)], with(0 \leq t \leq T) \tag{7.3}$$

It calculated pair-wise Pearson correlation matrices that were useful for pursuance trends. To track change in pattern over time, we calculated Dynamic Functional Connectivity (dFC) using a sliding window approach that computed average FC matrices at different time points. According to the previous study Arbabyazd *et al.* [2020], we denoted it as dFC-stream. Using equation 4, we calculated temporal frame of dFC stream for each individual Arbabyazd *et al.* [2020]:

$$FC_{ij}(t_k) = Corr[TS_i(t), TS_j(t)], with(t_k \leq t \leq t_k + W) \tag{7.4}$$

Where t_k indicates the kth temporal frame's start time and w represents fixed window length. We choose w= 30-60s, according to the literature Harlalka *et al.* [2019]. Sliding step $\Delta\tau$ is used to separate frame start time, such that $t_k = t_{k-1} + \Delta\tau = k\Delta\tau$.

Meta-Connectivity Analysis

Each dFC stream was considered as a collection of time series that defines the time dependency of individual FC pairwise couplings. An $M * M$ matrix of correlations between the time-dependent strengths of $M = N(N-1)$ FC links (N^2 pairs of regions, minus the self-loops) was produced by improving the calculation of FC matrix from regional nodes to inter-regional links Arbabyazd *et al.* [2020]. The resultant Meta-Connectivity (MC) matrix expresses the inter-link covariance analogous to how the FC matrix explains the inter-node covariance (Refer to Figure 7.2).

To calculate MC matrices, we used dFC stream and extracted $n = N(N-1)/2$ time-series of pairwise FC coupling, in which input was given as $FC_{ij}(t)$ for all pair of regions i and j with $i < j \leq N$ using equation (5) Arbabyazd *et al.* [2020]:

$$MC_{ij,kl} = Corr[FC_{ij}(t), FC_{kl}(t)] \tag{7.5}$$

These $MC_{ij,kl}$ entries are computed into a matrix format, making it possible to identify the pair of links contributed in each meta-link directly. Consequently, MC matrices had a dimension of $M * M$, where various rows correspond to different directed pairings of regions – that is, both the pair (i,j) and the pair (j,i) were included – and only linkages relating to self-loops – that is, of the type, i.e., (i,i) were eliminated from consideration Arbabyazd *et al.* [2020]. The resulting representation of MC matrices was large since:

$$MC_{ij,kl} = MC_{ji,kl} = MC_{ij,lk} = ... = MC_{lk,ji} \tag{7.6}$$

MC matrices captured higher level correlation between triplets or quadruplets of ROIs, as compared to FC matrices.

7.2.3 Identification of ASD Sub-groups using Rough Fuzzy C-Means Algorithm

We subjected MC matrices of each individual to an unsupervised Rough Fuzzy C-means clustering (RFCM) algorithm to identify sub-groups presented within ASD-samples Mitra *et al.* [2006]; Shi *et al.* [2016]; Raja *et al.* [2020]. The reason behind using RFCM was the ability of fuzzy set to handle overlapping partitions by applying gradual membership degrees to each sample in each cluster and ability of Rough set to deal with uncertainty and incompleteness, in which a centroid and two approximation regions were represented cluster Zhou *et al.* [2011]; Shi *et al.* [2016].

Background

A fuzzy set A in samples X is a class of events defined by continuous membership degree and membership function $\mu_A(x)$ Shi *et al.* [2016]. The value of $\mu_A(x)$ is associated with an interval [0,1] that indicates the degree of membership of x in A. The membership function can be defined using the following equation:

$$\mu_A(x) : X \rightarrow [0,1] \tag{7.7}$$

Where $\mu_A(x)$ represents the weighted coefficient that reflects vagueness in decision-making on belonging of x Shi *et al.* [2016]. Assume $x \in X$, and $\mu_A(x) = 1$ is indication higher degree for x belonging to A, whereas $\mu_A(x) = 0$ representing lower degree.

Rough set is used to reduce vagueness in a given set by applying some approximation. Let's assume U= nonempty set, R = equivalent relation on U, then U can be partitioned into disjoint classes of R:

$$\frac{U}{R} = \{X_1, X_2, ..., X_m\} \tag{7.8}$$

Where X represents the equivalence class of R, $i = 1, 2, ..., m$. Using obtained uncertain information, it is impossible to describe X in real applications. One way to resolve it using lower and upper approximation in a given arbitrary set $X \in 2^U$ using the following equation:

$$R(X) = \bigcup_{X_i \subseteq X} X_i; \;\; R(X) = \bigcup_{X_i \cap X \neq \emptyset} X_i \tag{7.9}$$

Where upper approximation $R(X)$ is defining object having a nonempty intersection with X, whereas lower approximation $R(X)$ represents objects that are a subset of X.

RFCM Architecture

To identify no. of sub-groups within ASD-samples, we implemented RFCM in which cluster was represented using fuzzy lower approximation and fuzzy upper approximation, and centroid was updated using the following way:

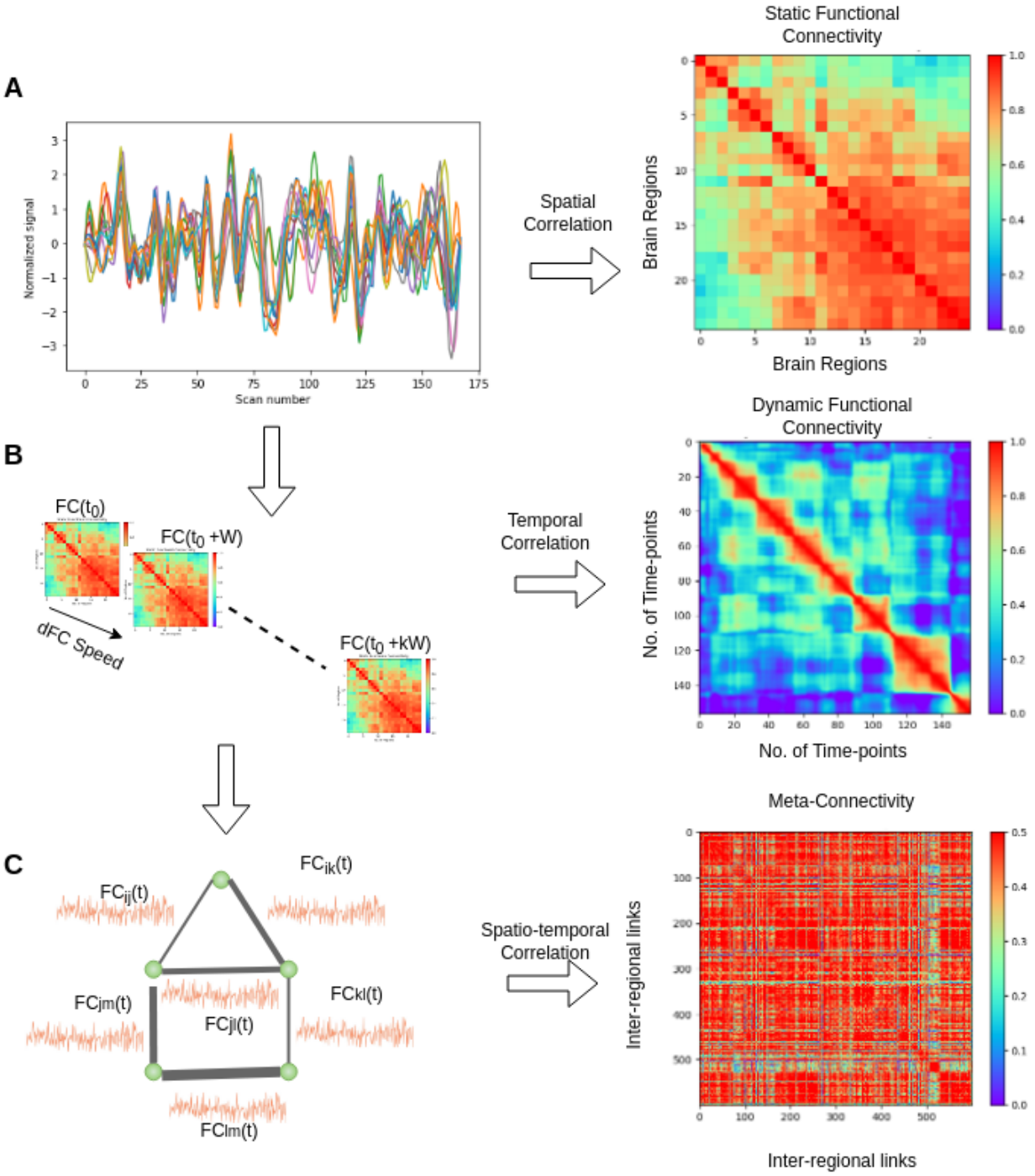

Figure 7.2: Calculation of Static, Dynamic functional connectivity, and Meta-connectivity: In the traditional approach **(A)**, neural activity is averaged over time to create a static Functional Connectivity (FC) matrix. We calculated a dynamic FC stream using shorter sliding windows **(B)**, leading to a dynamic FC matrix. In **(C)**, we considered each FC link as a dynamic variable, giving rise to a Meta-Connectivity (MC) matrix by capturing covariances between these dynamic links.

$$v_i = \begin{cases} w * A_1 + w^{'*} B_1 & \text{if} \quad R_1(\beta_i) \neq \emptyset, R_{1b}(\beta_i) \neq \emptyset \\ A_1 & \text{if} \quad R_1(\beta_i) \neq \emptyset, R_{1b}(\beta_i) \neq \emptyset \\ B_1 & \text{if} \quad R_1(\beta_i) \neq \emptyset, R_{1b}(\beta_i) \neq \emptyset \end{cases} \tag{7.10}$$

$$A_1 = \frac{1}{n_i} \sum_{X_j \in \underline{R}_1(\beta_i)} (\mu_{ij})^m X_j;\ n_i = \sum_{X_j \in \underline{R}_1(\beta_i)} (\mu_{ij})^m$$

$$B_1 = \frac{1}{n_j} \sum R_{1b}(\beta_i) (\mu_{ij})^m X_j;\ n_j = \sum_{X_{j1b}(\beta_i)} (\mu_{ij})^m$$

Here, c represents the number of clusters, β_i represents cluster i for $i = 2, \ldots, c$, $\underline{R}_1(\beta_i)$ is the lower approximation, $R_{1b}(\beta_i)$ is the boundary region, μ_{ij} is the fuzzy membership value, X_j is the data point, w is the weight for the lower approximation, w^* is the weight for the boundary region, and m is the fuzzy coefficient. If neither of two approximations was empty, then both the lower approximation and the boundary region were used to compute a new centroid. The relative significance of two approximations is determined by weighted parameters (w and w'). Since the lower approximation contributes more to the computation of a new centroid than the boundary region. $0 < w' < w < 1$ and $w + w' = 1$. Fuzzy membership degrees were used to identify which approximations an item belonged to. With the addition of fuzzy sets, the rough set-based clustering method could handle overlapping situations more effectively. The fuzzy membership function has the advantage of simultaneously making an item belong to all clusters with varying degrees of ownership, improving capability in dealing with uncertainty caused by overlapping boundaries.

Algorithm 2 Rough Fuzzy C-Means (RFCM) Clustering Algorithm

- Meta connectivity data X (size $M \times M$)
- Number of clusters c
- Fuzzy coefficient m ($m > 1$)
- Weighted parameters w and w' ($0 < w' < w < 1$ and $w + w' = 1$)
- Maximum iterations *max_iter*
- Convergence tolerance *tol*

Cluster assignments for each subject, cluster centroids

Initialize fuzzy membership matrix μ randomly such that $\sum_{i=1}^{c} \mu_{ij} = 1$ for all j Initialize iteration counter to 0 Initialize convergence flag to **false**

iteration counter < max_iter convergence flag is **false** Increment iteration counter each cluster i Compute lower approximation A_1:

$$A_1 = \frac{1}{n_i} \sum_{X_j \in \underline{R}_1(\beta_i)} (\mu_{ij})^m X_j$$

Compute boundary region B_1:

$$B_1 = \frac{1}{n_j} \sum_{X_j \in R_{1b}(\beta_i)} (\mu_{ij})^m X_j$$

Update centroid v_i:

$$v_i = \begin{cases} w \cdot A_1 + w' \cdot B_1 & \text{if} \quad \underline{R}_1(\beta_i) \neq \emptyset \text{ and } R_{1b}(\beta_i) \neq \emptyset \\ A_1 & \text{if} \quad \underline{R}_1(\beta_i) \neq \emptyset \text{ and } R_{1b}(\beta_i) = \emptyset \\ B_1 & \text{if} \quad \underline{R}_1(\beta_i) = \emptyset \text{ and } R_{1b}(\beta_i) \neq \emptyset \end{cases}$$

Store old membership matrix $\mu_{\text{old}} = \mu$

each data point X_j each cluster i Update membership value μ_{ij}:

$$\mu_{ij} = \left(\sum_{k=1}^{c} \left(\frac{d(X_j, v_i)}{d(X_j, v_k)} \right)^{\frac{2}{m-1}} \right)^{-1}$$

Check for convergence: $\|\mu - \mu_{\text{old}}\| < tol$ Set convergence flag to **true**

Assign each data point X_j to the cluster with the highest membership value μ_{ij} Cluster assignments, cluster centroids

7.3 RESULTS

7.3.1 Computation of Meta-Connectivity Matrices

In the current study, we extracted time-series signals from ToM, DMN, CEN, and SN brain networks and calculated functional connectivity matrices for each individual. The FC matrices have a unique symmetry, and despite their large N × N size, they only have $L = N(N-1)/2$ unique entries, representing the lower triangular part of the matrix. As the output, FC matrix appears as a $N * N$ matrix. We also calculated dFC stream, that output in a 3D tensor sized $N * N * F$, where F was

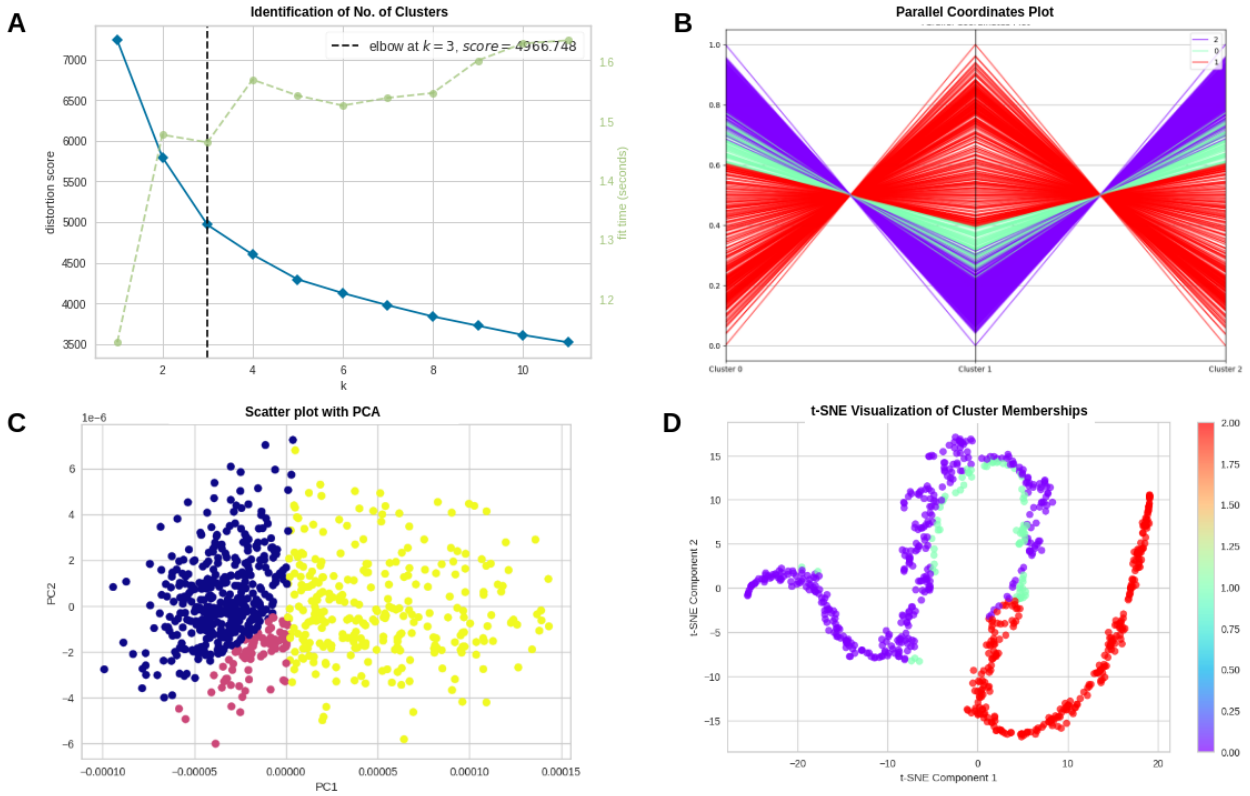

Figure 7.3: A) is showing identification of no. of clusters using Elbow approach. **B), C), and D)** are showing visualization of clusters using different approaches.

no. of frames. We chose window sizes 20 and 30 as per the literature Harlalka *et al.* [2019]. However, each FC matrix was shown in the 'vector' format that outputs a vector of size $L * 1$, and dFC stream yielded a 2D matrix sized $L * F$, with each frame in vector format. The 'vector' format is advantageous because it produces more memory-efficient output, which is essential for large datasets. It also naturally represents FCs as points in a vector space, making it easier to create dimensionally reduced representations of the dFC stream using methods like t-stochastic neighborhood embedding. After that, we calculated MC matrices for each individual of size $M * M$ (refer to Figure 7.2). To validate the results, we performed multiple one-sample t-tests with a p-value$<$0.01 and applied FDR correction.

7.3.2 Identification of ASD Sub-Groups using RFCM

We analyzed a total of 711 ASD samples, including data from ABIDE I and ABIDE II, to uncover hidden sub-groups within the ASD samples, aiming to better understand the heterogeneity within it. We applied the Rough Fuzzy C-Means algorithm in which MC matrices were subjected as input. Our analysis revealed three distinct sub-groups: the first subgroup included 361 samples, the second subgroup comprised 76 samples, and the third subgroup encompassed 274 samples.

1. **Clustering Methodology:** In this study, we applied RFCM to partition ASD individuals into $n = 3$ distinct clusters based on their meta connectivity matrices extracted from the mentioned brain networks. The no. of clusters was identified using the elbow approach shown in Figure 7.3A. The algorithm was configured with a fuzziness coefficient of $m = 2$, an error

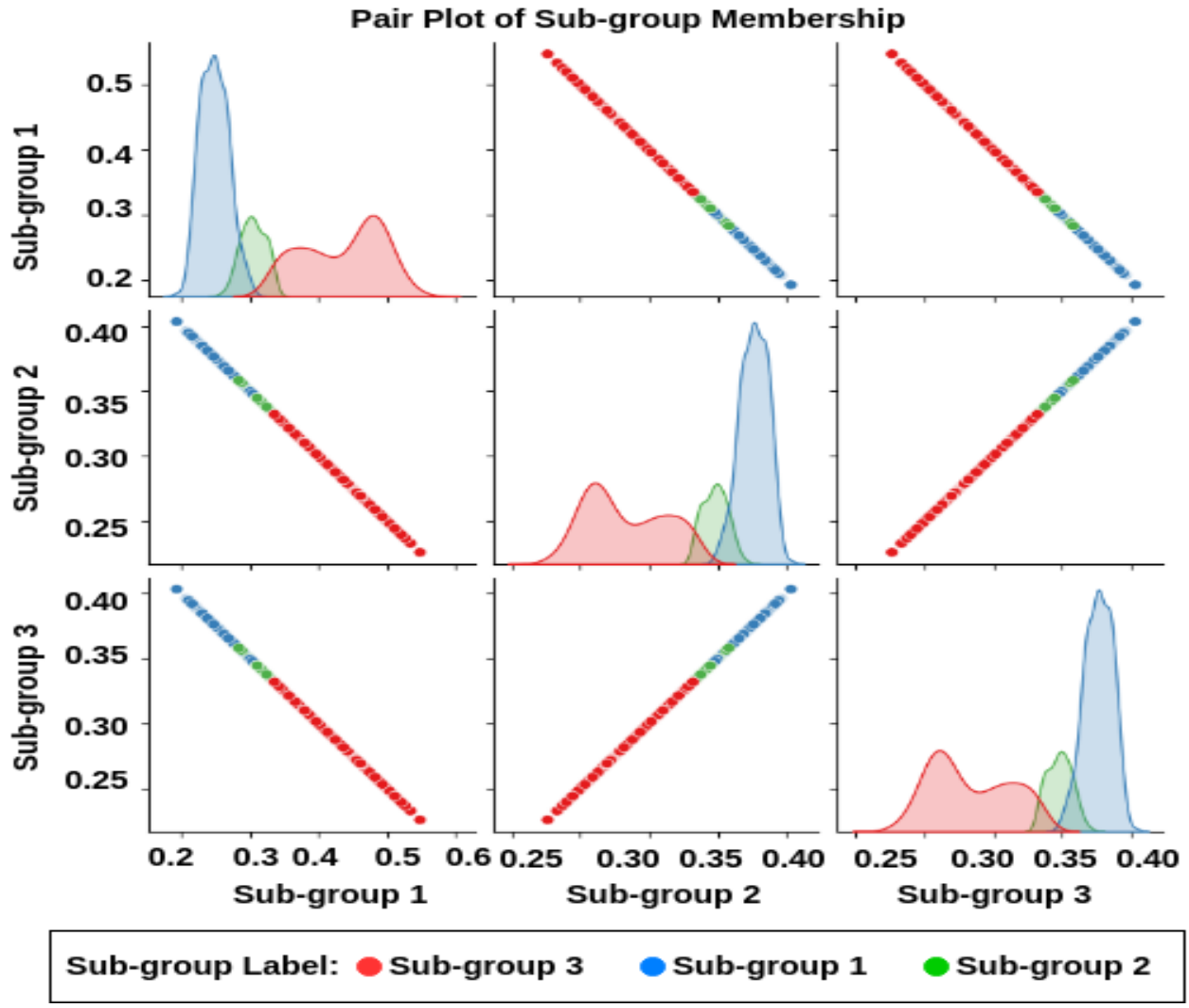

Figure 7.4: Illustrating the association between all three clusters,i.e., how one sample in a cluster is also associated with other clusters with a membership confidence value. The result suggested that sub-group 2 was highly associated with sub-groups 1 and sub-groups 3, whereas sub-group 1 and sub-group 3 were slightly associated with each other.

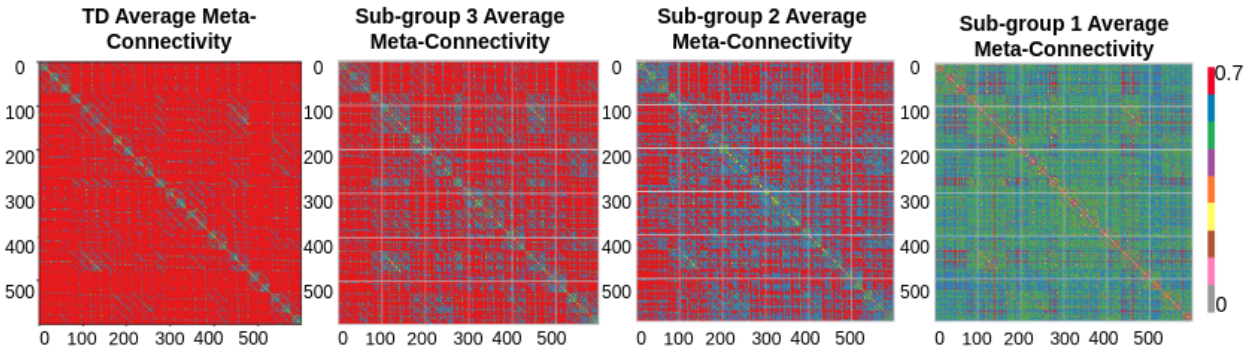

Figure 7.5: Average Meta-connectivity matrices for Neuro-typicals, sub-group 1, sub-group 2, and sub-groups 3 that show that sub-group 3 is more close to neuro-typicals and less associated with ASD symptoms. Whereas sub-group 1 is more associated with ASD symptoms.

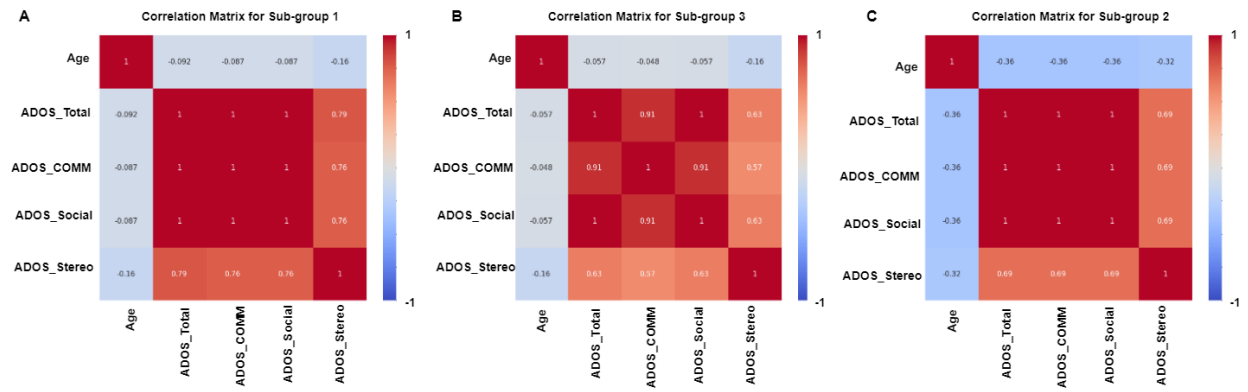

Figure 7.6: Figure illustrates the correlation matrix among behavioral scores for each subgroup. We found lower correlation for 3rd sub-group and higher correlation for 1st and 3rd sub-groups.

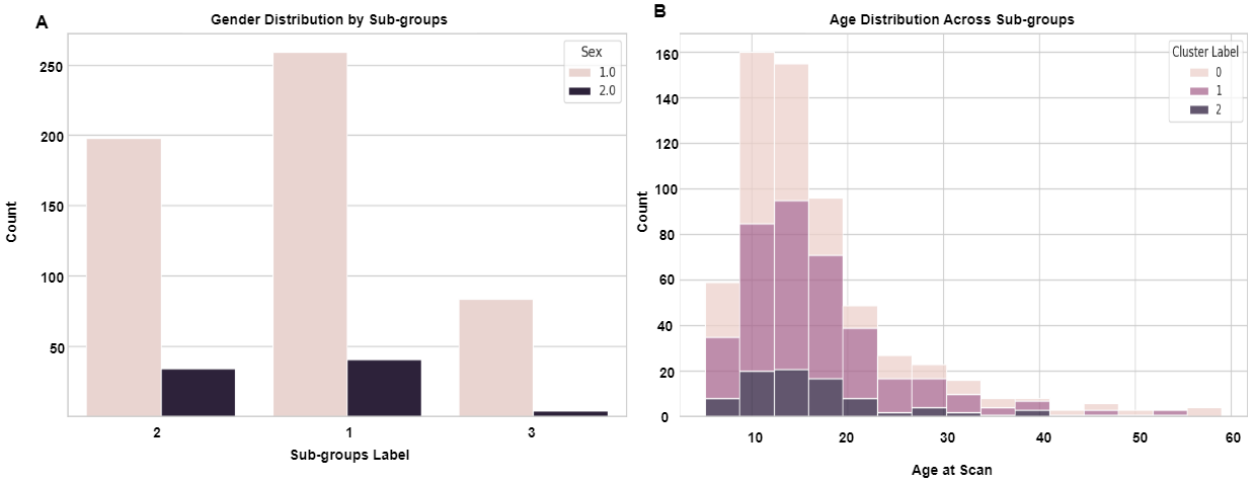

Figure 7.7: Figure illustrates gender and age distribution across clusters. In part B, we found that most of the samples belonged to the early developmental age range, and after 25 years of age, there were very few samples. In part A, we found that males were more included in the case of autism.

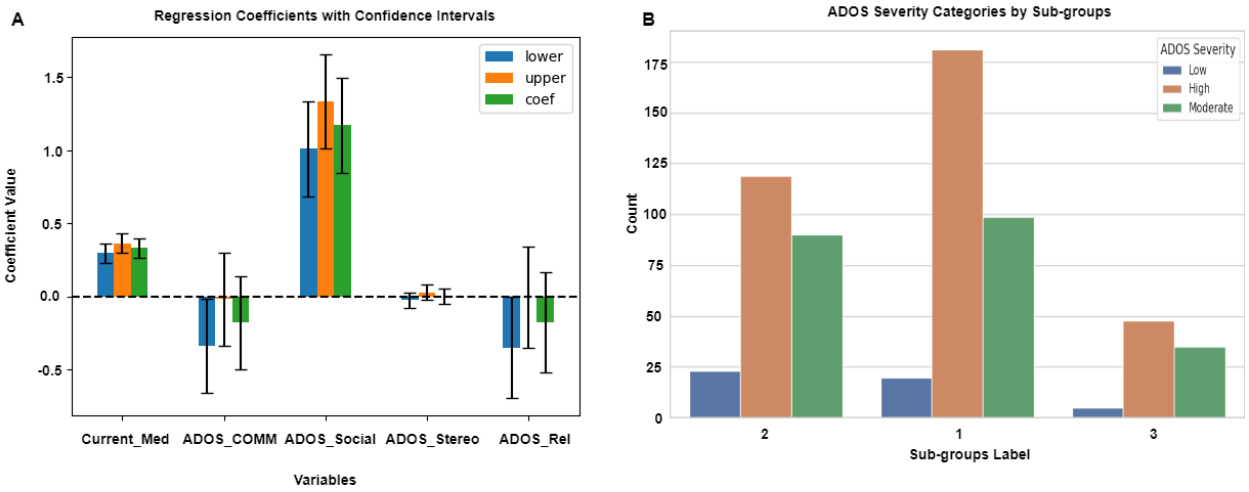

Figure 7.8: Figure illustrates regression coefficient with confidence interval and ADOS severity in each subgroup. We found more severity in 1st sub-group.

threshold of 0.005, and a maximum of 1000 iterations to ensure convergence (refer to Figure 7.3).

2. **Cluster Characteristics:** The clusters were characterized by their centroids, representing the average connectivity patterns within each cluster. The centroids were computed using fuzzy membership values (μ_{ij}), where higher values indicate stronger membership to a particular cluster. This approach allowed us to capture each sample's gradual degrees of belongingness to multiple clusters, reflecting the complex nature of ASD sub-groups.

3. **Lower and Upper Approximation:** In RFCM, each cluster is defined by its lower approximation ($\underline{R}(X)$) (We define the lower approximation threshold at $\mu_{\text{lower}} = 0.3$, indicating that samples with membership degrees $\mu_{ij} \geq 0.3$ are considered definitively belonging to cluster X_i.) and upper approximation ($\overline{R}(X)$) (The upper approximation threshold is set at $\mu_{\text{upper}} = 0.6$, including samples with membership degrees $0.6 \leq \mu_{ij} < 0.3$ in the cluster X_i. These samples exhibit some similarity with the cluster but do not meet the higher certainty threshold of the lower approximation) (refer to Figure 7.4). The lower approximation includes objects that belong to the cluster, while the upper approximation encompasses objects that may or may not belong, considering the uncertainty in membership assignment. This dual approximation framework provides a nuanced understanding of cluster boundaries and membership uncertainties within the ASD samples.

4. **Validation of Results:** To assess the validity of the identified clusters, we calculated several cluster validity indices: a) Silhouette Score: 0.65, b) Davies-Bouldin Index (DBI): 0.42, c) Dunn Index: 2.1. These indices measure the cohesion within clusters and separation between clusters. A higher silhouette score and Dunn Index, along with a lower DBI, suggest well-defined and distinct clusters in our analysis.

5. **Interpretation of Cluster Patterns:** To visually explore the identified clusters, Figure 7.4 presents a pair plot of cluster memberships. Each scatter plot depicts the relationship between membership values across pairs of clusters, with KDE plots along the diagonal showing the distribution of membership values within each cluster. The colors denote cluster labels, illustrating how individuals are distributed across the identified sub-groups based on

their connectivity profiles. Interestingly, we found that the 1st sub-group contained unique samples with lower associations with the other sub-groups. In contrast, the 2nd sub-group exhibited a high degree of association with the 1st sub-group and moderate association with 3rd sub-group (refer to Figure 7.4). Additionally, some samples showed associations with all three sub-groups. Lastly, the 3rd sub-group was slightly associated with the 1st sub-group. In the overall analysis, the 1st and 3rd sub-groups showed unique samples with less association.

6. **Analysis of Each Cluster:** To gain further insight into our results, we calculated the mean connectivity matrix for each subgroup. Notably, the 3rd subgroup displayed a notably high correlation (0.7, $p < 0.02$) between its edges, while the 2nd subgroup yielded a similar result, with a high correlation (0.6, $p < 0.03$). Conversely, the 1st subgroup demonstrated a moderate correlation (0.5, $p < 0.01$) between its edges. For validation purposes, we conducted multiple one-sample t-tests on each connectivity with a significance level set at $p < 0.01$ and applied FDR correction. We also calculated the mean MC matrix for the typical developing (TD) group for comparison purposes and observed a high correlation between edges in the TD group. We found that the 3rd sub-group was more associated with the TD group, whereas the 2nd sub-group was moderately associated, and the 1st sub-group was less associated with the TD group.

7.3.3 Association between Identified Clusters and Behavioral Scores

In this study, we successfully identified three distinct clusters within the ASD samples using RFCM, demonstrating the potential of this method for subgroup identification in ASD. Correlation analysis revealed insightful relationships, with current medication status (CURRENT-MED-STATUS) showing small negative correlations with communication (-0.100), social interaction (-0.091), stereotyped behaviors (-0.105), and research reliability (-0.102). Regression analysis provided robust findings, indicating that CURRENT-MED-STATUS ($coef = 0.3322, p < 0.001$) and ADOS-SOCIAL ($coef = 1.1727, p < 0.001$) are significant positive predictors of overall ADOS scores, while ADOS-COMM ($coef = -0.1776, p = 0.035$) and ADOS-RSRCH-RELIABLE ($coef = -0.1744, p = 0.048$) are significant negative predictors (refer to Figure 7.6). The high R-squared value (0.080) suggests that our model explains the variability in ADOS scores well, though it also indicates potential fitting. Diagnostic tests, including the Omnibus test ($6.952, p = 0.031$), Durbin-Watson statistic (1.441), Jarque-Bera test ($3.824, p = 0.048$), and condition number ($3.77e+03$), were used to assess model validity, highlighting areas for further refinement (refer to Figure 7.8).

In our study, we explored the association between clinically defined ASD sub-types and sub-types identified using meta-connectivity between brain regions. We found that: a) 1st sub-group, characterized by severe ASD, included 199 individuals with ATD, 66 with APD, and 14 with PDD. b) 2nd sub-group showing moderate ASD severity included 155 individuals with ATD, 51 with APD, and 20 with PDD. c) 3rd sub-group, more closely associated with typical development, included 37 individuals with ATD, 13 with APD, and 156 with PDD.

7.4 DISCUSSION

The primary challenge in analyzing ASD samples lies in identifying hidden sub-groups or heterogeneity among these samples Reiter *et al.* [2021]; Xu *et al.* [2021]; Akter *et al.* [2021]; Wang *et al.* [2022a]. To address this, our study aimed to identify distinct sub-groups within the autistic samples using meta-connectivity matrices as feature set and the Rough Fuzzy C-Means (RFCM) clustering algorithm. RFCM combines fuzzy set theory, which handles overlapping memberships,

and rough set theory, which manages uncertainty, providing a robust framework for characterizing heterogeneous sub-groups in ASD. This approach assigns a membership value to each sample, indicating its association with different clusters or sub-groups.

7.4.1 Analysis of Each Cluster

The traditional clustering approach like k-means clustering algorithm directly dividing data into clusters, but RFCM shows association between cluster on basis of shared features. Our analysis revealed three sub-groups within the ASD samples. 2nd sub-group showed a high association with both 1st sub-group and 3rd sub-group, whereas 1st sub-group and 3rd sub-group had less association with each other. 1st and 3rd sub-groups contained unique samples, in which samples which were coming under 1st sub-group were more severe to ASD severity, whereas samples that were coming under 3rd sub-group were slightly associated with ASD severity. To get insight the results, we calculated average meta-connectivity (MC) matrices for both TD and ASD groups (refer to Figure 7.5). The TD group exhibited hyper-connectivity between edges, while the ASD group showed hypo-connectivity. For each ASD sub-group, we also calculated an average meta-connectivity matrix. For 3rd sub-group, connectivity was similar to the TD group except for a few edges within the DMN and ToM networks, indicating a more significant association with ASD symptom severity. 2nd sub-group, with fewer samples, showed moderate connectivity, particularly between the DMN, Central Executive Network, and ToM networks, which exhibited lower connectivity. 1st sub-group displayed hypo-connectivity, especially between the DMN and CEN. Additionally, edges associated with the salience network were highly correlated with other networks across all three clusters. This finding aligns with previous reports of hypercorrelation in the salience network among autistics Chen *et al.* [2017]; Margolis *et al.* [2019].

7.4.2 Association between Identified Clusters and Behavioral Scores

To check the association between behavioral scores and sub-groups identified by the proposed approach, we analyzed the association between age, gender, and ADOS scores (i.e., ADOS-Social, ADOS-Total, ADOS-Communication, and ADOS-Stereo) with each cluster. We found severe ADOS scores in the first sub-group, moderate scores in the second sub-group, and lower scores in the third sub-group. The first sub-group had a higher number of males (over 250 samples) compared to females (approximately 40) and included the majority of samples aged 6-25 years (refer to Figure 7.7). Our analysis indicated that the first sub-group, identified using MC matrices of the mentioned brain networks, was the most severe ASD group, closely associated with higher symptom severity scores. Notably, many samples in the first sub-group underwent no medical treatment. In contrast, the second sub-group was moderately associated with ASD symptoms, and the third sub-group, which was more similar to the TD group, included samples with lower ASD severity.

We also identified an association between clinically identified ASD sub-types Sadiq *et al.* [2022] and sub-groups identified by the RFCM approach. The clinically defined sub-groups were identified mostly on the basis of behavioral scores. There was no role in the analysis of anatomical or functional MRI images. Our findings revealed that a majority of individuals with ATD were grouped into sub-groups 1 and 2, which showed stronger associations with ASD severity. Conversely, most individuals with PDD fell into sub-group 3, which exhibited closer associations with typically developing (TD) groups, indicating milder ASD symptoms. Figure 4 illustrates distinct sample distributions with minimal overlap between sub-group 1 and sub-group 3. Our findings suggest that while clinically defined sub-groups share some overlap in meta-connectivity within specific brain networks, differences in behavioral scores between ATD and APD do not necessarily correspond

to differences in functional brain development. Each group showed overlapping patterns in meta-connectivity. There is no framework in the literature, to the best of our knowledge, that shows an association between clinically identified sub-groups and sub-groups identified using higher-order correlation like meta-connectivity.

The RFCM algorithm effectively identified and validated three meaningful sub-groups within the autistic samples based on functional connectivity patterns. The combination of cluster size distribution, validity indices, visual representations, and external validation confirms the robustness and reliability of our clustering approach. These findings provide valuable insights into the heterogeneity of ASD and have implications for personalized diagnostics and interventions. However, the application of RFCM offers a novel approach by integrating fuzzy and rough set theories to address the challenges of overlapping and uncertain cluster boundaries in ASD research. This methodological novelty enhances our ability to delineate subtle variations in connectivity patterns, providing deeper insights into ASD sub-groups than traditional clustering approaches. our study demonstrates the potential of RFCM in unraveling the complex heterogeneity within the ASD samples. By identifying distinct sub-groups and their connectivity patterns, we can better understand the underlying mechanisms of ASD and tailor more effective, personalized interventions. These insights could lead to improved diagnostic accuracy and therapeutic strategies, ultimately enhancing the quality of life for individuals with ASD. As we continue exploring this approach, future research should focus on more extensive and diverse samples to validate and extend our findings, paving the way for more nuanced and inclusive ASD studies.

7.5 SUMMARY

Our study utilized meta-connectivity matrices and the Rough Fuzzy C-Means (RFCM) clustering algorithm to uncover distinct sub-groups within the ASD samples. RFCM, integrating fuzzy and rough set theories, proved effective in handling overlapping memberships and uncertainty, providing robust characterization of ASD heterogeneity. Our findings revealed three sub-groups differing in their associations with ASD severity. 1st sub-group exhibited severe symptoms and hypo-connectivity, particularly within the DMN and CEN brain networks. 2nd sub-group showed moderate connectivity, notably between DMN, CEN, and ToM networks. 3rd sub-group resembled TD connectivity patterns with minimal ASD-related alterations. Sub-group characterization also aligned with ADOS scores, where 1st sub-group had severe scores, 2nd sub-group moderate, and 3rd sub-group had lower scores. Males predominated in 1st sub-group, aged 6-25 years, with minimal medical intervention, emphasizing its severity. We observed that ATD predominantly clustered in 1st and 2nd sub-groups, correlating with higher ASD severity, while PDD samples clustered in 3rd sub-group, indicative of milder symptoms resembling TD traits. RFCM's integration of higher-order correlation analysis highlighted subtle connectivity variations, surpassing traditional clustering methods in capturing ASD sub-group nuances. Our approach underscores the potential for personalized diagnostics and interventions tailored to ASD sub-groups. These findings enhance our understanding of ASD heterogeneity and advocate for personalized treatment strategies. Future research should expand sample diversity to validate and extend these insights, paving the way for more comprehensive ASD studies and improved clinical outcomes.