Segment-Then-Connect: Change Point Dynamic Connectivity for Early MCI Detection

Martin Ondrus
Neuroscience and Mental Health Institute
University of Alberta

Ivor Cribben
Alberta School of Business
University of Alberta

Abstract

The most widely used inputs in classification models for resting-state functional magnetic resonance imaging (rs-fMRI) data are estimates of static-based functional connectivity (SFC) and sliding window dynamic functional connectivity (swDFC). Although these methods are computationally convenient, the resulting representations are highly simplified portrayals of a deeply integrated and dynamic process. Change point dynamic functional connectivity (cpDFC) methods offer an alternative to swDFC approaches with many advantages. In this study, we consider a classification task between controls and patients with eMCI using rs-fMRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) studies, ADNI2 and ADNIGO. Our results indicate that the DFC methods are generally superior to the SFC methods when used as inputs into the classification model. Most importantly, we find that cpDFC is generally superior to swDFC. We discuss how cpDFC methods offer greater parsimony of network features and ease of interpretability. Our empirical results indicate that functional brain network representations are dynamic, multiscale, and subject-specific, underscoring the need for a learning paradigm tailored to these properties.

1 Introduction

Resting-state functional magnetic resonance imaging (rs-fMRI) enables the investigation of brain connectivity without task-related variability, making it useful for detecting functional differences in neurological conditions [Biswal et al., 1995]. Functional connectivity (FC) quantifies statistical dependencies between regions of interest (ROIs) and is often represented as a graph [Muldoon and Bassett, 2016, Bassett and Bullmore, 2017]. While static FC (SFC) assumes constant connectivity, dynamic FC (DFC) captures time-varying interactions that better reflect neural processes [Hutchison et al., 2013]. The sliding-window approach (swDFC) is by far the most common method for DFC [Allen et al., 2014], but suffers from arbitrary window-size selection, redundancy from overlapping windows, and assumptions about uniform change rates [Zalesky and Breakspear, 2015, Leonardi and Van De Ville, 2015]. Furthermore, rs-fMRI time series are known to exhibit discrete state-switching rather than smooth, continuous drifts as assumed in swDFC [Allen et al., 2014].

We propose a segmentation-first approach that aligns more closely with the true generative process. Change point dynamic functional connectivity (cpDFC), which segments fMRI time series into non-overlapping stationary intervals in a data-driven manner. This approach reduces redundancy, captures subject-specific dynamics, and can produce more discriminative features for disease classification. We evaluated cpDFC against SFC and swDFC in distinguishing early mild cognitive impairment (eMCI) from healthy controls using Alzheimer's Disease Neuroimaging Initiative (ADNI) rs-fMRI data, and validated on an independent MCI dataset [Mascali et al., 2015]. Our work tackles key challenges in neuroimaging, namely, modeling complex, high-dimensional time series data (rs-fMRI), capturing distributional shifts, and generating interpretable, reliable representations for clinical

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

classification. Unlike fixed-window approaches that impose uniform temporal segmentation, cpDFC adaptively discovers subject-specific, stationary network regimes, parsing meso-scale dynamics that may correspond to evolving cognitive states or neural processes. Using change points, our approach instead extracts structured, meaningful segments that are more amenable to clinical deployment.

Our contributions are as follows: (1) the first application of change point detection to classify neurodegenerative diseases from rs-fMRI, (2) a systematic comparison of cpDFC, swDFC, and SFC, including step-size effects of swDFC, (3) the demonstration that cpDFC produces superior accuracy and interpretability, (4) evidence that combining multiple FC methods in an ensemble improves performance for eMCI classification and provides richer representations than any one method.

2 Methods

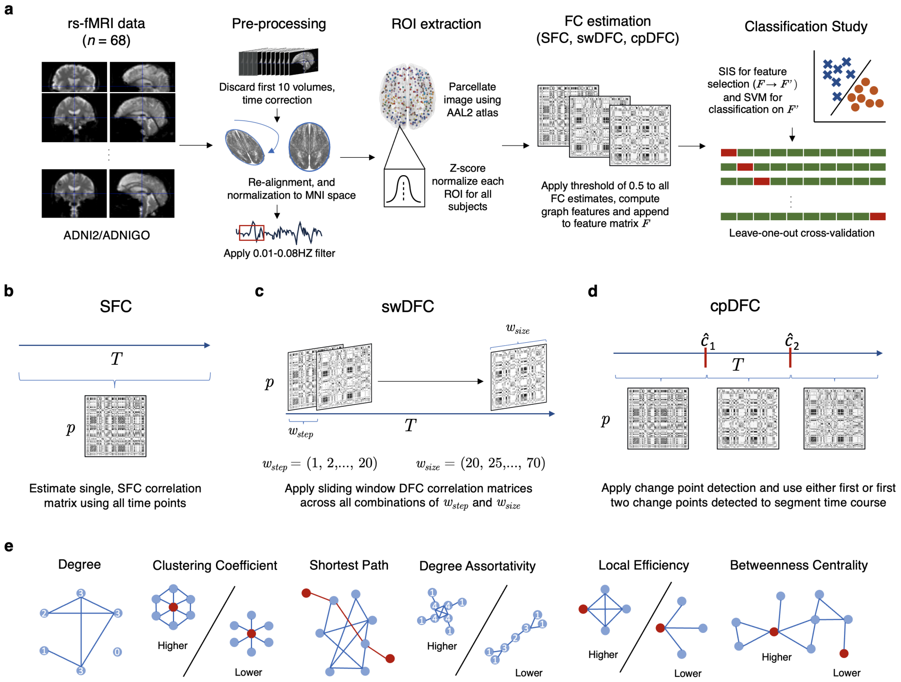


Figure 1: A schematic of CN/eMCI classification task starting from raw rs-fMRI data to eMCI/control classification for the ADNI rs-fMRI dataset. Panel (a) is the overall pipeline, while panels (b), (c), and (d) show the different FC methodologies (SFC, swDFC, cpDFC, respectively). Panel, (e), shows example graphs expressing the topological features used in the classification study.

We provide notation in the Supplementary Materials A. In ADNI's rs-fMRI dataset, we analyzed 33 subjects with early mild cognitive impairment (eMCI; mean age 72.3, 15M/18F) and 35 healthy controls (mean age 74.6, 14M/21F). Data were preprocessed in SPM [SPM, 2023] following a standard fMRI pipeline, including removal of initial volumes, motion correction, spatial normalization to MNI space, nuisance regression, spatial smoothing, detrending, and band-pass filtering (0.01–0.08 Hz). The brain was parcellated into 120 regions using the AAL2 atlas [Tzourio-Mazoyer et al., 2002], and Z-score normalized. The final data set of each subject was a $T = 130$ by $p = 120$ matrix (time points $\times$ ROIs). Full details are provided in the Supplementary Materials C.

For each stationary segment $\boldsymbol{S}$, FC between ROIs i and j is estimated using Pearson's correlation ρ_{ij}. An edge is kept if $|\rho_{ij}| > 0.5$, which yields graphs of strong statistical dependence [Adamovich et al., 2022]. For each subject, SFC assumes one stationary segment (entire scan), whereas DFC yields multiple segments. We estimate DFC under the assumption of unknown distributional shifts and rate of change, but with stationary segments existing in $\boldsymbol{X}$. Two approaches were used:

Sliding windows (swDFC): a fixed-length window (w) is shifted by step size (s), with FC computed in each (Algorithm 1 in Supplementary Materials; schematic in Figure 1c).

Change points (cpDFC): we use FaBiSearch [Ondrus et al., 2025], which detects multiple change points in multivariate time series based on network structure via non-negative matrix factorization [Lee and Seung, 1999]. The hyperparameters ($\delta, n_{\text{run}}, n_{\text{reps}}$) are established using recommendations from Ondrus and Cribben [2024]. Change points are ordered by p-value, and the first $k = 1$ or $k = 2$ are used to segment the time series before FC estimation. We also test NCPD [Cribben and Yu, 2017] and CRMT [Ryan and Killick, 2023] with comparable settings.

For each segment's graph, we compute standard node- and network-level metrics, degree, clustering coefficient, shortest path length, degree assortativity, local efficiency, and betweenness centrality, as candidate features [Rubinov and Sporns, 2010]. Formal definitions are provided in the Supplementary Materials G.

Labels are defined as $c = 0$ (control) and $c = 1$ (eMCI). The features of all segments are concatenated, and sure independence screening (SIS; Fan and Lv, 2008) selects a reduced set $\boldsymbol{F}'$. We train a linear SVM [Boser et al., 1992], chosen for robustness after testing RBF-SVM, logistic regression, and decision trees. We compare SFC, swDFC, and cpDFC in separate experiments. For SFC, all $T = 130$ time points are used. For swDFC, we test $w \in [10, 70]$ (step 5) and $s \in \{1, 2, 3, 5, 8, 10, 15, 20\}$. For cpDFC, we evaluate FBS_cpDFC1 and FBS_cpDFC2. SIS and SVM are applied within each fold of leave-one-out cross-validation (68 subjects total). The complete pipeline is shown in Figure 1a.

3 Results

We first evaluate SFC, swDFC, and cpDFC separately, and then combine the best-performing models in an ensemble. Figure 2 summarizes the results. For swDFC, we report only window/step size combinations achieving accuracy $\geq 63.57\%$ (one standard error above the null accuracy of 51.47%). Among all methods, FBS_cpDFC2 achieves the highest overall performance in two of four metrics: accuracy 77.94%, F1 80.00%, sensitivity 90.91%, and specificity 65.71%. The only metric in which FBS_cpDFC2 is significantly different from swDFC is sensitivity, but the two swDFC variants with higher sensitivity (60w_15s and 70w_10s) perform notably worse in other metrics. Compared to the best swDFC configuration (15w_3s: accuracy 72.06%, F1 74.67%), FBS_cpDFC2 performs slightly better (n.s., $p = 0.250$) but is significantly superior compared to the full swDFC set ($\mu = 0.5852$, $\sigma = 0.0483$; all metrics $p < 1.4 \times 10^{-5}$). cpDFC1 and SFC both underperform, with SFC near null accuracy. Across all (w, s) combinations of swDFC (Supplementary Materials B.2), no performance trend emerges and small parameter changes can cause large accuracy drops.

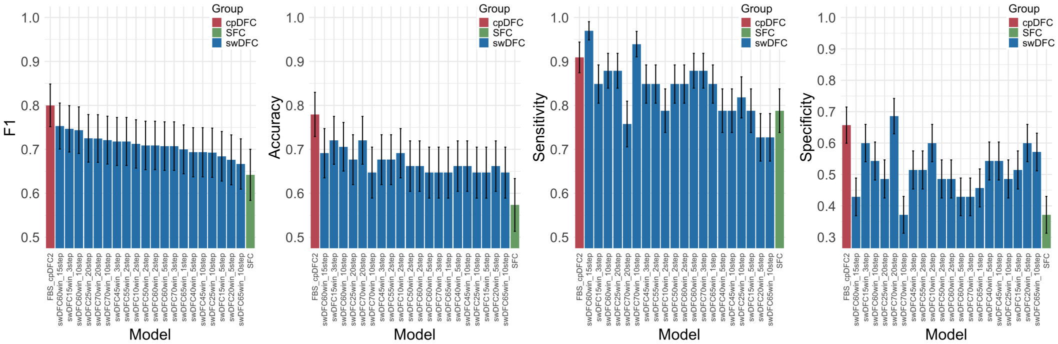


Figure 2: F1, accuracy, sensitivity, and specificity results from the classification study of CN subjects and subjects with eMCI in the ADNI rs-fMRI dataset. SFC, cpDFC, and swDFC correspond to static, change point, and window based dynamic functional connectivity, respectively.

In general, the difference between FBS_cpDFC2 and the best swDFC combination ($w = 60$, $s = 15$) can be summarized as such: F1 score improved by 6.26%, accuracy by 12.76%, and specificity by 53.31%, while the sensitivity was lower by 6.25%. swDFC exhibited substantial variability, depending on the choice of window and step sizes, with accuracy in the range of $22.06\% - 72.06\%$, F1 score $19.35\% - 75.29\%$, sensitivity $18.18\% - 96.97\%$, and specificity $14.29\% - 68.57\%$.

Figure 3 shows the performance of the ensemble when combining predictions from the top-2 to top-10 classifiers (ranked by F1 score). The best single model is FBS_cpDFC2. Combining additional

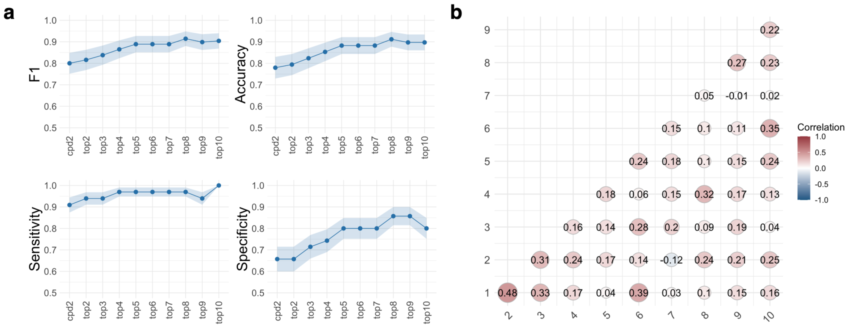


Figure 3: The ensemble model results from the classification study of CN subjects and subjects with eMCI in the ADNI dataset. Panel (a) shows the ensemble results, which combines the predictions from the top-2 to the top-10 classifier models as determined by F1 score. FBS_cpDFC2 is the best stand-alone classifier. Shaded regions indicate ± the standard error of proportion calculated. (b) shows the correlation between predicted probabilities of eMCI across the top-10 classifiers.

models does yield an improvement, likely due to minimal overlap in predictive information and differences in assumed rates of change. A correlation of predicted probabilities (Figure 3b) shows significant correlations (p-value = $4.00e-5$ and $1.09e-3$, respectively, t-test, $\alpha < 0.05$, adjusted for multiple comparisons using Benjamini and Hochberg, 1995) between cpDFC2 and two swDFC variants (swDFC40w_5s, swDFC35w_20s), indicating redundancy between these feature spaces.

4 Discussion

Our results show that subject-specific temporal segmentation via change point DFC via FBS_cpDFC2, yields the strongest and most stable eMCI classification performance in ADNI rs-fMRI. Although the locations of the change points vary widely between individuals, the rate of change itself is not directly discriminative; instead, cpDFC's advantage comes from capturing individualized temporal dynamics without imposing fixed window parameters. Across the entire grid, swDFC shows high variability, occasional performance below SFC, and unstable feature selection, underscoring the risks of exhaustive parameter search and multiple hypothesis testing. Furthermore, cpDFC achieves these results with 33% fewer selected features (28 vs. 42) and without costly hyperparameter tuning, making it less prone to overfitting than swDFC in small-sample, high-dimensional settings.

Post-hoc ensemble analysis combining the best models yields further gains, supporting the idea that multi-timescale information improves classification. Significant correlations between some model predictions indicate partial redundancy, but diverse temporal scales still provide complementary features. Performance gains plateau after the top-8 ensemble, suggesting diminishing returns. Replication on an independent MCI dataset (Supplementary Materials C.2) confirmed the superior performance of cpDFC over all swDFC variants, despite differences in acquisition and preprocessing, suggesting robustness to dataset-specific factors.

Overall, our findings suggest: (1) individualized temporal segmentation is more robust than fixed-window approaches; (2) change point detection offers a principled alternative to swDFC for dynamic FC estimation; and (3) integrating multiple temporal resolutions can capture complementary neural dynamics for downstream tasks. Our cpDFC approach exemplifies a data-driven, segmentation-first approach that circumvents the need to tune hyperparameters such as window size, enabling stable, interpretable features from high-dimensional brain signals. Our ensemble findings highlight how multi-timescale fusion can enhance classification by aggregating complementary information.

We acknowledge that our work has some limitations. One is that the smaller sample size ($n = 68$) may limit the ability to generalize to larger populations, particularly given the heterogeneity of eMCI. Additionally, our study uses some fixed pre-processing and parcellation choices which could influence downstream classification. Future works could explore larger studies and jointly learn segment boundaries and within-segment representations end-to-end using deep-learning approaches.

References

Timofey Adamovich, Ilya Zakharov, Anna Tabueva, and Sergey Malykh. The thresholding problem and variability in the EEG graph network parameters. *Scientific Reports*, 12(1):18659, 2022.

Elena A Allen, Eswar Damaraju, Sergey M Plis, Erik B Erhardt, Tom Eichele, and Vince D Calhoun. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, 24(3):663–676, 2014.

Feng Bai, David R Watson, Hui Yu, Yongmei Shi, Yonggui Yuan, and Zhijun Zhang. Abnormal resting-state functional connectivity of posterior cingulate cortex in amnestic type mild cognitive impairment. *Brain Research*, 1302:167–174, 2009.

Danielle S Bassett and Edward T Bullmore. Small-world brain networks revisited. *The Neuroscientist*, 23(5):499–516, 2017.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34(4):537–541, 1995.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

Mathew R Brier, Jewell B Thomas, Abraham Z Snyder, Tammie L Benzinger, Dongyang Zhang, Marcus E Raichle, David M Holtzman, John C Morris, and Beau M Ances. Loss of intranetwork and internetwork resting state functional connections with Alzheimer's disease progression. *Journal of Neuroscience*, 32(26):8890–8899, 2012.

Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124 (1):1–38, 2008.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT press, 2022.

Ivor Cribben and Yi Yu. Estimating whole-brain dynamics by using spectral clustering. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 66(3):607–627, 2017. ISSN 14679876. doi: 10.1111/rssc.12169.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

Evan M Gordon, Timothy O Laumann, Babatunde Adeyemo, Jeremy F Huckins, William M Kelley, and Steven E Petersen. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral Cortex*, 26(1):288–303, 2016. ISSN 1047-3211. doi: 10.1093/cercor/bhu239. URL `http://10.0.4.69/cercor/bhu239https://dx.doi.org/10.1093/cercor/bhu239`.

Daniel A Handwerker, Vinai Roopchansingh, Javier Gonzalez-Castillo, and Peter A Bandettini. Periodic changes in fMRI connectivity. *NeuroImage*, 63(3):1712–1719, 2012.

R Matthew Hutchison, Thilo Womelsdorf, Elena A Allen, Peter A Bandettini, Vince D Calhoun, Maurizio Corbetta, Stefania Della Penna, Jeff H Duyn, Gary H Glover, Javier Gonzalez-Castillo, et al. Dynamic functional connectivity: promise, issues, and interpretations. *NeuroImage*, 80: 360–378, 2013.

David T Jones, Prashanthi Vemuri, Matthew C Murphy, Jeffrey L Gunter, Matthew L Senjem, Mary M Machulda, Scott A Przybelski, Brian E Gregg, Kejal Kantarci, David S Knopman, et al. Non-stationarity in the "resting brain's" modular architecture. *PLOS One*, 7(6):e39731, 2012.

Vito Latora and Massimo Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701, 2001.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Nora Leonardi and Dimitri Van De Ville. On spurious and real fluctuations of dynamic functional connectivity during rest. *NeuroImage*, 104:430–436, 2015.

Daniele Mascali, Mauro DiNuzzo, Tommaso Gili, Marta Moraschi, Michela Fratini, Bruno Maraviglia, Laura Serra, Marco Bozzali, and Federico Giove. Intrinsic patterns of coupling between correlation and amplitude of low-frequency fMRI fluctuations are disrupted in degenerative dementia mainly due to functional disconnection. *PLOS One*, 10(4):e0120988, 2015.

Sarah Feldt Muldoon and Danielle S Bassett. Network and multilayer network approaches to understanding human brain dynamics. *Philosophy of Science*, 83(5):710–720, 2016.

Mark EJ Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

Martin Ondrus and Ivor Cribben. fabisearch: A package for change point detection in and visualization of the network structure of multivariate high-dimensional time series in R. *Neurocomputing*, 578: 127321, 2024.

Martin Ondrus, Emily Olds, and Ivor Cribben. Factorized binary search: change point detection in the network structure of multivariate high-dimensional time series. *Imaging Neuroscience*, 2025.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *arXiv preprint arXiv:1603.05766*, 2016.

Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010.

Sean Ryan and Rebecca Killick. Detecting changes in covariance via random matrix theory. *Technometrics*, 65(4):480–491, 2023.

Reisa A Sperling, Bradford C Dickerson, Maija Pihlajamaki, Patrizia Vannini, Peter S LaViolette, Ottavio V Vitolo, Trey Hedden, J Alex Becker, Dorene M Rentz, Dennis J Selkoe, et al. Functional alterations in memory networks in early Alzheimer's disease. *Neuromolecular Medicine*, 12:27–43, 2010.

SPM. Statistical parametric mapping, 2023. URL `https://www.fil.ion.ucl.ac.uk/spm/`.

Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

Stanley Wasserman and Katherine Faust. *Social Network Analysis in the Social and Behavioral Sciences*, page 3–27. Structural Analysis in the Social Sciences. Cambridge University Press, 1994.

Susan Whitfield-Gabrieli and Alfonso Nieto-Castanon. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectivity*, 2(3):125–141, 2012.

Andrew Zalesky and Michael Breakspear. Towards a statistical test for functional connectivity dynamics. *NeuroImage*, 114:466–470, 2015.

Yongxia Zhou, John H Dougherty Jr, Karl F Hubner, Bing Bai, Rex L Cannon, and R Kent Hutson. Abnormal connectivity in the posterior cingulate and hippocampus in early Alzheimer's disease and mild cognitive impairment. *Alzheimer's & Dementia*, 4(4):265–270, 2008.

A Notation

We denote a matrix with a bold capital letter $\boldsymbol{A}$ and vectors as bold lowercase letters $\boldsymbol{a}$. The entry corresponding to the ith row and the jth column in $\boldsymbol{A}$ is $\boldsymbol{A}_{ij}$. The vector of the ith column in $\boldsymbol{A}$ is given by $\boldsymbol{A}_i$. We denote a single subject's fMRI data by $\boldsymbol{X}$, where each data matrix $\boldsymbol{X} \in \mathbb{R}^{T \times p}$ has T time points and p ROIs. A time point t is an element of the time index set $\{1, \ldots, T\}$. A stationary segment in $\boldsymbol{X}$ is denoted by $\boldsymbol{S} = \{x_t \in \mathbb{R}^p : t_1 \leq t \leq t_2\}$ where $t_1, t_2 \in \{1, \ldots, T\}$. A graph $\mathcal{G}$ is defined as a collection of vertices and edges $\mathcal{G} = (V, E)$. We use the terms “graph" and “network" interchangeably throughout. The number of samples or subjects in each study is denoted by n.

B Additional results

B.1 Additional classification results

In Figure 4 we report the area under the receiver operating curve for all swDFC analyses. We include the results for SFC and multiple change points (FBS_cpDFC2) plotted as vertical lines at their respective values, in order to better contextualize the differences between each technique.

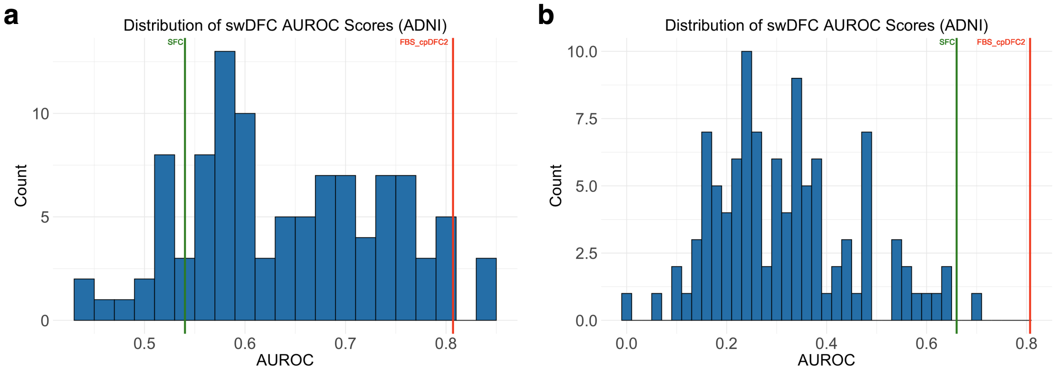


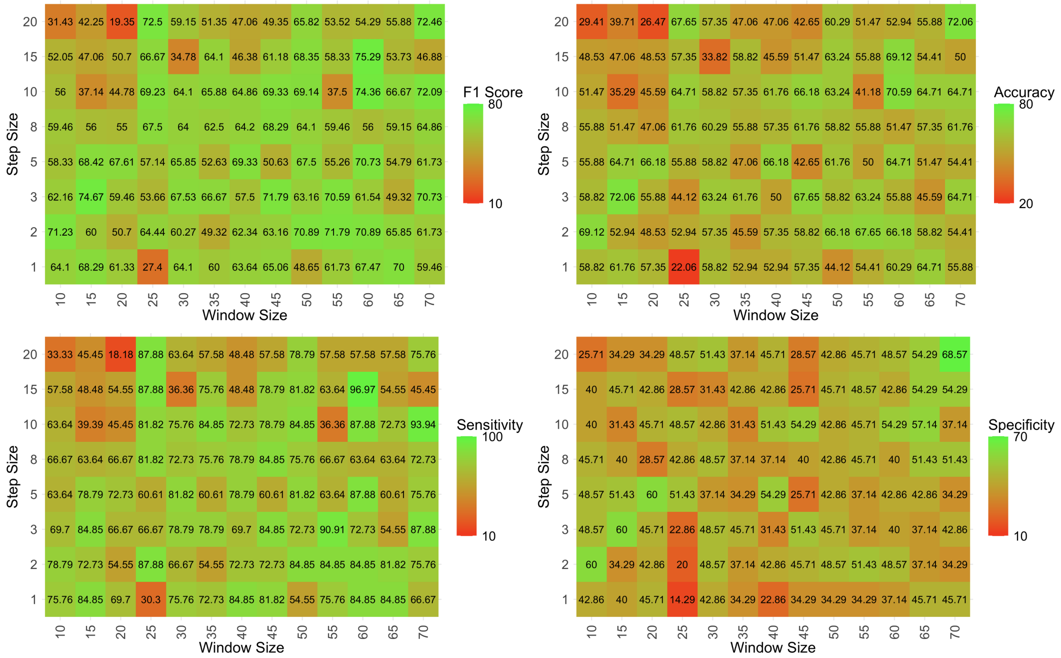
Figure 4: Histograms of the swDFC AUROC scores for the (a) ADNI, and (b) Mascali et al. [2015] rs-fMRI datasets. In each plot, we also provide a vertical line corresponding to the SFC result (green) and FBS_cpDFC2 result (red) for context.

To evaluate whether improvements to classification performance are attributable to the change point estimation procedure rather than randomly segmenting the time series at two discrete time locations, we generate a random segmentation (or “sham” change point) baseline. In particular, for each subject, we uniformly sample two “sham” change point locations which adhere to the pre-specified minimum distance change points $\delta = 30$ used in FBS_cpDFC2. This means that for each subject, the “sham” change points must be at least 30 time points away from the beginning of the time series ($t = 0$), the end of the time series ($t = T$), and from each other ($|t_2 - t_1|$). We then proceed with the same remaining steps for feature generation, selection, and classification. This procedure is then repeated 1000 times to generate a null distribution of classification performance under random segmentation.

The resulting distribution of AUROC values is worse than chance-level performance ($\mu = 0.3815, \sigma = 0.1229$) In contrast, multiple change point estimation via FBS_cpDFC2 achieved an AUROC of 0.8069, above the mean of the permutation distribution (permutation based one-sided p-value: $9.99e - 4$, corrected via Phipson and Smyth [2016]). This evidence supports the notion that estimated change points under a formalized change point detection procedure provide non-random segmentation that captures meaningful temporal structure, and that this is not simply an artifact of segmenting the time series.

B.2 Sliding window results

Figure 5 shows the results of different combinations of window size and step size for swDFC. Across the various combinations, there is no clear relationship between step size and window size. Additionally, there are several combinations that are plagued by instability, as shown by the large variability in step and window sizes in neighboring combinations (or tiles). For example, the combination of window size 15 and step size 3 achieved the best accuracy of 72.06%. However, changing the window size by 5 time points in either direction (to either 10 or 20), drastically reduces the performance of these models (to 58.82% and 55.88%, respectively). The same is also evident when the step size is changed to 2 or 8: the performance of the models reduces to 52.94% and 64.71%, respectively.



Step Size	10	15	20	25	30	35	40	45	50	55	60	65	70
20	31.43	42.25	19.35	72.5	59.15	51.35	47.06	49.35	65.82	53.52	54.29	55.88	72.46
15	52.05	47.06	50.7	66.67	34.78	64.1	46.38	61.18	68.35	58.33	75.29	53.73	46.88
10	56	37.14	44.78	69.23	64.1	65.88	64.86	69.33	69.14	37.5	74.36	66.67	72.09
8	59.46	56	55	67.5	64	62.5	64.2	68.29	64.1	59.46	56	59.15	64.86
5	58.33	68.42	67.61	57.14	65.85	52.63	69.33	50.63	67.5	55.26	70.73	54.79	61.73
3	62.16	74.67	59.46	53.66	67.53	66.67	57.5	71.79	63.16	70.59	61.54	49.32	70.73
2	71.23	60	50.7	64.44	60.27	49.32	62.34	63.16	70.89	71.79	70.89	65.85	61.73
1	64.1	68.29	61.33	27.4	64.1	60	63.64	65.06	48.65	61.73	67.47	70	59.46

Step Size	10	15	20	25	30	35	40	45	50	55	60	65	70
20	29.41	39.71	26.47	67.65	57.35	47.06	47.06	42.65	60.29	51.47	52.94	55.88	72.06
15	48.53	47.06	48.53	57.35	33.82	58.82	45.59	51.47	63.24	55.88	69.12	54.41	50
10	51.47	35.29	45.59	64.71	58.82	57.35	61.76	66.18	63.24	41.18	70.59	64.71	64.71
8	55.88	51.47	47.06	61.76	60.29	55.88	57.35	61.76	58.82	55.88	51.47	57.35	61.76
5	55.88	64.71	66.18	55.88	58.82	47.06	66.18	42.65	61.76	50	64.71	51.47	54.41
3	58.82	72.06	55.88	44.12	63.24	61.76	50	67.65	58.82	63.24	55.88	45.59	64.71
2	69.12	52.94	48.53	52.94	57.35	45.59	57.35	58.82	66.18	67.65	66.18	58.82	54.41
1	58.82	61.76	57.35	22.06	58.82	52.94	52.94	57.35	44.12	54.41	60.29	64.71	55.88

Step Size	10	15	20	25	30	35	40	45	50	55	60	65	70
20	33.33	45.45	18.18	87.88	63.64	57.58	48.48	57.58	78.79	57.58	57.58	57.58	75.76
15	57.58	48.48	54.55	87.88	36.36	75.76	48.48	78.79	81.82	63.64	96.97	54.55	45.45
10	63.64	39.39	45.45	81.82	75.76	84.85	72.73	78.79	84.85	36.36	87.88	72.73	93.94
8	66.67	63.64	66.67	81.82	72.73	75.76	78.79	84.85	75.76	66.67	63.64	63.64	72.73
5	63.64	78.79	72.73	60.61	81.82	60.61	78.79	60.61	81.82	63.64	87.88	60.61	75.76
3	69.7	84.85	66.67	66.67	78.79	78.79	69.7	84.85	72.73	90.91	72.73	54.55	87.88
2	78.79	72.73	54.55	87.88	66.67	54.55	72.73	72.73	84.85	84.85	84.85	81.82	75.76
1	75.76	84.85	69.7	30.3	75.76	72.73	84.85	81.82	54.55	75.76	84.85	84.85	66.67

Step Size	10	15	20	25	30	35	40	45	50	55	60	65	70
20	25.71	34.29	34.29	48.57	51.43	37.14	45.71	28.57	42.86	45.71	48.57	54.29	68.57
15	40	45.71	42.86	28.57	31.43	42.86	42.86	25.71	45.71	48.57	42.86	54.29	54.29
10	40	31.43	45.71	48.57	42.86	31.43	51.43	54.29	42.86	45.71	54.29	57.14	37.14
8	45.71	40	28.57	42.86	48.57	37.14	37.14	40	42.86	45.71	40	51.43	51.43
5	48.57	51.43	60	51.43	37.14	34.29	54.29	25.71	42.86	37.14	42.86	42.86	34.29
3	48.57	60	45.71	22.86	48.57	45.71	31.43	51.43	45.71	37.14	40	37.14	42.86
2	60	34.29	42.86	20	48.57	37.14	42.86	45.71	48.57	51.43	48.57	37.14	34.29
1	42.86	40	45.71	14.29	42.86	34.29	22.86	34.29	34.29	34.29	37.14	45.71	45.71

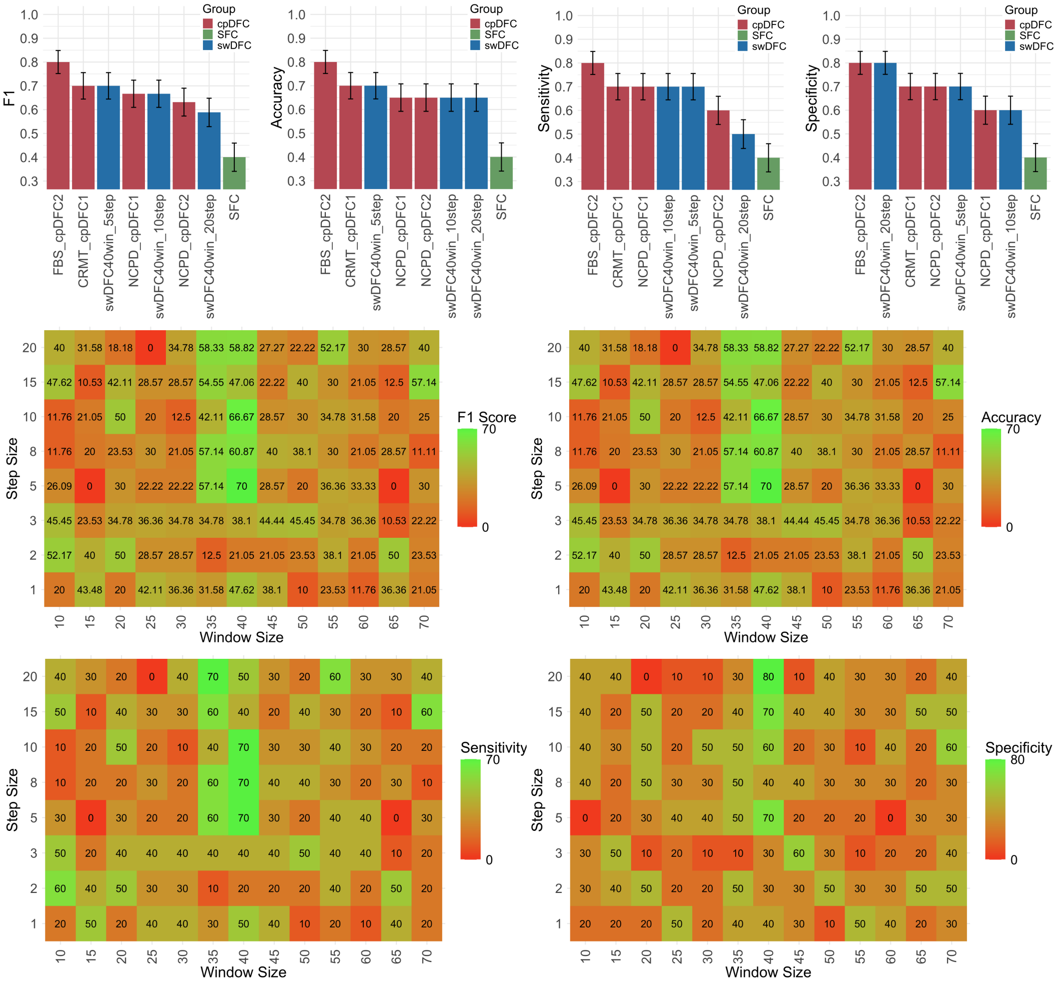
Figure 5: Heatmaps of F1 score, accuracy, sensitivity, and specificity from the classification study of CN subjects and subjects with eMCI from the ADNI rs-fMRI dataset using swDFC. Results are shown across window sizes [10, 70] in increments of 5 and step sizes [1, 2, 3, 5, 8, 10, 15, 20].

Figure 5 shows the results of different combinations of window size and step size for swDFC. Across the various combinations, there is no clear relationship between step size and window size. Additionally, there are several combinations that are plagued by instability, as shown by the large variability in step and window sizes in neighboring combinations (or tiles). For example, the combination of window size 15 and step size 3 achieved the best accuracy of 72.06%. However, changing the window size by 5 time points in either direction (to either 10 or 20), drastically reduces the performance of these models (to 58.82% and 55.88%, respectively). The same is also evident when the step size is changed to 2 or 8: the performance of the models reduces to 52.94% and 64.71%, respectively.

B.3 Secondary dataset

We present the classification results from the Mascali et al. [2015] dataset. We also include two additional change point detection methodologies, specifically Network Change Point Detection (NCPD: Cribben and Yu, 2017) and Covariance Change Points through Random Matrix Theory (CRMT: Ryan and Killick, 2023). Figure 6 (top panel) shows the classification results using SFC, swDFC and cpDFC methods. Similar to the ADNI data set, FaBiSearch with two change points has a superior performance compared to the other methods. In addition, it appears that change point methods outperform swDFC methods. In Figure 6 (bottom panels), we again find that there is no particular patterns between the different combinations of step size and window size and classification performance. The best performing models have a window size of approximately 40 and step size between 5 and 20 depending on the evaluation metric. For swDFC, performance on this dataset mirrors ADNI but is often inferior. Given the smaller size of this study ($n = 20$) compared to the

ADNI data set ($n = 68$), it is possible that the performance discrepancies between the combinations of window and step size are further exacerbated in smaller sample size settings. Figure 7(a) shows the ensemble analysis for the Mascali et al. [2015] dataset. The lack of a monotonically increasing performance pattern as more models are combined may stem from the small sample size. Alternatively, it could be due to the larger number of higher correlations between the models (Figure 7(b)).



F1 Score

Step Size	10	15	20	25	30	35	40	45	50	55	60	65	70
20	40	31.58	18.18	0	34.78	58.33	58.82	27.27	22.22	52.17	30	28.57	40
15	47.62	10.53	42.11	28.57	28.57	54.55	47.06	22.22	40	30	21.05	12.5	57.14
10	11.76	21.05	50	20	12.5	42.11	66.67	28.57	30	34.78	31.58	20	25
8	11.76	20	23.53	30	21.05	57.14	60.87	40	38.1	30	21.05	28.57	11.11
5	26.09	0	30	22.22	22.22	57.14	70	28.57	20	36.36	33.33	0	30
3	45.45	23.53	34.78	36.36	34.78	34.78	38.1	44.44	45.45	34.78	36.36	10.53	22.22
2	52.17	40	50	28.57	28.57	12.5	21.05	21.05	23.53	38.1	21.05	50	23.53
1	20	43.48	20	42.11	36.36	31.58	47.62	38.1	10	23.53	11.76	36.36	21.05

Accuracy

Step Size	10	15	20	25	30	35	40	45	50	55	60	65	70
20	40	31.58	18.18	0	34.78	58.33	58.82	27.27	22.22	52.17	30	28.57	40
15	47.62	10.53	42.11	28.57	28.57	54.55	47.06	22.22	40	30	21.05	12.5	57.14
10	11.76	21.05	50	20	12.5	42.11	66.67	28.57	30	34.78	31.58	20	25
8	11.76	20	23.53	30	21.05	57.14	60.87	40	38.1	30	21.05	28.57	11.11
5	26.09	0	30	22.22	22.22	57.14	70	28.57	20	36.36	33.33	0	30
3	45.45	23.53	34.78	36.36	34.78	34.78	38.1	44.44	45.45	34.78	36.36	10.53	22.22
2	52.17	40	50	28.57	28.57	12.5	21.05	21.05	23.53	38.1	21.05	50	23.53
1	20	43.48	20	42.11	36.36	31.58	47.62	38.1	10	23.53	11.76	36.36	21.05

Sensitivity

Step Size	10	15	20	25	30	35	40	45	50	55	60	65	70
20	40	30	20	0	40	70	50	30	20	60	30	30	40
15	50	10	40	30	30	60	40	20	40	30	20	10	60
10	10	20	50	20	10	40	70	30	30	40	30	20	20
8	10	20	20	30	20	60	70	40	40	30	20	30	10
5	30	0	30	20	20	60	70	30	20	40	40	0	30
3	50	20	40	40	40	40	40	40	50	40	40	10	20
2	60	40	50	30	30	10	20	20	20	40	20	50	20
1	20	50	20	40	40	30	50	40	10	20	10	40	20

Specificity

Step Size	10	15	20	25	30	35	40	45	50	55	60	65	70
20	40	40	0	10	10	30	80	10	40	30	30	20	40
15	40	20	50	20	20	40	70	40	40	30	30	50	50
10	40	30	50	20	50	50	60	20	30	10	40	20	60
8	40	20	50	30	30	50	40	40	30	30	30	20	30
5	0	20	30	40	40	50	70	20	20	20	0	30	30
3	30	50	10	20	10	10	30	60	30	10	20	20	40
2	30	40	50	20	20	50	30	30	50	30	30	50	50
1	20	20	20	50	20	40	40	30	10	50	40	20	30

Window Size

Figure 6: The classification results of CN subjects and subjects with MCI from the Mascali et al. [2015] rs-fMRI dataset. (Top panel) F1, accuracy, sensitivity, and specificity results of all methods. SFC, cpDFC, and swDFC correspond to static, change point, and window based dynamic functional connectivity, respectively. (Bottom panels) Heatmaps of F1 score, accuracy, sensitivity, and specificity. Results are shown across window sizes [10, 70] in increments of 5 and step sizes [1, 2, 3, 5, 8, 10, 15, 20].

B.4 Change point detection and stationary segments

To estimate cpDFC, we apply FabiSearch [Ondrus et al., 2025] to each subject in the ADNI dataset. Figure 8 shows both the detected change points for the control group (CN) (Figure 8(a)) and the eMCI group (Figure 8(b)). Within each group, the change points are not consistently clustered around specific time points during the scan; rather, they are distributed relatively uniformly, aside from some edge effects near the start and end of the session. This observation suggests that the timing of change points is largely subject-specific, rather than being driven by a group-level pattern. A non-parametric Kolmogorov-Smirnov test on the change point locations between the two groups suggests that they are not significantly different ($p = 0.413$). A one-sided t-test on the means of the location of the first

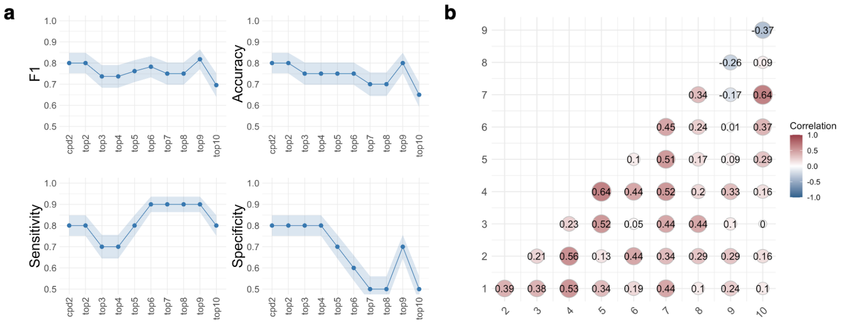


Figure 7: The ensemble model results from the classification study of CN subjects and subjects with eMCI in the Mascali et al. [2015] dataset. Panel (a) shows the ensemble results. For the ensemble model, we combined the predictions from the top-2 to the top-10 classifier models, as determined by their highest F1 score. The FBS_cpDFC2 model is the best stand-alone classifier. Shaded regions indicate ± the standard error of proportion calculated as $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. (b) shows the correlation between predicted probabilities of eMCI across the top-10 classifiers, as determined by the highest F1 score.

detected change points of the CN group and the eMCI groups suggests that the first change point for the CN group is significantly smaller (in time) than the eMCI group ($p = 0.028$).

Figure 8 also shows the stationary FC states (or modes) between each change point, where $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3$ refer to the first, second, and third stationary segments estimated by change point detection, respectively. Across all segments and even between the CN and subjects with eMCI, there is strong FC between the frontal and cerebellar regions. In the first FC state, $\boldsymbol{S}_1$, CN have stronger FC between the parietal regions and the rest of the brain, and also engage the temporal regions with the rest of the brain more as well. Subjects with eMCI, have stronger FC between the frontal and parietal regions.

In $\boldsymbol{S}_2$, CN subjects exhibit strong FC between the frontal and parietal regions. In contrast, subjects with eMCI show FC primarily concentrated in the frontal regions, with connections extending to the occipital and limbic systems. Notably, a small number of occipital nodes in subjects with eMCI mediate much of the FC to other brain regions. In $\boldsymbol{S}_3$, CN patients display robust FC among the frontal, occipital, and cerebellar regions, forming a tightly coupled triad. Conversely, subjects with eMCI exhibit more diffuse FC, with the cerebellum acting as a hub, extensively connected to the rest of the brain, particularly the parietal lobe.

B.5 Graph structure changes across time segments

In Figure 9(a), we show the ROIs that were selected by SIS in all folds in leave-one-out cross-validation for FBS_cpDFC2. Figure 9(b) includes more information on the selected features and the corresponding ROIs, such as node ID, feature type, stationary segment, and differences in the mean value for the features of the CN and the eMCI groups. We find that the most consistently chosen features correspond to the frontal, parietal, and cerebellum regions. Degree, betweenness, clustering coefficient, local efficiency, and shortest path were all key features that were selected across LOOCV folds. We also find that eMCI is associated with lower degree in the parietal region, and that paths were more efficient, shorter, and tightly clustered in the cerebellum and frontal regions. Later segments ($\boldsymbol{S}_2$, and $\boldsymbol{S}_3$) were chosen more often across these folds, which may be related to the subjects being more settled and closer to a true "resting" state compared to the beginning of the fMRI experiment.

As the results in the main article show, there is a concentration of strongly differentiating features in the later time segments (Figure 9) for the classification task. This is highlighted also in Figure 10, especially in the ROIs close to the main diagonal, and also in the squared difference row, which are the largest for time segments 2 and 3. It is evident that the group-wise characteristics become more stable and contribute most to the differentiation in later time windows. Furthermore, it is clear from Figure 10 that global changes in graph structure are subtle, since the averaged adjacency matrices

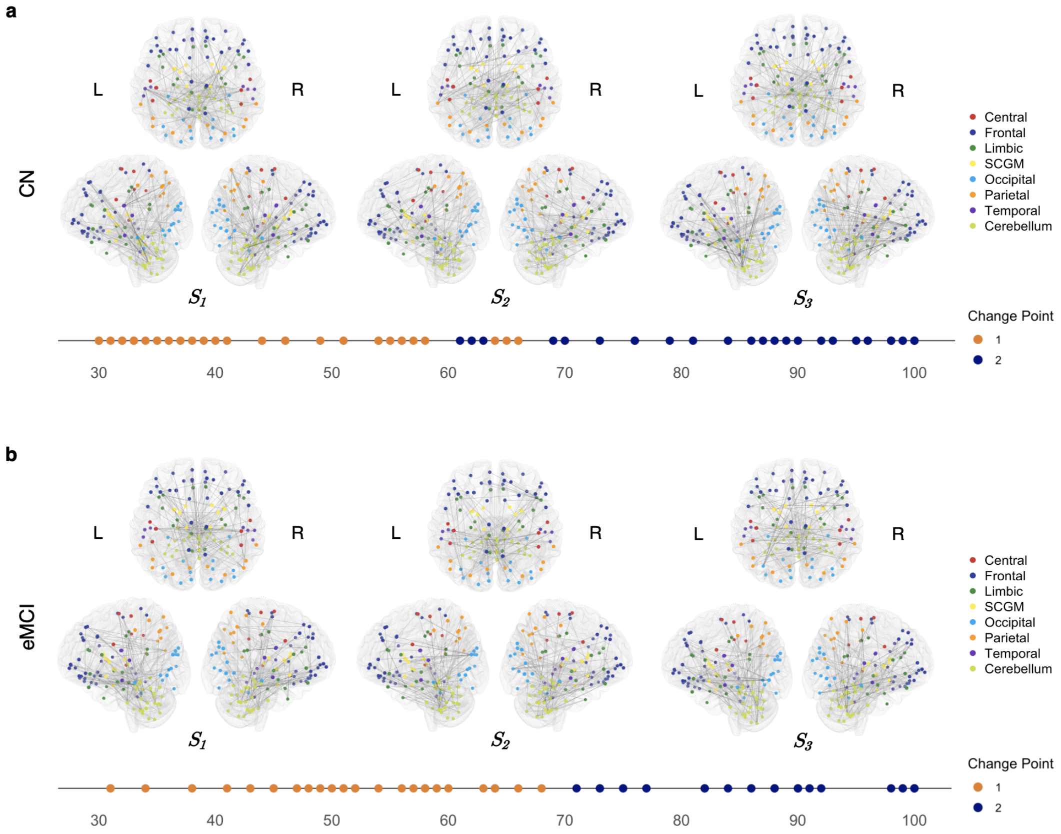
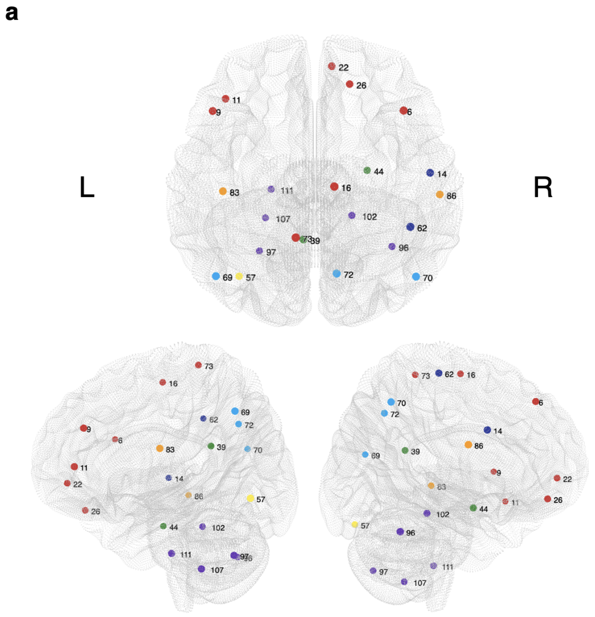


Figure 8: The change point detection results from applying FaBiSearch (taking only the first two change points, FBS_cpDFC2) and corresponding stationary FC states for (a) controls and (b) subjects with eMCI for the ADNI rs-fMRI dataset. For each panel, FC plots are shown for each stationary segment, where $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3$ correspond to the first, second, and third stationary segments. Individual change points are used to segment the time series, and then correlation matrices are averaged across subjects. The top 100 edges as determined by the absolute value of this averaged correlation are shown for each stationary segment. Below the FC plots, orange and blue points indicate the first and second detected change points, respectively, for each subject such that the points are pooled within each group (CN or eMCI).

are generally relatively similar, but localized differences in functional connectivity have widespread effects related to the global graph structure. For example, we observed differences in the shortest path metric, which can be strongly influenced by changes in one or a few edges, especially if there is a lack of redundancy in connections between nodes. Therefore, these subtle differences in the graph structure have strong consequences in distinguishing CN from patients with eMCI.

As we restrict each subject to two change points, each subject has 3 stationary segments, or partitions between change points (or stationary modes). We can then compare group-wise differences in the adjacency matrices between consecutive segments. We calculate the average adjacency matrix for each segment and class (CN or eMCI), using the following procedure. For each group (CN: $n = 35$, eMCI: $n = 33$), $\mu\boldsymbol{A}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{A}_{ij}$. We can then calculate the entry-wise difference between the groups using $\mu\boldsymbol{A}^{\text{diff}} = \mu\boldsymbol{A}^{\text{CN}} - \mu\boldsymbol{A}^{\text{eMCI}}$. Figure 10 (rows 1 and 2) shows the average adjacency matrices grouped by segment (1, 2, or 3) and class (CN or eMCI), while Figure 10 (row 3) shows the differences in class between segments for all 120 ROIs. The differences appear to concentrate close to the diagonal of the adjacency matrices. This difference is especially pronounced between ROIs 50-60. In addition, subjects with eMCI have stronger off-diagonal connections, notably at the intersection of ROIs 10-20 and 80-90.



b

Node ID	Folds	Region	Feature	Segment	CN	eMCI
70	68	Parietal	Degree	S_2	17.029	15.727
107	68	Cerebellum	Betweenness	S_3	79.675	83.258
6	68	Frontal	Clustering coefficient	S_3	0.413	0.427
73	68	Frontal	Clustering coefficient	S_3	0.396	0.439
6	68	Frontal	Local efficiency	S_3	0.675	0.690
97	68	Cerebellum	Shortest Path	S_3	3.943	6.030
73	66	Frontal	Local efficiency	S_3	0.667	0.663
57	42	Occipital	Betweenness	S_1	86.251	80.899
83	37	Temporal	Betweenness	S_3	70.976	57.667
111	34	Cerebellum	Shortest Path	S_3	5.286	10.061
44	29	Limbic	Local efficiency	S_1	0.515	0.666
9	13	Frontal	Betweenness	S_1	93.772	60.663
26	10	Frontal	Shortest Path	S_1	5.229	5.636
9	7	Frontal	Clustering coefficient	S_1	0.347	0.462
16	6	Frontal	Betweenness	S_2	52.337	79.090
22	5	Frontal	Local efficiency	S_1	0.580	0.707
86	4	Temporal	Shortest Path	S_3	5.000	8.970
73	3	Frontal	Clustering coefficient	S_1	0.384	0.524
62	3	Central	Local efficiency	S_3	0.695	0.727
86	2	Temporal	Local efficiency	S_3	0.676	0.687
62	1	Central	Clustering coefficient	S_1	0.401	0.508
73	1	Frontal	Local efficiency	S_1	0.588	0.723
39	1	Limbic	Shortest Path	S_1	7.000	6.788
14	1	Central	Betweenness	S_2	84.906	87.841
97	1	Cerebellum	Betweenness	S_2	84.796	67.681
102	1	Cerebellum	Clustering coefficient	S_2	0.416	0.488
72	1	Parietal	Betweenness	S_3	64.736	84.901
62	1	Central	Clustering coefficient	S_3	0.438	0.493
11	1	Frontal	Local efficiency	S_3	0.612	0.669
69	1	Parietal	Local efficiency	S_3	0.685	0.690
96	1	Cerebellum	Shortest Path	S_3	5.086	6.848

Figure 9: (a) The associated regions of interest (ROIs) of the features selected across all folds by SIS in leave-one-out cross validation for FBS_cpDFC2 in the classification study of CN subjects and subjects with eMCI from the ADNI rs-fMRI dataset. (b) The node ID, the number of LOOCV folds that the feature was chosen, region from the AAL atlas, the graph theoretic feature type, the stationary segment (i.e., 1 = first, 2 = second, 3 = third), and the mean values of the features of the CN and eMCI groups. The features are ordered in descending order based on how often they were selected across the LOOCV folds.

C Data and pre-processing

C.1 ADNI data

The first dataset we obtained from the ADNI database (`http://adni.loni.usc.edu`). ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary objective of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological evaluations can be combined to measure the progression of MCI and early AD. For up-to-date information, see `www.adni-info.org`. In ADNI's rs-fMRI experiments, subjects were instructed to remain still and relaxed in the scanner. Subjects level data were obtained using Phillips scanners and include 33 subjects with eMCI (mean age 72.3, 15M/18F) and 35 healthy controls (mean age 74.6, 14M/21F).

Data were pre-processed using SPM [SPM, 2023], following a standard fMRI preprocessing pipeline. The first 10 volumes were discarded to account for initial scanner and subject noise. Next, a slice timing correction was performed (`spm_slice_timing`) to correct for timing differences between slices within each volume. The images were then realigned using `spm_realign`, which applies a rigid-body transformation to align each volume with the mean functional image. Motion parameters were estimated using least-squares with 2nd degree B-spline interpolation. The estimated motion parameters were subsequently applied to reslice all volumes using 4th degree B-spline interpolation to minimize resampling artifacts. The images were then normalized to the Montreal Neurological Institute (MNI) space with 3mm×3mm×3mm voxels (`spm_normalise`). Nuisance covariates (Friston 24, cerebrospinal fluid, white matter, and global mean) were regressed out. The voxels were then spatially smoothed using a Gaussian kernel (FWHM = 6mm) in `spm_smooth`. A linear trend was removed from each time series using ordinary least squares regression (`spm_detrend`) and a fourth order Butterworth low-pass filter (0.01–0.08 Hz) was applied to attenuate high-frequency noise and low-frequency drift outside the typical BOLD signal range (`spm_filter`). We used the Automated Anatomical Labeling (AAL2) atlas to subdivide the brain into 120 anatomical regions [Tzourio-

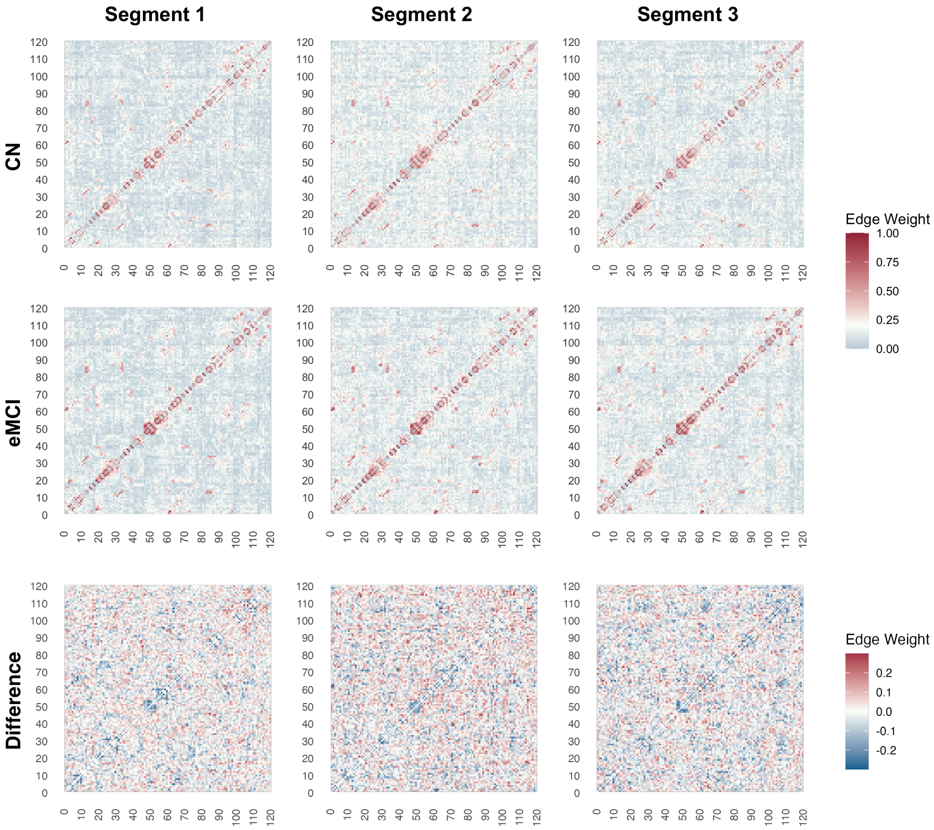


Figure 10: Averaged adjacency matrices across temporal segments (columns), and subject class (rows) for the FaiSearch (cpDFC2) method in the classification study of CN subjects and subjects with eMCI from the ADNI rs-fMRI dataset. Node labels (1 to 120) correspond to the AAL atlas [Tzourio-Mazoyer et al., 2002] labels. The difference between the CN and the eMCI adjacency matrices computed for each segment is shown in the third row.

Mazoyer et al., 2002]. We chose the AAL2 atlas because of its widespread use in neuroimaging studies, particularly in Alzheimer's disease and MCI research, allowing for comparisons with other works. AAL2 provides whole-brain coverage with a moderate number of parcels, providing a balance between spatial resolution and statistical power. The region of interest (ROI) time series for a given region was defined as the average of the time series of all voxels within that region. Finally, each time series was z-score normalized. The final preprocessed dataset for each subject was a time series of dimension $T = 130$ time points by $p = 120$ ROIs.

C.2 Secondary dataset

In order to check the robustness of our results, we apply our methods to a second study of MCI classification [Mascali et al., 2015], which includes 10 patients with mild cognitive impairment and 10 healthy elderly controls. Participants were told to lie quietly with their eyes closed without falling asleep. A 3T MRI system (Magnetom Allegra, Siemens, Erlangen, Germany) was used to acquire images, with the following properties: TR = 2080 ms, TE = 30 ms, 32 axial slices parallel to the AC-PC plane, matrix = 64 x 64, in plane resolution = 3x3 mm^2, slice thickness = 2.5 mm, 50% skip, flip angle = 70°. Functional images were preprocessed using the Connectivity toolbox [Whitfield-Gabrieli and Nieto-Castanon, 2012]. The initial four volumes were discarded for signal and scanner stabilization, resulting in 216 time points per subject, and images were slice-time corrected and realigned to the first image. More detailed information on the preprocessing steps can be found in Mascali et al. [2015]. Finally, the atlas of Gordon et al. [2016] was used to parcellate the brain into ROIs. To explore the generalizability of our approach while maintaining comparability with the main study (120 ROIs), we selected the Default [Sperling et al., 2010, Buckner et al., 2008], Frontoparietal [Brier et al., 2012], Cinguloparietal [Bai et al., 2009], and Dorsal Attention [Zhou et al., 2008] communities, resulting in a reduced dimensionality of 102 ROIs. This allowed us to test the method on another dataset with slightly different preprocessing and experimental conditions, while keeping the parcellation scale comparable, to assess the robustness of the results.

D Algorithms for time-varying functional connectivity estimation

In this section, we provide a brief overview of the algorithms used to estimate and analyze dynamic functional connectivity, including a standard sliding-window correlation approach and the FaBiSearch change point detection method Ondrus et al. [2025]. For further technical details are available on FaBiSearch, we refer readers to the original publication [Ondrus et al., 2025].

Algorithm 1: Algorithm for estimating dynamic functional connectivity using sliding windows.

```
Inputs: fMRI data matrix $\boldsymbol{X}$, window size $w$, step size $s$
Initialize the starting point and end points of the window as $t_{\text{start}} = 1$ and $t_{\text{end}} = w$, respectively.
while $t_{\text{start}} + w - 1 \leq T$ do
    $t_{\text{end}} \leftarrow t_{\text{start}} + w - 1$
    Estimate FC within the window $\boldsymbol{X}_{t_{\text{start}}\,:\,t_{\text{end}}}$
    Advance the window, $t_{\text{start}} \leftarrow t_{\text{start}} + s, \quad t_{\text{end}} \leftarrow t_{\text{start}} + w - 1$
end
```

Algorithm 2: Change point detection algorithm using FaBiSearch which can be applied recursively to find multiple change points.

```
Inputs: data matrix $\boldsymbol{X}$, minimum distance $\delta$, number of runs of NMF $n_{run}$, number of
        permutation repetitions $n_{reps}$
Find optimal rank, $r$, over all of $\boldsymbol{X}$
Evaluate fitments of $\boldsymbol{X}$ to NMF using Kullback-Leibler divergence to navigate over $T$.
Continue until convergence to a single $t$ which becomes the candidate change point to be
  evaluated.
Calculate the p-value of the candidate change point to determine whether splitting at this point
  improves Kullback-Leibler divergence over the null distribution.
```

E Functional connectivity through graph estimation

We begin by introducing the concept of stationarity, where a stationary segment in the multivariate BOLD time series is defined as one in which the statistical properties are assumed to not change with respect to time [Jones et al., 2012, Handwerker et al., 2012]. The goal of DFC then, is to either model the dynamics of these changing properties, or, to segment the time series to isolate stationary segments. From the DFC approaches outlined in the next section, we obtain a series of stationary segments for each subject, while for SFC, stationary is assumed over the entire experimental time course, and thus one segment is obtained for each subject. For each stationary segment **S**, we estimate the functional connectivity (FC), or linear dependency, between the ith and jth time series using Pearson's correlation coefficient ρ_{ij}. In the brain fMRI setting, the graph $\mathcal{G} = (V, E)$ is defined such that each vertex $v_i \in V$ represents a region of interest (ROI) or parcellated brain region, and each edge $e_{ij} \in E$ is weighted by the corresponding correlation ρ_{ij} between ROI time series i and j. Because we are primarily interested in strong functional connections, we apply an absolute correlation threshold of 0.5, consistent with the range (0.1–0.8) commonly used in the literature [Adamovich et al., 2022]. Thus, an edge e_{ij} remains in $\mathcal{G}$ if $|\rho_{ij}| > 0.5$. A schematic of the SFC estimation procedure is shown in Figure 1(b). The resulting adjacency matrix encodes the network structure for the segment, with retained edge weights reflecting the strength of functional coupling between brain regions as measured by correlation. More formally, let $\tau = 0.5$. The edge set is

$$E = \big\{\{i, j\} : i < j,\ |\rho_{ij}| > \tau\big\},$$

and the weighted adjacency is $A_{ij} = \rho_{ij}$ if $\{i, j\} \in E$ and $A_{ij} = 0$ otherwise (with $A_{ii} = 0$).

F Estimating DFC

To estimate DFC, we assume that we have no prior knowledge of the distribution of $\boldsymbol{X}$, and that there may be an unknown number of distributional shifts, with the number and locations in $\{1, \ldots, T\}$ and the rate of change between them is assumed unknown. We do however assume that there exist segments in $\boldsymbol{X}$ which *are* stationary. We consider two approaches, sliding windows (swDFC) and change points (cpDFC), for estimating time-varying connectivity in rs-fMRI. swDFC is a naive approach in which no data information is used to determine the temporal structure. Here, a window of a particular length, w, is used to subselect the set of time indices in $\{1, \ldots, T\}$. As its name suggests, the window is then slid over a predetermined length, s, to define the next window. Consequently, the sliding window captures dynamic information by subsampling the entirety of the distribution through a predetermined and overlapping set of time-dependent steps, where within each of these windows, FC is estimated as described in Section E. We describe the sliding-window method in Algorithm 1 (in the Appendix) and show a schematic in Figure 1(c).

The choices for w and s are crucial, as they determine the granularity and the amount of overlap between snapshots that this technique captures. However, the values for these parameters are context dependent and it is not possible to derive them directly from $\boldsymbol{X}$. As a consequence of this, we vary the combinations of w and s in our experiments and test the classification performance.

We implement cpDFC using FaBiSearch, a change point detection technique in the network structure between (high-dimensional) multivariate time series [Ondrus et al., 2025]. More formally, we consider the multiple change point detection problem in a multivariate time series data $\boldsymbol{X}$. We seek to find the time points where the network structure of $\boldsymbol{X}$ changes. FaBiSearch utilizes non-negative matrix factorization (NMF: Lee and Seung, 1999) to identify multiple change points in high-dimensional time series data. In particular, change points are identified in a sequential manner through successive splitting of the time series. No limit is imposed on the number of change points; instead, detection proceeds until further splitting cannot be performed. Candidate change points are then evaluated through a permutation test procedure. The method requires a series of hyperparameters; $\boldsymbol{X}$, α, δ, n_{run}, n_{reps}, r which refer to the input multivariate time series data, significance for the permutation test, minimum distance between change points, maximum number of runs of the NMF algoirthm, and the number of permutation in the permutation test. The rank of NMF to use in the procedure, r, can be determined from the data by an iterative permutation procedure. Algorithm 2 (in the Appendix) summarizes FaBiSearch, and we utilize a loss based on Kullback-Leibler divergence to assess the fit of the model. For more details on FaBiSearch, see Ondrus et al. [2025].

In contrast to the sliding window method, which can be highly sensitive to the precise choice of parameters such as window length and step size, Change point detection does not require parameter values to be so finely tuned for optimal performance. Instead, it only requires sufficient settings, meaning enough data samples, permutation iterations, and related hyperparameters, to reliably detect change points, without the need to identify a single "best" configuration. Ondrus and Cribben [2024] show through a sensitivity analysis on simulated data that the accuracy of change point detection using FaBiSearch plateaus beyond sufficient values of the input hyperparameters. In their simulations, Ondrus and Cribben [2024] found that performance plateaued once $\delta \geq 30$, $nrun \geq 100$, $nreps \geq 100$, denoting the minimum distance between change points, the number of NMF runs for convergence, and the number of permutations for the significance test, respectively. Increasing these values beyond these thresholds produced only marginal improvements in detection accuracy.

For each subject, we first estimate the change points using FaBiSearch and then order the change points $\hat{q}$ by their respective p-value from smallest to largest. After defining the set of change points, $\hat{q}$, we partition $\boldsymbol{X}$ into stationary time segments between the change points. In the classification study, we used $k = 1$ and $k = 2$ change points to define the stationary segments. Finally, for each of the stationary segments, we estimate FC as described in Section E.

We also estimate cpDFC using two other change point detection methods, specifically Network Change Point Detection (NCPD: Cribben and Yu, 2017) and Covariance Change Points through Random Matrix Theory (CRMT: Ryan and Killick, 2023). For NCPD and CRMT, we used similar hyperparameters to FaBiSearch, that is, a similar minimum distance between the change points and the optimal rank. All other hyperaparameters for these methods were set to their default values. For CRMT, we further pre-processed ROI time series by performing a truncated SVD with the same optimal rank as used in FaBiSearch to satisfy the theoretical condition that $p \leq n$ for the estimator to be well-behaved. A schematic of change point detection is shown in Figure 1(d).

G Graph Theoretic Features

Degree

The degree of a node can be calculated from the following definition:

$$d_i = \sum_{j \in V} a_{ij} \ ,$$

where $d_i \in \mathbb{N}_0$ is the degree of the node i, V is the set of all nodes, and a_{ij} is the intersection of the nodes i and j in the adjacency matrix.

Clustering coefficient

From Wasserman and Faust [1994], it is given by:

$$C_i = \frac{2e_i}{d_i(d_i - 1)} \ ,$$

where $C_i \in [0, 1]$ is the clustering coefficient for node i, d_i is the degree of node i, and e_i is the number of edges between node i and neighbors d_i.

Shortest path

We use the notation $\ell_{ij} \in \mathbb{N}_0$ as the fewest number of edges that connect nodes i and j together to define the shortest path. The shortest path can be calculated using different algorithms, although we use breadth-first search (BFS: Cormen et al., 2022) in our implementation.

Degree assortativity

Newman [2002] define degree assortativity as

$$r_{jk} = \frac{m^{-1} \sum_i j_i k_i - \left[m^{-2} \sum_i \frac{1}{2}(j_i + k_i)\right]^2}{m^{-1} \sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[m^{-2} \sum_i \frac{1}{2}(j_i + k_i)\right]^2}$$

where $r \in [-1, 1]$ is the degree assortativity of the graph, m is the total number of edges, and j_i and k_i are the degree of the nodes j and k that are connected through i.

Local efficiency

The definition of local efficiency from Latora and Marchiori [2001] is

$$E_i = \frac{1}{d_i(d_i - 1)} \sum_{j,k \in G_i, j \neq k} \frac{1}{\ell_{jk}} \ ,$$

where $E_i \in [0, 1]$ is the local efficiency measure of node i, d_i is the degree of node i, and ℓ_{jk} is the shortest distance between nodes j and k in the sub-graph of $\mathcal{G}_i$.

Betweenness centrality

The definition of betweenness centrality [Freeman, 1977] follows:

$$B_i = \sum_{j \neq i \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} ,$$

where $B_i \in [0, 1]$ is the betweenness centrality of node i, σ_{jk} is the total number of shortest paths between nodes j and k, and $\sigma_{jk}(i)$ is the total number of shortest paths between j and k that pass through i.

H Evaluation metrics

We evaluated the performance of our models from the classification study using the following four metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where $\quad \text{Precision} = \frac{TP}{TP + FP}$, $\text{Recall} = \frac{TP}{TP + FN}$ and ,

TP denotes True Positives; The number of correctly labeled eMCI.

TN denotes True Negatives; The number of correctly labeled control.

FP denotes False Positives; The number of control incorrectly labeled as eMCI.

FN denotes False Negatives; The number of eMCI incorrectly labeled control.

I Data, code, and computational resources statement

Due to the sensitive nature of the data used in this study, as well as the terms of use for both sources, we are unable to directly share the data used. The ADNI and Mascali et al. [2015] datasets used were derived from the following public domains `http://adni.loni.usc.edu/` and `https://dataverse.harvard.edu/dataverse/restAD`, respectively. All `R` code implementing experiments is available on Anonymous GitHub. All experiments were performed using 48 core machines with 2 Intel Platinum 8260 Cascade Lake at 2.4Ghz and 187GB of memory.

J Potential societal impacts

Our study provides evidence of more accurate and robust estimates of early neurodegeneration, which can accelerate fundamental neuroscience research, improve biomarkers for neurological and psychiatric disorders, and ultimately inform better diagnostics and therapies. However, there are potential negative impacts. For one, there is a risk of misinterpretation of models. Treating the edges of a correlational network as causal may prompt unsafe interventions. Another concern is privacy. High-resolution connectomes can, in principle, carry individual-specific signatures. Sharing or pooling data without adequate safeguards risks misuse of participants' brain data. There are also risks in using this method in unintended ways, such as outside clinical or research contexts (e.g., surveillance of cognitive states). Lastly, there are considerations regarding fairness. If the method is applied to heterogeneous populations without proper care, estimates can systematically misrepresent under-studied groups (e.g., age, ethnicity), leading to biased conclusions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we provide clear statements about background information, previous studies, and our contribution through the paper, both in the abstract as well as throughout the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide commentary on the limitations of our study in the Discussion portion of the paper. In particular, we state:

   "We acknowledge that our work has some limitations. One is that the smaller sample size ($n = 68$) may limit the ability to generalize to broader populations, particularly given the heterogeneity of eMCI. Additionally, our study uses some fixed pre-processing and parcellation choices which could influence downstream classification. Future works could explore larger studies, and jointly learn segment boundaries and within-segment representations end-to-end using deep-learning approaches."

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not consider theoretical implications of the proposed framework/method, as we are more concerned with validating the approach empirically.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a full methodological pipeline from raw data to final predictions detailed in both the main paper, as well as the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

   (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide a statement at the end of the supplementary material which details where open access data was retrieved from, and we also provide a link to an Anonymous Github repository which includes comprehensive experimental pipelines.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide full preprocessing, data splitting, and hyperparameter selection details in the paper, with additional specifics in the supplementary material. The main text summarizes the overall experimental setup, while the supplement contains exact parameter values, algorithm settings, and training/testing procedures to ensure full reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We report all hypothesis testing results, statistical significance provide error bars in figures.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the compute environment in the supplementary material. Execution time varies by experiment, but as a guideline, processing from raw data through preprocessing, FC estimation, feature generation, and classifier training/testing takes a few hours per subject. Experiments were parallelized across subjects using CPU cores.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and ensured full compliance. Anonymity is preserved by omitting author-identifying information and using an Anonymous GitHub repository for code access.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide a statement of potential societal impacts of the work available in the supplementary materials.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The authors do not believe the methodology proposed in this paper has a high risk for misuse. Pretrained model weights are not provided in any of the code.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We include appropriate citations for all code and data sets used for the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include all code and experiments as an Anonymous Github.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Although we did not collect any human data as part of our neuroimaging experiments, we did use an open source data set which includes human subjects. Details are available in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Although our real data experiments include human subjects, we were not involved in the data collection process at all, and are using the open source and de-identified version of the data set.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.