


A comprehensive benchmark of graph neural networks, graph kernels, and classical machine learning approaches on rs-fMRI brain graphs

Razan Mhanna^{1,2} 

RAZAN.MHANNA@INRIA.FR

Sophie Achard¹ 

SOPHIE.ACHARD@INRIA.FR

Alexander Petersen³

PETERSEN@STAT.BYU.EDU

Jonas Richiardi⁴ 

JONAS.RICHIARDI@CHUV.CH

¹ Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, Grenoble, France

² Univ. Grenoble Alpes, Inserm U1216, CHU Grenoble Alpes, Institut des Neurosciences, France

³ Department of Statistics, Brigham Young University, Provo, UT, USA

⁴ Lausanne University Hospital and University of Lausanne, Switzerland

Editors: Under Review for MIDL 2026

Abstract

Resting-state functional MRI (rs-fMRI) provides a powerful lens through which large-scale brain organization can be examined by modeling functional connectivity as a graph. These functional brain graphs now form the basis of machine-learning applications in neuroscience, ranging from relatively straightforward classification problems to more challenging behavioral and cognitive prediction tasks. While graph neural networks (GNNs) have gained increasing attention in neuroimaging, the absence of a unified, reproducible benchmark comparing GNNs with classical machine-learning models and graph kernel methods, across heterogeneous datasets and tasks, has made it difficult to assess their relative strengths. In this work, we introduce a comprehensive benchmarking framework spanning four heterogeneous cohorts ($N = 1513$) and multiple classification tasks, including clinical diagnosis and phenotypic prediction. We systematically evaluate classical models, graph kernels, and representative GNN architectures under a rigorous repeated nested cross-validation design and assess pairwise differences using the corrected repeated k-fold test with false-discovery-rate control. Our results show that, for this class of relatively small graphs with fixed vertex ordering, well-tuned classical ML approaches and graph kernels are competitive with GNNs, while requiring substantially fewer computational resources. For instance, the Shortest-Path graph kernel achieves 0.98 accuracy on the COMA dataset, logistic regression reaches 0.81 accuracy and 0.63 MCC on HCP sex prediction, and all model families cluster closely on multi-site datasets such as ABIDE and ADHD, where no statistically significant differences emerge. All code, seeds, cross-validation folds, fold-specific hyperparameters, full prediction logs and computational-cost measurements are publicly released at <https://gitlab.inria.fr/rmhanna/benchmark-study> to ensure full transparency and reproducibility. This benchmark provides practical guidance for model selection in rs-fMRI connectome analysis.

Keywords: Resting-state fMRI, brain networks, Graph kernels, Graph neural networks, benchmarking, reproducibility, computational efficiency.

1. Introduction

A popular approach used to investigate brain function is resting-state functional magnetic resonance imaging (rs-fMRI), a noninvasive neuroimaging technique that measures fluctuations in the blood-oxygenation-level-dependent (BOLD) signal as a correlate of brain activity. fMRI data inherently possess a complex spatio-temporal structure: each voxel (volumetric pixel) provides a time series of BOLD measurements, resulting in high-dimensional, noisy observations that inherently reside on non-Euclidean domains. To reduce the spatial complexity, voxels can be collated into user-specified regions of interest (ROI), after which aggregation is performed across voxels within the same ROI by averaging. Based on these regional signals, a functional connectome is commonly modeled as a graph, where nodes correspond to ROIs and edges represent estimated pairwise functional connectivity. Such graph-based representations have been successfully applied across a broad spectrum of network neuroscience studies, including investigations of neurodegenerative conditions such as Alzheimer’s disease and mild cognitive impairment, autism spectrum disorder, disorders of consciousness, and the prediction of psychometric, cognitive, and behavioral phenotypes in healthy individuals (Dadi et al., 2019; Di Martino et al., 2014; Li et al., 2021; Cui et al., 2022). However, despite substantial progress in neuroimaging analysis and machine learning, significant challenges persist due to the intrinsic complexity of functional connectivity data—including high dimensionality, sensitivity to parcellation schemes, variability across acquisition sites, and limited sample sizes. As a result, developing reliable computational methods that are robust, generalizable across cohorts, and reproducible remains an open and pressing problem in functional connectivity modeling.

Prior benchmarking efforts in brain functional connectivity have mainly focused on either GNNs or classical ML pipelines. BrainGB (Cui et al., 2022) evaluates message-passing GNN architectures across three functional datasets (HIV disease classification, PNC and ABCD gender prediction) and one structural dataset (PPMI Parkinson’s disease classification), using a modular design to compare a large family of GNN operators. In parallel, the benchmark of Dadi et al. (2019) systematically assessed classical ML models over 6 datasets, 8 atlases, and 3 connectivity profiles, highlighting the strong performance of tangent-space parametrization and ℓ_2 -penalized classifiers. We focus on these two benchmarks because they represent the most comprehensive and methodologically rigorous evaluations currently available: BrainGB provides the most extensive assessment of GNN architectures, whereas Dadi et al. (2019) remains the reference point for classical ML pipelines. Other studies typically investigate a single dataset, a single prediction task, or a narrowly defined model family, making broad methodological comparisons difficult. Nonetheless, these benchmarks remain limited to either GNNs or shallow ML, and do not provide a unified comparison spanning ML, graph kernels, and GNNs across heterogeneous datasets and tasks.

In this work, we develop a systematic benchmark that integrates multiple cohorts with distinct demographic and clinical profiles, including PNC, HCP, ABIDE, and a clinical COMA dataset. This multi-cohort design allows us to assess model performance across a broad range of prediction tasks: clinical case-control classification, phenotypic sex prediction, ASD diagnosis, and ADHD classification, mirroring the diversity of applications commonly encountered in network neuroscience. To the best of our knowledge, no prior

study has provided a unified comparison of classical machine-learning models, graph-kernel methods, and graph neural networks across such heterogeneous datasets and task paradigms.

Our contributions are threefold:

- Multi-dataset, multi-task evaluation: We evaluate models across four cohorts covering both healthy and clinical populations, enabling assessment on diverse classification tasks (clinical diagnosis, ASD/ADHD classification, sex prediction).
- Comprehensive comparison of modeling families: We benchmark classical ML methods (logistic regression, SVM, XGBoost), graph-kernel approaches (Shortest-Path, Weisfeiler-Lehman), and GNN architectures (GCN, GraphSAGE, GAT) under a unified and fully reproducible experimental pipeline.
- Robust evaluation strategy: We employ a repeated nested cross-validation framework in which, for each of the 50 outer folds, one fold is reserved exclusively for testing, whereas the remaining folds are further partitioned into training and validation subsets. This procedure has been demonstrated to provide more stable model rankings (Eve et al., 2025).
- Statistical significance analysis: We assess whether performance differences between models are statistically significant using the corrected repeated k-fold test of Bouckaert and Frank (2004), combined with Benjamini and Hochberg (1995) FDR correction test across pairwise comparisons.
- Practical efficiency assessment: We report computational aspects including training time, and CO_2 emissions, for CPU-operated classical machine learning and graph kernels models. This provides a clear picture of the computational footprint of lightweight methods within our benchmark.

Taken together, this benchmark offers a transparent and systematic comparison spanning classical machine-learning models, graph-kernel methods (which remain underexplored in the neuroimaging community) and more recent GNN architectures, evaluated across multiple datasets and prediction tasks. Our goal is to provide a clearer picture of the relative strengths of each modeling family in terms of accuracy, robustness, and computational efficiency.

2. Materials and Methods

2.1. Datasets

In this study, we selected publicly available resting-state fMRI datasets that offer comparable preprocessing pipelines and compatible brain parcellations, enabling methodological coherence across datasets. This choice allows us to assess the generalizability of state-of-the-art (SOTA) models under varying sample sizes, scanner characteristics, and population demographics. We first use data from the publicly available, fully de-identified Neuro Bureau ADHD-200 dataset (Bellec et al., 2017), which originally includes participants recruited across eight sites: Peking University, Bradley Hospital (Brown University), Kennedy Krieger Institute, the Donders Institute, New York University Child Study Center, Oregon

Dataset	Source	Atlas	#ROIs	Classification task	# Subjects
HCP	Dadi et al. (2019)	AAL	116	Sex classification	443
ADHD	Bellec et al. (2017)	AAL	116	ADHD prediction	160
ABIDE	Dadi et al. (2019)	AAL	116	ASD prediction	866
COMA	Oujamaa et al. (2023)	AAL	105	DoC prediction	44

Table 1: Summary of dataset statistics. ADHD = Attention Deficit Hyperactivity Disorder; ASD = Autism Spectrum Disorder; DoC = Disorders of Consciousness.

Health and Science University, the University of Pittsburgh, and Washington University in St. Louis. All cohorts received approval from their respective institutional review boards, and written informed consent was obtained from all participants or their legal guardians. Individuals had no history of psychiatric, neurological, or medical conditions other than ADHD.

We also include the ABIDE dataset, released through the Autism Brain Imaging Data Exchange initiative ([Di Martino et al., 2014](#)), which aggregates rs-fMRI acquisitions from multiple sites to study Autism Spectrum Disorder. We rely on rs-fMRI time series provided by the Preprocessed Connectome Project (PCP) ([Craddock et al., 2013](#)), as used in [Dadi et al. \(2019\)](#), and use them to perform ASD vs. control classification. Then, we incorporate data from the Human Connectome Project (HCP), which offers high-quality imaging and behavioral assessments for healthy young adults ([Van Essen et al., 2013](#)). We use the preprocessed rs-fMRI time series from the HCP900 release, also recovered from the preprocessing distributed in [Dadi et al. \(2019\)](#). This dataset enables experiments on sex classification, providing a complementary setting with substantially longer acquisitions. Finally, we use a clinical dataset of 44 subjects acquired at Grenoble Alpes University Hospital ([Oujamaa et al., 2023](#)). This cohort comprises 24 patients who had sustained acute severe traumatic brain injury; at the time of scanning, 15 had recovered consciousness, whereas 9 remained in a minimally conscious state (MCS). A control group of 20 age-matched healthy volunteers was collected under comparable acquisition conditions ([Job et al., 2020](#)). All rs-fMRI scans were obtained on the same MRI system using identical acquisition parameters and were parcellated using a 105-region modified AAL3 atlas.

A summary of all datasets, associated prediction tasks, and sample sizes is provided in Table 1.

2.2. Graph Construction

For each subject, regional BOLD time courses were collected from preprocessed datasets reported in Section 2.1. Based on these signals, pairwise Pearson correlation coefficients were computed using the ConnectivityMeasure function from the Nilearn library ([Abraham et al., 2014](#)), resulting in one weighted connectivity matrix per subject. Depending on the model requirements, two graph variants were derived: (i) a weighted matrix used for models that can handle weighted edge values, and (ii) a thresholded binary graph obtained by first extracting the minimum spanning tree (MST) to ensure connectivity, followed by retaining the top 10% of strongest remaining correlations for models that require unweighted graphs.

2.3. Benchmark Models

2.3.1. CLASSICAL ML

As baseline classifiers, we employ three widely used machine-learning algorithms: Logistic Regression, Random Forests, and XGBoost. Contrary to many prior neuroimaging studies, including the benchmark of [Dadi et al. \(2019\)](#), we perform systematic hyperparameter optimization independently for each dataset. The full search space and optimization details are reported in Section 2.5. These classical models have consistently demonstrated strong performance across multiple domains, including neuroimaging studies.

2.3.2. GRAPH KERNELS

Kernel methods provide a powerful mathematical framework for measuring similarities between structured objects, such as graphs, by implicitly mapping them from the original feature space to a (possibly infinite-dimensional) Hilbert space, where it corresponds to an inner product between transformed samples. Their main advantage lies in their computational efficiency: many kernels admit closed-form expressions for these inner products, removing the need to compute the explicit transformation. As established by Mercer’s theorem, a function qualifies as a valid kernel if it satisfies the conditions of positive semi-definiteness, ensuring it represents a legitimate inner product in some feature space. Formally, for a non-empty set χ and a function $k : \chi \times \chi \rightarrow \mathbb{R}$, there exists a Hilbert space \mathcal{H}_k and a feature map $\phi : \chi \rightarrow \mathcal{H}_k$ such that

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}_k}, \quad x, y \in \chi.$$

Once the kernel function is defined, kernel-based algorithms such as the Support Vector Machine (SVM) can be applied directly for classification or regression tasks ([Hofmann et al., 2008](#)).

In this study, we evaluate two graph kernels implemented in the GraKeL library ([Siglidis et al., 2020](#))—the Shortest-Path and Weisfeiler–Lehman kernels.

- The Shortest-Path kernel measures graph similarity by comparing the lengths and endpoint labels of all shortest paths between node pairs. For each graph, the shortest-path distances are computed, and two graphs are considered similar when their node pairs are connected through paths of comparable lengths. This kernel captures the overall topological layout of a network and is particularly effective when global path structure differentiates the graphs. It is later referred to as *GK – SP* throughout the paper.
- The Weisfeiler–Lehman kernel relies on the iterative node-label refinement procedure proposed in the WL test of graph isomorphism. At each iteration, a node’s label is updated by combining its current label with those of its neighbors, generating progressively enriched node representations. By comparing graphs across multiple refinement steps, the kernel captures hierarchical structural information and neighborhood similarity. Its efficiency and scalability make it a strong baseline for graph classification tasks. It is referred to as *GK – WL* in the rest of the paper.

2.3.3. GRAPH NEURAL NETWORKS

Graph Neural Networks (GNNs) have gained significant attention in the field of network neuroscience (Cui et al., 2022; Li et al., 2021; Comparini et al., 2026; Xu et al., 2023, 2024) due to their ability to effectively model and analyze complex graph structures. Most modern architectures can be expressed under the message passing neural network (MPNN) framework, in which node representations are iteratively updated by aggregating information from their neighborhoods. A generic MPNN layer can be written as:

$$\mathbf{h}_i^{(l+1)} = \phi\left(\mathbf{h}_i^{(l)}, \square_{j \in \mathcal{N}(i)} \psi\left(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}\right)\right),$$

where $\mathbf{h}_i^{(l)}$ is the feature vector of node i at layer l , ψ and ϕ are learnable functions, and \square denotes a permutation-invariant aggregation operator.

In this study, we focus on three representative GNN architectures widely used in many applications including brain graph analysis:

- Graph Convolutional Network (GCN): A baseline architecture that updates node representations by combining features from adjacent nodes through a predefined, normalized aggregation strategy (Jiang et al., 2019).
- Graph Attention Network (GAT): Incorporates an attention mechanism that adaptively weights neighboring nodes by learning how much each one should contribute during the feature aggregation process, allowing the model to emphasize the most informative interactions (Veličković et al., 2017).
- GraphSAGE (SAmple and aggreGatE): A sampling-based architecture that learns node embeddings by aggregating information from a fixed number of sampled neighbors, enabling inductive generalization to unseen nodes and graphs (Hamilton et al., 2017).

These architectures provide complementary mechanisms for learning from functional brain connectivity graphs and form suitable candidates for comparative evaluation in this study.

2.4. Evaluation metrics

The performance of our classification models is assessed using accuracy (ACC), balanced accuracy (BACC), and the Matthews correlation coefficient (MCC). While ACC quantifies the overall proportion of correctly classified samples, BACC accounts for class imbalance by averaging sensitivity and specificity. MCC provides a more comprehensive and balanced summary of performance by incorporating all four entries of the confusion matrix. It is defined as:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

where TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively. MCC values range from -1 to 1 , with higher values indicating stronger agreement between predicted and true labels and offering a reliable assessment of classifier performance under class imbalance.

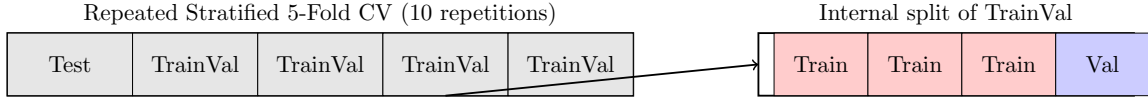


Figure 1: Repeated 5-fold nested cross-validation: Across 50 outer folds(5 folds repeated 10 times), one outer fold is held out as the test set, while the remaining four folds are internally split into three training folds and one validation fold for hyperparameter tuning.

2.5. Implementation details and hyperparameter optimization

To minimize the risk of data leakage and obtain stable performance estimates on relatively small neuroimaging datasets, all models were trained and evaluated within a repeated stratified K-fold cross-validation framework (10 repetitions, 5 folds). This choice is supported by recent empirical analyses demonstrating that repeated k-fold cross validation yields more reliable model rankings than single-split or non-repeated procedures, especially when sample sizes are limited within each outer fold (Eve et al., 2025). Hyperparameters were tuned independently using an internal stratified split of the training data, as illustrated in Figure 2.5, ensuring that the test fold remains completely unseen during both training and hyperparameter selection. The implementations were developed using PyTorch, NetworkX, PyTorch Geometric, GraKeL and CodeCarbon libraries (Fey and Lenssen, 2019; Siglidis et al., 2020; Courty et al., 2024a). All experiments were conducted on a Dell workstation running Ubuntu 22.04, equipped with an Intel Core i7 processor and an NVIDIA RTX A500 GPU.

A systematic hyperparameter search was conducted to identify the optimal configuration for the classical machine-learning models and graph kernel classifiers. The search spaces were defined as follows: for Logistic Regression, the regularization parameter was $C \in [10^{-4}, 10^4]$ (log-uniform prior); for the Random Forest classifier, $n_{\text{estimators}} \in [100, 500]$, $\text{max_depth} \in [2, 20]$, $\text{min_samples_leaf} \in [1, 10]$, and $\text{max_features} \in \{\text{"sqrt"}, \text{"log2"}\}$. For XGBoost, we optimized $\text{max_depth} \in [2, 6]$ and $\text{min_child_weight} \in [0.1, 10]$ (log-uniform prior).

Since the SP kernel produces a precomputed Gram matrix in GraKeL, only the SVM regularization parameter C is tuned. For the Weisfeiler–Lehman (WL) kernel, we optimized both $C \in [10^{-3}, 10^3]$ and the WL height $h \in [2, 4]$. Hyperparameters were optimized separately for each dataset using Bayesian optimization on an inner validation split, with balanced accuracy as the criterion to maximize. Each search was limited to 20 Bayesian optimization calls. All seeds, fold-specific hyperparameters, and full logs are provided in our public GitHub repository to ensure transparency and reproducibility.

Due to the substantially higher computational cost of training GNNs compared to classical graph kernels, we adopt fixed hyperparameter settings that follow common practice in prior neuroimaging benchmarks, including BrainGB (Cui et al., 2022) and the work of Comparini et al. (2026). This approach is standard in multi-cohort evaluations, where exhaustive hyperparameter tuning for deep models is prohibitively expensive and provides

Table 2: Classification performance across datasets. LR = Logistic Regression, RF = Random Forest, Acc = Accuracy, MCC = Matthews correlation coefficient.

Model's family	Model	COMA		HCP		ADHD		ABIDE	
		Acc	MCC	Acc	MCC	Acc	MCC	Acc	MCC
Classical ML	LR	<u>0.84±0.1</u>	<u>0.71 ±0.19</u>	0.81±0.03	0.63±0.07	0.62±0.07	<u>0.11±0.17</u>	0.63±0.03	0.26±0.06
	RF	0.8 ± 0.16	0.63±0.31	0.7±0.04	0.4±0.09	<u>0.67±0.04</u>	0.02±0.15	0.58±0.03	0.16±0.07
	XGBoost	0.71 ± 0.14	0.42 ±0.35	<u>0.73±0.03</u>	<u>0.46±0.07</u>	0.62±0.06	0.01±0.17	<u>0.6±0.03</u>	<u>0.19± 0.06</u>
Graph kernels	GK-SP	0.98±0.04	0.97±0.07	0.71±0.04	0.41±0.09	0.67±0.03	0.04±0.16	<u>0.6±0.04</u>	<u>0.19±0.08</u>
	GK-WL	0.63±0.11	0.25±0.25	0.6±0.04	0.17±0.1	0.64±0.1	0.01±0.06	0.53±0.03	0.05±0.06
GNN	GCN	0.74 ± 0.15	0.49±0.32	0.68±0.05	0.35±0.1	0.61±0.1	0.1±0.2	0.53±0.04	0.06±0.07
	GAT	0.71±0.15	0.45±0.31	0.71±0.04	0.42±0.09	0.6±0.08	<u>0.11±0.16</u>	0.56 ± 0.03	0.13± 0.06
	GraphSAGE	0.75±0.13	0.52±0.27	0.73±0.04	0.45±0.09	0.64±0.07	0.16±0.17	0.57±0.04	0.13±0.08

limited benefit for small- to medium-sized neuroimaging datasets. All GNN models were trained with the Adam optimizer and a maximum of 200 epochs, with early stopping triggered after 30 consecutive epochs without improvement in the validation loss. For the GCN architecture, we used a learning rate of 0.001 with 32 hidden channels and three graph convolutional layers. The GAT model was configured with 16 hidden channels, four attention heads, and two layers, while GraphSAGE employed 32 hidden channels and three layers. All GNNs shared the same learning rate, number of epochs, optimizer, and early-stopping criterion

3. Results

3.1. Classification results

Table 2 summarizes the classification performance across all datasets in terms of accuracy (Acc) and Matthews correlation coefficient (MCC), while Figure 2 reports the corresponding balanced accuracy averaged over repeated cross-validation splits, with error bars indicating the standard deviation. Overall, we observe that classical machine learning methods, graph kernels, and graph neural networks achieve broadly comparable performance across cohorts. Importantly, graph kernels—particularly the Shortest-Path kernel (GK-SP)—achieve some of the highest performance levels, most notably on the COMA dataset, where GK-SP attains an accuracy of approximately 0.98 and an MCC of 0.97 (Table 2). On the ADHD dataset, GK-SP also yields competitive accuracy (0.67), with lower MCC values, reflecting the difficulty of the task, likely due to the heterogeneity introduced by multi-site data acquisition. In addition, logistic regression achieves strong performance on both the HCP (Acc 0.81, MCC 0.63) and ABIDE (Acc 0.63, MCC 0.26) datasets, highlighting the effectiveness of simple linear models in these settings. In contrast, graph neural network models, including GCN, GAT, and GraphSAGE, generally exhibit similar or slightly lower mean performance across datasets and cross-validation splits. This behaviour is also reflected in the balanced accuracy results shown in Figure 2, where GNNs typically match or underperform classical machine learning and graph kernel approaches.

The COMA dataset exhibits the clearest separation between model families, with graph kernels achieving the highest accuracy and MCC values, highlighting their ability to capture global alterations in functional brain connectivity associated with disorders of consciousness. In contrast, graph neural networks yield lower average performance in this setting,

suggesting limitations in learning stable and discriminative representations from very small clinical datasets. As shown in Figure 2, balanced accuracy values are high overall but are accompanied by increased variability across cross-validation folds. This variability is likely related to the inter-patient heterogeneity within the COMA cohort, where patterns of functional alteration can differ substantially across individuals, leading to sensitivity to the specific composition of training and test folds.

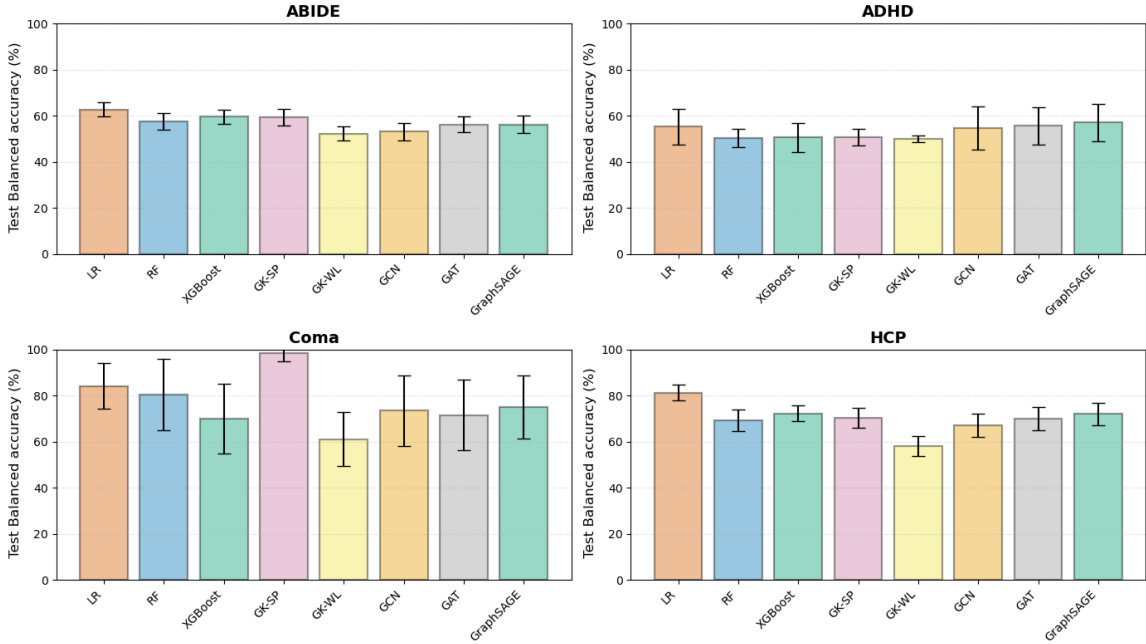


Figure 2: Repeated cross-validation balanced accuracy. The height of each bar corresponds to the mean balanced accuracy across all folds and repetitions, with error bars representing the standard deviation.

The HCP dataset exhibits the most homogeneous performance across models, with several approaches achieving similar accuracy values. Notably, logistic regression attains strong performance, comparable to or exceeding that of more complex graph-based models. This suggests that the high quality, consistency, and relatively low noise of the HCP data allow even simple linear models to capture discriminative information effectively. For comparison with prior work, the NeuroGraph benchmark (Said et al., 2023) reports an accuracy of 69.9% for sex classification on the HCP Young Adult S1200 dataset (1,078 subjects) using a Random Forest classifier, which is broadly consistent with the performance of classical machine-learning baselines observed in our study. In contrast, graph neural networks in NeuroGraph achieve higher accuracy, with reported values of 75.46% for GCN, 77.69% for GraphSAGE, and 76.20% for GAT. However, these results are obtained under different experimental conditions, namely using static functional connectivity graphs constructed with the Schaefer atlas at 1000 ROIs and a single stratified 70/10/20 train-validation-test split, which limits direct comparability with our repeated cross-validation setting.

For the ADHD dataset, all methods exhibit relatively modest performance, with substantial overlap in balanced accuracy across model families, as shown in Figure 2. Graph kernels, particularly the Shortest-Path kernel (GK-SP), achieve some of the highest accuracy values (0.67 in Table 2). In terms of MCC, the best values are also attained by classical approaches, although MCC remains low overall across all models, reflecting the limited class separability in this dataset. This pattern reflects the subtle and heterogeneous nature of ADHD-related functional connectivity alterations, as well as the variability introduced by multi-site data collection. In this context, increased model complexity does not provide a clear advantage, and performance differences between approaches remain limited and often not statistically significant. A similar performance pattern is observed on the ABIDE dataset, where the absence of clear gains from graph neural networks suggests that the high inter-site heterogeneity of ABIDE, combined with limited sample sizes per site, may hinder the learning of robust graph-level representations. In contrast, simpler models appear less sensitive to this variability, resulting in comparable or even superior performance. While careful hyperparameter tuning can modestly improve results, it does not fundamentally alter the observed patterns. In comparison with prior work, the ContrastPool study (Xu et al., 2024) reports, on the ABIDE dataset (989 subjects) evaluated using 10-fold cross-validation, an accuracy of $65.82 \pm 3.51\%$ for logistic regression and $61.18 \pm 5.01\%$ for random forest, values that are consistent with the performance of classical pipelines observed in our experiments. General-purpose GNNs achieve comparable or slightly lower accuracy, including GCN ($60.97 \pm 2.84\%$), GAT ($60.87 \pm 5.02\%$), and GraphSAGE ($63.09 \pm 3.11\%$). Importantly, GNNs in that study benefit from explicit hyperparameter tuning via grid search, ensuring a fair comparison across model families.

3.2. Statistical comparison of models

To assess whether observed performance differences between models were statistically significant, we followed the recommendations of Bouckaert and Frank (2004), who showed that standard significance tests applied to cross-validation results can be overly optimistic due to dependencies between folds. We therefore employed the corrected repeated k-fold cross-validation test. For each dataset and each pair of models, predictions collected on identical test splits across all folds and repetitions were compared using a loss function derived from the evaluation metric. Statistical significance was then assessed using a two-sided Student’s t-test on the mean loss difference. Furthermore, to account for multiple pairwise comparisons, Benjamini–Hochberg false discovery rate (FDR) correction was applied within each dataset (Benjamini and Hochberg, 1995). Raw p-values from the corrected repeated k-fold test are shown in the upper triangular part of each matrix, while FDR-adjusted p-values are reported in the lower triangular part of Figure 3. Starred cells indicate comparisons that remain statistically significant after FDR correction.

Across datasets, the statistical analysis largely supports the conclusion that most models achieve comparable performance, with limited evidence of robust pairwise differences. In particular, the ADHD dataset shows no statistically significant differences between models, as reflected by uniformly large p-values in both the raw and FDR-adjusted matrices. This lack of statistical separation is likely due to several factors, including the limited sample size retained from the ADHD-200 initiative (160 subjects in our study) and the multi-

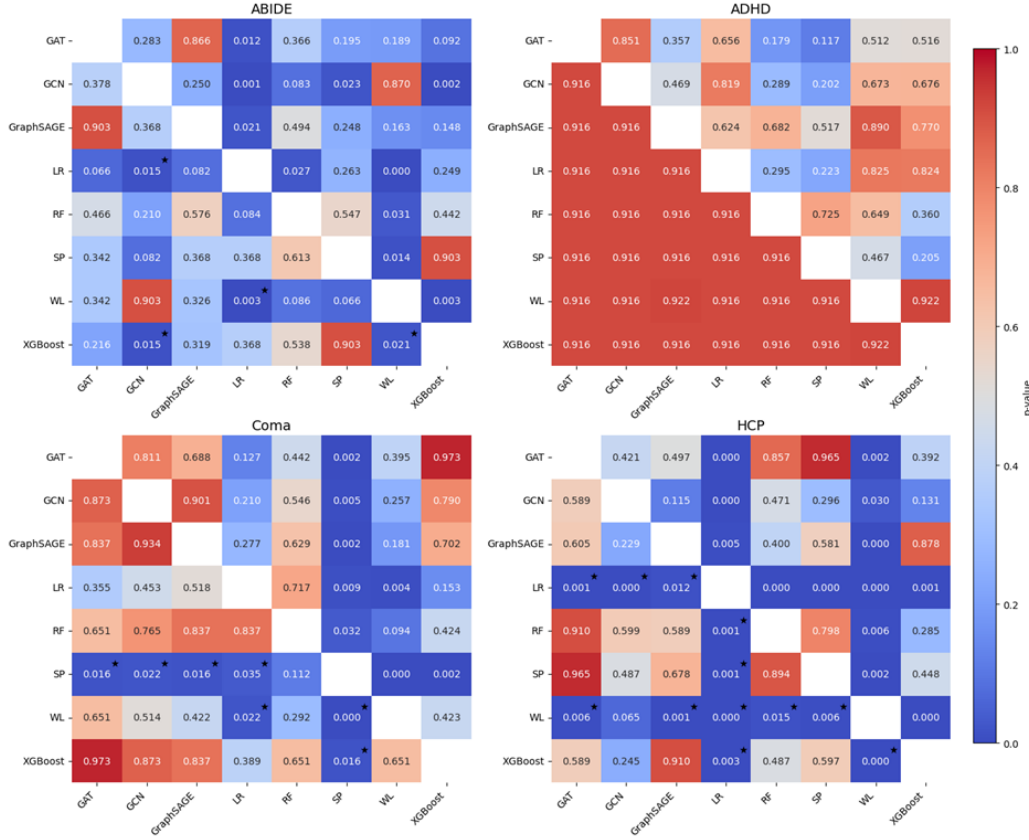


Figure 3: Pairwise statistical comparison of models across all datasets (ABIDE, ADHD, Coma, and HCP) based on accuracy. Each cell reports the p-value obtained from the corrected repeated k-fold cross-validation test (Bouckaert and Frank, 2004). Raw p-values are displayed in the upper triangular part of each matrix. The lower triangular part displays p-values after Benjamini–Hochberg false discovery rate (FDR) correction applied across all pairwise model comparisons within each dataset. Diagonal entries correspond to self-comparisons. Cells marked with a star indicate comparisons that remain statistically significant after FDR correction at level $\alpha = 0.05$.

site heterogeneity, which results in small effective sample sizes per site. A similar pattern is observed for the ABIDE dataset, where only a small number of isolated pairwise comparisons remain significant after correction. As shown in Table 2 and Figure 2, most performance differences across models in ABIDE are small and statistically fragile, indicating substantial overlap between model families once cross-validation dependence and multiple testing are properly accounted for.

For the HCP dataset, the statistical comparisons reveal more consistent differences between model families than in ABIDE or ADHD. Logistic regression achieves the strongest performance, and several comparisons between LR and GNNs yield very small p-values

Table 3: Total training time (sec) and CO₂ emissions (kg) aggregated across the full cross-validation protocol (all folds and repetitions).

Model	ABIDE		ADHD		COMA		HCP	
	Time (s)	CO ₂ (kg)	Time (s)	CO ₂ (kg)	Time (s)	CO ₂ (kg)	Time (s)	CO ₂ (kg)
LR	15 850.97	1.26×10^{-2}	1 499.40	5.09×10^{-4}	369.74	1.51×10^{-4}	8 202.85	6.76×10^{-3}
RF	7 916.05	1.27×10^{-3}	1 395.58	2.14×10^{-4}	386.46	5.95×10^{-5}	4 467.13	6.80×10^{-4}
XGBoost	450 497.07	2.94×10^{-1}	5 210.19	8.32×10^{-4}	559.06	1.14×10^{-5}	47 518.22	5.89×10^{-3}
GK-SP	4 537.77	6.97×10^{-4}	810.98	1.22×10^{-4}	228.32	3.47×10^{-5}	2 332.29	3.56×10^{-4}
GK-WL	9 429.97	1.38×10^{-3}	1 012.94	1.61×10^{-4}	230.97	3.49×10^{-5}	2 317.81	3.58×10^{-4}

in the upper triangular matrix (e.g., LR vs. GAT and LR vs. GCN with raw p-values 0.000), with some remaining significant after FDR correction. In addition, the Weisfeiler–Lehman (WL) graph kernel shows statistically significant differences when compared to most other models; however, these differences correspond to inferior performance, as WL consistently achieves the lowest accuracy and MCC values across models (Table 2). This indicates that, while WL captures graph representations that differ significantly from those learned by other approaches, these representations are less aligned with sex-related discriminative patterns in the HCP dataset. Overall, these results suggest that in a high-quality, homogeneous cohort such as HCP, simple linear models are sufficient to capture relevant population-level effects, whereas increased structural expressiveness—as in WL kernels or GNNs—does not necessarily translate into improved predictive performance.

In contrast, for the COMA dataset, the Shortest-Path graph kernel exhibits the strongest and most statistically consistent differences when compared with other model families, (e.g., raw p-values as low as $p < 0.02$ for SP versus GraphSAGE or GAT, with comparisons remaining significant after FDR correction at $p < 0.05$), while differences among the remaining models are less pronounced. These results indicate that shortest-path-based features capture discriminative network patterns that are not effectively exploited by other approaches. This finding is consistent with the neurobiological characteristics of disorders of consciousness: as shown by Achard et al. (2012), comatose patients exhibit a pronounced reorganization of hub structure and path-length-related properties, despite a largely preserved global network topology.

3.3. Computational and environmental costs

Classical machine-learning models and graph kernel methods were executed on the CPU of a Dell workstation running Ubuntu 22.04 and equipped with an Intel Core i7 processor, whereas graph neural network (GNN) models were trained using an NVIDIA RTX A500 GPU on the same machine. Because GPU acceleration constitutes a fundamentally different computational regime, raw wall-clock runtimes obtained from CPU-based and GPU-based implementations should not be interpreted as direct algorithmic comparisons. Moreover, GPU-accelerated training typically entails higher power consumption, potentially leading to increased environmental costs. Accordingly, we report in Table 3 the computational runtime and CO₂ emissions exclusively for classical machine-learning and graph kernel methods; CO₂ emissions were estimated using CodeCarbon (Courty et al., 2024b).

4. Conclusion

In this work, we conducted a comprehensive benchmarking study of classical machine-learning methods, graph kernel approaches, and graph neural networks (GNNs) on four resting-state fMRI brain graphs across multiple classification tasks. By evaluating predictive performance, statistical significance, and computational cost within a unified experimental framework, we provide a principled assessment of the practical trade-offs between model families. Across the evaluated datasets, well-tuned classical machine-learning methods and graph kernels achieve performance comparable to GNN architectures. In particular, for small clinical datasets, classical approaches such as logistic regression, random forests, and kernel-based methods exhibit strong predictive accuracy while incurring substantially lower computational and environmental costs. Although GNNs are often presented as the dominant paradigm for graph-structured neuroimaging data, our results indicate that their empirical advantages over classical baselines are limited and strongly task-dependent. Several limitations should be acknowledged. First, our analysis is restricted to a set of relatively small, publicly available datasets, and the findings may not directly generalize to large-scale graphs or multimodal settings. Second, given the higher computational demands of GNNs, we relied on standard hyperparameter configurations commonly used in prior neuroimaging studies. More exhaustive hyperparameter tuning may lead to slight performance improvements. Overall, our findings question the assumption that increased model complexity necessarily leads to superior performance in brain graph analysis. In line with recent calls for responsible and reproducible machine learning, we emphasize the importance of transparent reporting, fair baselines, and cost-aware evaluation in future applications of GNNs to neuroimaging.

Acknowledgments

This work was supported by the Agence Nationale de la Recherche under the France 2030 programme, reference ANR-23-IACL-0006.

References

- Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.
- Sophie Achard, Chantal Delon-Martin, Petra E Vértes, Félix Renard, Maleka Schenck, Francis Schneider, Christian Heinrich, Stéphane Kremer, and Edward T Bullmore. Hubs of brain functional networks are radically reorganized in comatose patients. *Proceedings of the National Academy of Sciences*, 109(50):20608–20613, 2012.
- Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The neuro bureau adhd-200 preprocessed repository. *Neuroimage*, 144:275–286, 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)*, 57(1):289–300, 1995.
- Remco R Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 3–12. Springer, 2004.
- Alessio Comparini, Léa Schmidt, Vanessa Siffredi, Damien Marie, Clara James, and Jonas Richiardi. Late and early fusion graph neural network architectures for integrative modeling of multimodal brain connectivity graphs. In *Proceedings of the 22nd International Workshop on Mining and Learning with Graphs*, Porto, Portugal, 2026.
- Benoit Courty, Victor Schmidt, Boris Feld, Jérémy Lecourt, Mathilde Léval, Luis Blanche, Alexis Cruveiller, Franklin Zhao, Aditya Joshi, Alexis Bogroff, et al. mlco2/codecarbon: v2. 4.1. *Zenodo*, 2024a.
- Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, Marc Alencon, Michał Stechly, Christian Bauer, Lucas Otávio N. de Araújo, JPW, and MinervaBooks. mlco2/codecarbon: v2.4.1, May 2024b. URL <https://doi.org/10.5281/zenodo.11171501>.
- Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li,

- Michael Milham, et al. The neuro bureau preprocessing initiative: open sharing of pre-processed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7(27):5, 2013.
- Hejie Cui, Wei Dai, Yanqiao Zhu, Xuan Kan, Antonio Aodong Chen Gu, Joshua Lukemire, Liang Zhan, Lifang He, Ying Guo, and Carl Yang. Braingb: a benchmark for brain network analysis with graph neural networks. *IEEE transactions on medical imaging*, 42(2):493–506, 2022.
- Kamalaker Dadi, Mehdi Rahim, Alexandre Abraham, Darya Chyzhyk, Michael Milham, Bertrand Thirion, Gaël Varoquaux, Alzheimer’s Disease Neuroimaging Initiative, et al. Benchmarking functional connectome-based predictive models for resting-state fmri. *NeuroImage*, 192:115–134, 2019.
- Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.
- Célestin Eve, Thomas Moreau, and Gaël Varoquaux. De l’importance de la validation croisée. In *Actes du XXXè Colloque francophone de traitement du signal et des images (GRETSI 2025)*, pages 753–756, Strasbourg, France, 2025.
- Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
- Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. 2008.
- Bo Jiang, Ziyang Zhang, Doudou Lin, Jin Tang, and Bin Luo. Semi-supervised learning with graph learning-convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11313–11320, 2019.
- Agnès Job, Chloé Jaroszynski, Anne Kavounoudias, Assia Jaillard, and Chantal Delon-Martin. Functional connectivity in chronic nonbothersome tinnitus following acoustic trauma: a seed-based resting-state functional magnetic resonance imaging study. *Brain connectivity*, 10(6):279–291, 2020.
- Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74: 102233, 2021.
- Lydia Oujamaa, Chantal Delon-Martin, Chloé Jaroszynski, Maite Termenon, Stein Silva, Jean-François Payen, and Sophie Achard. Functional hub disruption emphasizes consciousness recovery in severe traumatic brain injury. *Brain Communications*, 5(6):fcad319, 2023.

- Anwar Said, Roza Bayrak, Tyler Derr, Mudassir Shabbir, Daniel Moyer, Catie Chang, and Xenofon Koutsoukos. Neurograph: Benchmarks for graph machine learning in brain connectomics. *Advances in Neural Information Processing Systems*, 36:6509–6531, 2023.
- Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, Konstantinos Skianis, and Michalis Vazirgiannis. Grakel: A graph kernel library in python. *Journal of Machine Learning Research*, 21(54):1–5, 2020.
- David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- Jiaxing Xu, Yunhan Yang, David Huang, Sophi Shilpa Gururajapathy, Yiping Ke, Miao Qiao, Alan Wang, Haribalan Kumar, Josh McGeown, and Eryn Kwon. Data-driven network neuroscience: On data collection and benchmark. *Advances in Neural Information Processing Systems*, 36:21841–21856, 2023.
- Jiaxing Xu, Qingtian Bian, Xinhang Li, Aihu Zhang, Yiping Ke, Miao Qiao, Wei Zhang, Wei Khang Jeremy Sim, and Balázs Gulyás. Contrastive graph pooling for explainable classification of brain networks. *IEEE Transactions on Medical Imaging*, 43(9):3292–3305, 2024.