UAB
Universitat Autònoma
de Barcelona

Dipòsit digital
de documents
de la UAB

This is the **published version** of the bachelor thesis:

Canal Fort, Ramon. *Advancing Graph Neural Networks for Brain Connectivity Analysis in Multiple Sclerosis*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337317

under the terms of the  license.

Advancing Graph Neural Networks for Brain Connectivity Analysis in Multiple Sclerosis

Ramon Canal Fort

Resum– Aquest treball presenta un pipeline complet de Xarxes Neuronals de Graf (GNN) per a l'anàlisi de la connectivitat cerebral multimodal en Esclerosi Múltiple (EM), ampliant la recerca prèvia de Subirà-Cribillers et al. (2025). A partir d'un conjunt de dades de 270 participants (71 controls sans, 199 pacients amb EM) de dues cohorts, es construeixen grafs de 76 nodes per modalitat (DTI, morfologia de matèria grisa, rs-fMRI). L'aportació central és l'arquitectura EdgeTypeAware, que aplica convolucions independents per modalitat per evitar que la major densitat d'arestes de la rs-fMRI domini l'agregació. Un estudi ablatiu de 28 condicions identifica la configuració óptima (top-20% arestes, tres *features* topològiques: strength, betweenness i local efficiency). La millor arquitectura assoleix AUC-ROC = 0.839 ± 0.054 en classificació binària HC (*healthy controll*) vs. EM. S'avaluen addicionalment la classificació de subtipus (F1-macro = 0.474), la predicció de *Expanded Disability Status Scale* (EDSS) per regressió (MAE = 1.31, ICC = 0.33) i l'explicabilitat per gradient saliency, que identifica l'eficiència local com la feature més discriminativa.

Paraules clau– Xarxes Neuronals de Graf, Esclerosi Múltiple, connectivitat cerebral multimodal, EdgeTypeAware, gradient saliency, classificació binària, predicció EDSS.

Abstract– This work presents a complete Graph Neural Network (GNN) pipeline for multimodal brain connectivity analysis in Multiple Sclerosis (MS), extending prior research by Subirà-Cribillers et al. (2025). Using a dataset of 270 participants (71 healthy controls, 199 MS patients) from two cohorts, graphs with 76 nodes per modality (DTI, grey matter morphology, rs-fMRI) are constructed. The central contribution is the EdgeTypeAware architecture, which applies independent convolutions per modality to prevent the higher edge density of rs-fMRI from dominating aggregation. A systematic 28-condition ablation identifies the optimal configuration (top-20% edges, three topological node features: strength, betweenness and local efficiency). The best architecture achieves AUC-ROC = 0.839 ± 0.054 in binary HC (Healthy Control) vs. MS classification. MS subtype classification (F1-macro = 0.474), Expanded Disability Status Scale (EDSS) regression (MAE = 1.31, ICC = 0.33), and gradient saliency explainability are additionally evaluated; local efficiency is the most discriminative feature, with frontal, temporal, and cingulate regions most salient.

Keywords– Graph Neural Networks, Multiple Sclerosis, multimodal brain connectivity, EdgeTypeAware, gradient saliency, binary classification, EDSS prediction.

1 Context

Multiple Sclerosis (MS) is a chronic inflammatory and neurodegenerative disease of the central nervous system affecting millions worldwide. It leads to progressive physical and cognitive disability and is characterised by demyelinating lesions detectable via brain MRI. Clinical management relies heavily on accurate diagnosis, monitoring of disease progression, and treatment evaluation.

Brain connectivity networks derived from multimodal MRI (including diffusion tensor imaging (DTI), grey matter morphology (GM), and resting-state functional MRI (rs-fMRI)) provide a rich representation of brain organisation that is altered in MS. These networks can be naturally modelled as graphs, where nodes represent anatomical brain regions and edges encode the strength of their connections.

Recent work by Subirà-Cribillers et al. [1] demonstrated the potential of Graph Neural Networks (GNNs) for MS classification using multimodal brain connectivity data, achieving an AUC-ROC of up to 0.82 with a DTI-based Graph Attention Network. However, key limitations remain: the DTI-only architecture showed high sensitivity to hyperparameter choices, and multimodal architectures did not consistently outperform single-modality ones. Furthermore, MS subtype classification and disability score (EDSS) prediction remain insufficiently explored with GNN-based approaches. The complementary analyses

• Contact e-mail: 1674310@uab.cat
• Work supervised by: Jordi Casas-Roma (Departament de Ciències de la Computació, UAB)
• Curs 2025/26

of Casas-Roma & Lozano-Bagen [2] and Lozano-Bagen et al. [3] further consolidated the multimodal framework used here as a direct baseline.

This TFG builds directly on that prior work, aiming to advance the state of the art by addressing the modality density imbalance in the multimodal graph, systematically ablating graph construction parameters, comparing multiple GNN families, exploring EDSS regression, and adding gradient-based explainability analysis.

2 OBJECTIVES

The main objective of this TFG is to develop an improved GNN-based pipeline for brain connectivity analysis in MS, surpassing the performance and robustness of existing models. The specific objectives are:

- **O1.** Evaluate and compare novel GNN architectures (GraphSAGE, GATv2, and Graph Transformers) against the existing GCN and GAT baselines.
- **O2.** Improve model robustness and generalisation through regularisation strategies and systematic hyperparameter optimisation.
- **O3.** Address the MS subtype classification task (HC; RRMS, Relapsing-Remitting Multiple Sclerosis; SPMS, Secondary Progressive Multiple Sclerosis; and PPMS, Primary Progressive Multiple Sclerosis) with improved performance over the current baseline.
- **O4.** Explore EDSS prediction as a regression task using GNN-derived graph embeddings.
- **O5.** Investigate more informative graph construction strategies, including alternative node feature definitions and edge weighting schemes.

The extent to which O1 and O3 were ultimately addressed differs from this initial scope; these deviations, and their justification, are discussed explicitly in the Conclusions.

3 METHODOLOGY

The project adopts a **cascade (waterfall) methodology**, structured in well-defined sequential phases where each phase produces a concrete deliverable before the next begins. This choice is appropriate given the academic timeline with fixed submission deadlines, and the fact that each phase depends heavily on the outputs of the previous one.

Within each phase, a lightweight iterative approach is applied: individual components are implemented, tested, and refined in short cycles. This hybrid structure balances the predictability of a cascade plan with the flexibility needed for experimental research.

Phase 1 covers the literature review, dataset characterisation, and the full data loading and preprocessing pipeline.

Phase 2 reproduces the existing GCN/GAT baseline and explores new GNN architectures under identical experimental conditions, using 5-fold stratified cross-validation and AUC-ROC as the primary metric.

Phase 3 pursues systematic ablation over graph construction parameters and node features, and extends the pipeline to clinically relevant tasks beyond binary HC vs. people with Multiple Sclerosis (pwMS) classification.

Phase 4 consolidates all results, addresses EDSS regression and explainability, and produces this final report.

Phase 5 is reserved for oral defence preparation and final dossier submission.

Data governance and ethics. The connectivity matrices and clinical variables (MS subtype, EDSS, age, sex) used in this work were provided by the clinical partner cohorts in fully anonymised form: patients are identified only by numeric codes, with no names, dates of birth, or other direct identifiers present in any file accessed during this project. Acquisition of the underlying MRI data and collection of clinical scores were carried out by the originating hospitals under their own institutional ethics approval and informed-consent procedures, as reported in the prior work this TFG builds on [1, 2]. No additional ethics approval was required for the present analysis, which is limited to the secondary processing of this pre-existing, anonymised connectivity and clinical data.

4 STATE OF THE ART

4.1 GNN Architectures

GNNs are deep learning models designed to learn from graph-structured data via *message passing*: each node iteratively aggregates information from its neighbours, producing progressively richer representations. Five core architectures define this design space; of these, GCN, GAT, GATv2 and GraphSAGE are evaluated empirically in this work (Section 8), while GIN and Graph Transformers are reviewed below for context but were not implemented, for the reasons discussed in the Conclusions.

GCN [5] applies spectral convolutions via a normalised aggregation of neighbour features ($\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$). It is simple and effective on medium-sized graphs, but prone to oversmoothing with many layers.

GAT [6] introduces self-attention to weight each neighbour's contribution dynamically. Multi-head attention improves stability but increases computational cost and data requirements.

GATv2 [14] corrects a theoretical limitation of GAT by computing attention dynamically rather than statically, yielding strictly greater expressivity.

GraphSAGE [7] learns aggregation functions over sampled neighbourhoods, enabling inductive inference suitable for settings where the model must generalise across patient cohorts.

GIN [8] maximises discriminative power by achieving theoretical equivalence to the Weisfeiler-Lehman graph isomorphism test via sum aggregation.

Graph Transformers [9] adapt the transformer mechanism to graphs and are the most expressive but also the most data-hungry architectures.

4.2 GNNs in Neuroimaging

GNNs have found increasing application in brain network analysis. The three main MRI modalities are rs-fMRI (functional connectivity via BOLD signal correlations), dMRI/DTI (structural connectivity via tractography), and

EEG/MEG (sensor-level connectivity). In all cases, nodes correspond to brain regions defined by parcellation atlases (e.g., AAL, Desikan-Killiany) and edges encode connectivity strength after thresholding.

Representative studies include Wein et al. [15], who proposed a diffusion convolutional RNN combining structural and functional connectivity for fMRI dynamics prediction; Sujitha et al. [16], who combined GCN, GAT and GraphSAGE in an ensemble for major depression classification (70–78% accuracy); and Ktena et al. [17], who applied a spectral GNN for autism biomarker discovery via a population-level graph. Common challenges include small cohort sizes, inter-subject variability, graph construction parameter sensitivity, and limited model interpretability.

4.3 Brain Connectivity in MS

Graph-theoretic studies of MS consistently reveal network reconfiguration. Kocevar et al. [10] found higher assortativity, longer path lengths, and reduced global efficiency in the structural DTI networks of pwMS, achieving F1 $> 70\%$ on subtype classification with an SVM. Chen et al. [12] applied GCN to T1-derived morphological graphs from 91 pwMS, achieving F1 scores of 72–78% for subtype discrimination, outperforming 3D-CNNs. The most directly relevant work is that of Casas-Roma et al. [1, 2, 3], which integrates multimodal connectivity (DTI, GM, fMRI) with GNNs for MS classification, constituting the direct baseline for this TFG. An earlier multilayer-network analysis by the same group [4] had already identified complementary disease-related alterations across morphological, structural and functional connectivity layers without a GNN-based approach, providing further motivation for the multimodal graph design adopted here.

4.4 MS Subtype Classification and EDSS Prediction

For MS subtype classification, Marzullo et al. [11] demonstrated that GCN applied to weighted structural DTI networks achieves strong 5-class classification, with weighted matrices consistently outperforming binarised ones. For EDSS prediction, Valsasina et al. [13] trained tree-based models on volumetric MRI features from 1,554 pwMS, reporting ICC of 0.56–0.76. GNN-specific approaches to EDSS regression remain sparse in the literature, representing an open opportunity for this TFG.

5 Dataset Description

5.1 Clinical Characteristics

The dataset comprises 270 participants from two clinical cohorts: BCN ($n = 191$) and NAP ($n = 79$). Of these, 71 are healthy controls (HC) and 199 are people with MS (pwMS). The distribution of MS subtypes is highly imbalanced, reflecting natural disease prevalence. As shown in Table 1, the majority of participants have Relapsing-Remitting MS (RRMS, $n = 155$, 57.4%), followed by HC ($n = 71$, 26.3%), Secondary Progressive MS (SPMS, $n = 31$, 11.5%), and Primary Progressive MS (PPMS, $n = 13$, 4.8%). The two cohorts have markedly different compositions: BCN is dominated by RRMS (60.7%), while NAP has a much higher proportion of HC (31.6%). This heterogeneity adds complexity to model generalisation and motivates the use of stratified cross-validation.

TABLE 1: Class distribution by MS subtype and cohort.

Subtype	BCN	NAP	Total	%
RRMS	116	39	155	57.4
HC	46	25	71	26.3
SPMS	21	10	31	11.5
PPMS	8	5	13	4.8
Total	**191**	**79**	**270**	100

The EDSS ranges from 0 to 7.5 across all participants, with a right-skewed distribution reflecting the inclusion of HC (EDSS = 0) and early-stage patients. Restricted to the 198 pwMS with a valid EDSS measurement, the range is 1.0–7.5 (mean 3.12 ± 1.87).

5.2 Brain Graph Structure

Each participant is represented by three 76×76 adjacency matrices. The 76 nodes correspond to brain regions defined by the Desikan-Killiany cortical parcellation atlas (68 cortical + 8 subcortical). Fig. 1 illustrates the structure of the three matrices; Table 2 summarises raw graph statistics per modality.

TABLE 2: Graph statistics per modality (before thresholding). Density > 1.0 in RS-fMRI arises from self-loops in the raw matrices, removed during preprocessing.

Modality	Density (mean $\pm$ std)
DTI	0.668 ± 0.062
GM (morph.)	0.847 ± 0.142
RS (fMRI)	1.027 ± 0.000

The RS-fMRI graphs are almost fully connected (density ≈ 1.0), reflecting the inclusion of negative functional correlations taken as absolute values. DTI graphs are considerably sparser (density ≈ 0.67), consistent with the biological constraint that only physically connected white matter tracts contribute to structural connectivity.

5.3 Edge Thresholding

Dense, noisy graphs hinder GNN learning by saturating degree-based normalisation. Table 3 shows the effect of retaining only the top $k\%$ of edges by connectivity weight. A top-20% threshold (mean density ≈ 0.14, avg. degree ≈ 10.4) represents a biologically plausible and computationally tractable starting point, consistent with prior work [10, 1]. The sensitivity of model performance to this parameter is investigated systematically in Section 8. Table 3 uses a coarse 5–40% sweep to characterise how den-

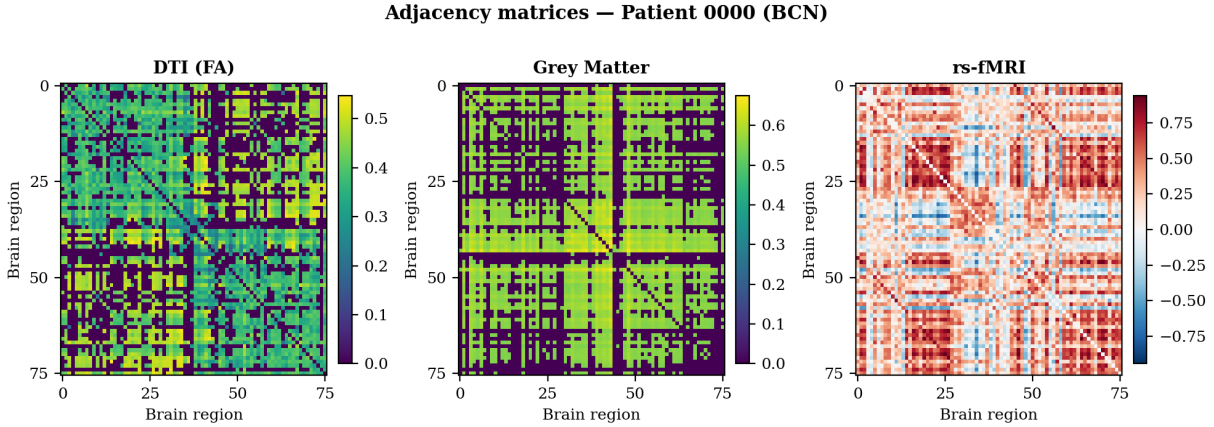

Fig. 1: Adjacency matrices for the three MRI modalities of a real BCN patient (ID 0000), raw values before thresholding. DTI and GM use a sequential colour scale; RS-fMRI uses a diverging scale since functional correlations can be negative. Each cell (i, j) encodes connectivity strength between brain regions i and j.

sity scales with the threshold in general; the ablation in Section 8 instead uses a finer 10–25% sweep centred on the 20% value identified here as the most plausible starting point.

TABLE 3: EFFECT OF EDGE THRESHOLDING ON GRAPH DENSITY AND AVERAGE NODE DEGREE.

Top %	Density (mean)	Avg. degree
5%	0.035	2.6
10%	0.069	5.2
20%	0.138	10.4
30%	0.208	15.6
40%	0.277	20.7

6 PIPELINE AND EDGETYPEAWARE DESIGN

6.1 Data Loading and Graph Construction

The preprocessing pipeline reads clinical metadata from CSV files for each cohort, then loads the three adjacency matrices per participant. Each matrix is cleaned before graph construction: self-connections are removed and all values are converted to their absolute value, since functional MRI correlations can be negative but the relevant quantity for edge weighting is the magnitude of the connection. Edge thresholding is then applied per modality, retaining the top $k\%$ of connections by weight.

6.2 Node Feature Computation

Five graph-theoretic features are computed per node per modality:

- **Degree**: number of connections.
- **Strength**: total weighted connectivity.
- **Betweenness centrality**: $b(v) = \sum_{s \neq v \neq t} \sigma_{st}(v)/\sigma_{st}$, a proxy for inter-regional hub role.
- **Closeness centrality**: $C(v) = \dfrac{N-1}{\sum_{u \neq v} d(v,u)}$, the inverse of the mean shortest-path distance from v to every other node.
- **Local efficiency**: $E_{\text{loc}}(v) = \frac{1}{|\mathcal{N}_v|} \sum_{j \in \mathcal{N}_v} \frac{1}{d(i,j)}\big|_{G[\mathcal{N}_v]}$, the global efficiency of the neighbourhood subgraph of v.

All features are normalised to $[0, 1]$. In multimodal mode they are computed independently per modality and concatenated, yielding a 15-dimensional feature vector per node.

6.3 Multimodal Graph Fusion

The three modality-specific graphs are merged into a single **concatenated multimodal graph**. Edges from each modality are pooled together, and each edge carries a type attribute (0 = DTI, 1 = GM, 2 = RS) allowing the model to distinguish the source of each connection. Node features are concatenated across modalities. Fig. 2 illustrates this structure with a 10-node example; a real patient graph with all 76 nodes is included in the Appendix.

6.4 EdgeTypeAware Architecture

The key problem with naive multimodal concatenation is *density imbalance*: after top-20% thresholding, rs-fMRI edges account for the majority of all edges in the unified graph, since rs-fMRI matrices have near-unit density before thresholding. Standard spectral convolution normalises by node degree, so the degree-based signal is dominated by the rs-fMRI subgraph and the gradients from the biologically informative but sparse DTI and GM subgraphs are effectively diluted.

The **EdgeTypeAware** design solves this by applying an independent graph convolution to each modality subgraph within every layer, then fusing representations in latent space:

$$\mathbf{h}^{(l+1)} = \text{BN}\Big(\mathbf{W}_{\text{proj}}\Big[\mathbf{h}_{\text{DTI}}^{(l+1)} \,\|\, \mathbf{h}_{\text{GM}}^{(l+1)} \,\|\, \mathbf{h}_{\text{RS}}^{(l+1)}\Big]\Big) \quad (1)$$

where $\mathbf{h}_m^{(l+1)} = \text{ReLU}(\text{GConv}_m(\mathbf{h}^{(l)}, \mathcal{E}_m))$ and BN denotes batch normalisation. Each modality uses its own de-

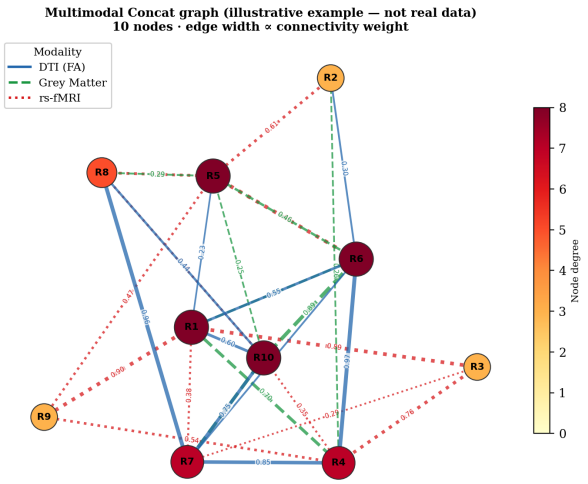

Fig. 2: Illustrative multimodal graph (10 nodes, not real data). Edge colour and style encode modality (blue solid = DTI, green dashed = GM, red dotted = rs-fMRI); edge width is proportional to connectivity weight; node colour encodes total degree. See Appendix for a real patient graph.

gree normalisation, so RS cannot suppress DTI or GM. The learned linear projection lets the model discover the optimal relative weighting of each modality for the task at hand. Four variants were implemented, replacing the per-modality convolution with GCN [5], GAT [6], GraphSAGE [7], and GATv2 [14] operators respectively. A standard GCN treating all edges as equivalent serves as a lower-bound baseline.

7 Experimental Setup

All experiments follow a 5-fold stratified cross-validation protocol with a fixed random seed, ensuring consistent splits across all conditions. AUC-ROC is the primary evaluation metric, as it is robust to class imbalance (HC:pwMS ratio of approximately 1:2.8). Secondary metrics include sensitivity, specificity, F1-macro, AUC-PR, and the Matthews Correlation Coefficient (MCC), a balanced measure even on highly imbalanced classes, computed from the confusion matrix as

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

which ranges from -1 (total disagreement) to $+1$ (perfect prediction), with 0 corresponding to random guessing. Unless stated otherwise, $\pm$ values reported throughout this work denote the standard deviation across the 5 CV folds.

For binary tasks, threshold calibration is applied per fold: rather than using a fixed decision boundary of 0.5, the threshold that maximises balanced accuracy on the validation set is selected. Because this threshold is tuned on the same fold used for evaluation, the resulting sensitivity and specificity values are optimistic and should be read as upper bounds rather than out-of-sample estimates; AUC-ROC, which is threshold-independent, is therefore used as the primary metric throughout. For regression, metrics are MAE, RMSE, Pearson r, and ICC(A,1) computed post-hoc on out-of-fold predictions.

Class imbalance is addressed by computing per-class inverse-frequency weights passed to CrossEntropyLoss. Early stopping (patience of 20 epochs) is applied based on the primary validation metric.

8 Results

8.1 Standard GCN Baseline vs. EdgeTypeAware

The natural first step was to establish a GCN baseline on the multimodal graph treating all edges, regardless of their modality of origin, as equivalent. With a default configuration (2 layers, hidden dimension 64, dropout 0.5, lr = 1e-3) and a fixed threshold of 0.5, this baseline achieves AUC-ROC = 0.767 ± 0.104. Inspecting the graph structure explains the mediocre result: rs-fMRI edges dominate the unified graph because rs-fMRI matrices have near-unit density before thresholding, and degree-based normalisation amplifies this imbalance, diluting the DTI and GM signal.

Switching to the EdgeTypeAware design under the same configuration achieves AUC-ROC = 0.809 ± 0.082 ($+0.042$). Applying per-fold threshold calibration further improves both models, with the EdgeTypeAware reaching MCC = 0.568 ± 0.079 and F1-macro = 0.772 ± 0.034—a dramatic reduction in variance compared to the standard baseline (F1 std = 0.117), indicating that the EdgeTypeAware architecture learns more consistently across folds.

TABLE 4: Standard GCN vs. EdgeTypeAware GCN, with and without threshold calibration. Se. denotes sensitivity and Sp. denotes specificity.

Model	AUC	MCC	Se.	Sp.
Std. GCN, thr = 0.5	0.767	0.328	0.626	0.714
Std. GCN, calibrated	0.770	0.469	0.675	0.832
EdgeAware, thr = 0.5	0.809	0.327	0.514	0.818
EdgeAware, calibrated	**0.821**	**0.568**	**0.809**	0.803

8.2 Hyperparameter Optimisation

A 30-trial Bayesian search (TPE sampler, median pruning, 5-fold CV with 50 epochs per trial) explored hidden dimensions {32,64,128}, number of layers {2,3}, dropout in [0.3,0.6], learning rate in [1e-4, 1e-2], weight decay in [1e-5, 1e-3], and batch size {8,16,32}. The best configuration achieves AUC-ROC = 0.831 ($+0.010$ over the default), confirming the default was already well-dimensioned. The most informative finding is that 3 message-passing layers appear in 21 of 30 trials, suggesting that three propagation hops better capture multi-hop connectivity disruptions in MS. The optimal settings are: hidden = 64, layers = 3, dropout = 0.378, lr = 2.11×10^{-3}, weight_decay = 4.20×10^{-5}, batch = 16. These are fixed for all subsequent experiments.

Note on experimental ordering. The edge-density ablation reported next (Phase A) was carried out earlier in the project, under the default configuration described above

(Table 4), prior to this hyperparameter search. Its top-20% result (AUC-ROC = 0.822) therefore reflects the default hyperparameters rather than the tuned ones, and is consequently slightly below the values obtained afterwards with the tuned configuration (Table 6). The 20% threshold itself remained the best choice once tuning was applied: the corresponding row of Table 6 (min3, DTI+GM+RS, top-20%, AUC-ROC = 0.833) is the best result from the Phase B ablation sweep. A small further gap remains between this value and the 0.839 reported in Table 7 for the nominally identical configuration. The fixed random seed used throughout this work controls the 5-fold cross-validation split, which is therefore identical across Tables 5, 6 and 7; it does not separately fix the network weight initialisation at the start of each individual training run. The Phase B sweep trains this configuration as the 1 of 28 conditions executed sequentially within one script, whereas the architecture comparison of Section 8.4 trains it as a standalone run; in both cases the resulting weight initialisation differs, which is sufficient to produce a 0.006 AUC-ROC shift at this sample size. The 0.839 figure is reported as the final result because it was obtained under the same standalone-run conditions as the other three architectures compared in Table 7, making it the directly comparable value; the 0.833 figure remains the reference used to rank the 28 Phase B configurations in Table 6.

8.3 Ablation Study

Phase A: Edge Density. Table 5 shows that performance increases monotonically from 10% to 20%, then drops at 25%: sparser graphs carry too little connectivity information, while edges above 20% introduce noise that degrades the degree-normalisation signal. Top-20% achieves the best AUC-ROC (0.822) and MCC (0.566) with the lowest fold variance (AUC std = 0.058). This value is fixed for all subsequent experiments.

TABLE 5: PHASE A: EFFECT OF EDGE DENSITY THRESHOLD.

top-%	AUC-ROC	MCC	Sens.	Spec.
10%	0.766±0.074	0.487	0.759	0.772
15%	0.808±0.090	0.532	0.694	0.900
20%	**0.822±0.058**	**0.566**	0.698	0.914
25%	0.800±0.069	0.516	0.728	0.830

Phase B: Feature × Modality. A full factorial experiment crossed four feature subsets (*all5*: degree, strength, betweenness, closeness, local efficiency; *no_degree*; *topologic*: betweenness, closeness, local efficiency; *min3*: strength, betweenness, local efficiency) with seven modality combinations ($2^3 - 1 = 7$), yielding 28 conditions. The top results are in Table 6; the complete table is in Appendix B.

Three conclusions emerge. First, the full trimodal combination (DTI+GM+RS) consistently outperforms all bimodal and unimodal alternatives: removing any modality hurts performance. GM alone is the weakest (AUC ≈ 0.62) but contributes complementary information in the trimodal setting. Second, the feature set matters less than the modality choice: all four subsets perform similarly when all three modalities are included. Third, **min3** is selected as the reference: it ties with no_degree in AUC-ROC (0.833) but achieves the highest MCC of all 28 conditions (0.616) and is more parsimonious. Its neurological justification is solid: strength captures overall weighted connectivity, betweenness identifies inter-regional bridging hubs whose disruption is documented in MS [10], and local efficiency quantifies local information integration, known to be reduced in MS networks.

TABLE 6: PHASE B: TOP-5 RESULTS FROM THE 28-CONDITION ABLATION.

Features	Mod.	AUC-ROC	MCC
no_degree	DTI+GM+RS	0.833±0.075	0.561
min3	**DTI+GM+RS**	**0.833±0.086**	**0.616**
all5	DTI+GM+RS	0.832±0.058	0.551
min3	DTI+GM	0.821±0.103	0.532
topologic	DTI+GM+RS	0.819±0.065	0.568

8.4 Architecture Comparison

All four EdgeTypeAware variants were evaluated under the optimal data configuration (Table 7). GAT and GraphSAGE were initially tested with the GCN-optimised hyperparameters (acknowledged as slightly favouring GCN). A dedicated 25-trial search for GraphSAGE confirmed the ranking: tuned GraphSAGE (AUC = 0.806) still falls below EdgeTypeAware GCN (0.839). A 2×2 factorial experiment additionally tested GATv2 and two community-detection-based features (participation coefficient, clustering coefficient): neither addition improved results. GATv2's dynamic attention requires more data than $n = 270$ provides; the stochastic community detection (Louvain algorithm) adds noise that outweighs any added signal at this sample size.

TABLE 7: ARCHITECTURE COMPARISON, BINARY HC VS. MS (MIN3, DTI+GM+RS, TOP-20%).

Model	AUC	MCC	Se.	Sp.
Standard GCN	0.770	0.469	0.675	0.832
EdgeAware GCN	**0.839**	**0.595**	0.799	0.846
EdgeAware GAT	0.787	0.488	0.734	0.803
EdgeAware SAGE	0.799	0.568	0.718	0.871
SAGE (tuned)	0.806	0.572	0.779	0.790
GATv2 + min3	0.808	0.543	0.769	0.830
GATv2 + ext. feat.	0.798	0.535	0.754	0.817

EdgeTypeAware GCN leads on all primary metrics with the lowest fold variance (AUC std = 0.054). The GAT result is instructive: attention is theoretically appealing (not all inter-regional connections should contribute equally) but with $n = 270$ the attention coefficients lack sufficient diversity to converge to meaningful weights. GraphSAGE shows competitive MCC (0.568–0.572) but higher sensitivity variance (std = 0.223). Notably, Fold 2 is consistently the hard-

est across all models (AUC-ROC 0.60–0.64), suggesting its participant distribution poses a challenge independent of architecture. From this point, EdgeTypeAware GCN with min3 features, top-20% threshold, and all three modalities is the consolidated reference model.

8.5 MS Subtype Classification

Two subtype tasks were evaluated excluding HC (Table 8).

For *3-class subtype classification* (RRMS vs. SPMS vs. PPMS), F1-macro reaches 0.474 ± 0.076. RRMS is classified acceptably (F1 = 0.791), which is expected given its dominance in training data. PPMS performance is poor and highly variable (F1 = 0.276 ± 0.196): with only $n = 13$ samples and a RRMS:PPMS ratio of approximately 12:1, some validation folds contain too few PPMS samples to evaluate reliably. A dedicated hyperparameter search for this task confirmed the ceiling of F1-macro ≈ 0.47 is a data-size limitation, not a configuration one (tuned model: 0.442 ± 0.033).

For *RRMS vs. Progressive MS* (SPMS+PPMS grouped), the minority class rises from $n = 13$ to $n = 44$ (ratio 3.5:1). This grouping is medically principled: both progressive subtypes share continuous neurological deterioration without remissions, in contrast to RRMS. AUC-ROC improves to 0.710 and F1-macro to 0.627. The high sensitivity (0.836) shows the model consistently identifies the progressive pattern; the moderate specificity (0.607) indicates that some RRMS patients with more advanced connectivity disruption are misclassified as progressive.

TABLE 8: COMPARISON ACROSS ALL CLASSIFICATION TASKS (EDGETYPEAWARE GCN, OPTIMAL CONFIG.).

Task	AUC-ROC	MCC	F1
HC vs. MS	**0.839±0.054**	**0.595**	**0.747**
RRMS vs. Prog.	0.710±0.100	0.395	0.627
3-class	—	0.223	0.474

The gradient in performance across tasks is itself a scientific finding: HC vs. MS is tractable because the connectivity gap between healthy and diseased brains is large and multidimensional; RRMS vs. Progressive is harder because both groups are already MS patients; and 3-class discrimination is hardest because MS subtypes differ primarily in temporal progression profile; a dimension that static cross-sectional connectivity matrices cannot capture directly.

8.6 EDSS Regression

EDSS prediction was evaluated on pwMS only ($n = 198$ with valid EDSS; range 1.0–7.5, mean 3.12 ± 1.87), excluding HC to avoid the trivial EDSS = 0 signal. A regression variant of the EdgeTypeAware architecture was used, identical to the classifier but with a single scalar output unit and mean squared error training loss. A non-stratified 5-fold split was used; ICC(A,1) was computed post-hoc on all 198 out-of-fold predictions.

The MAE of 1.31 EDSS points is below the overall dataset standard deviation (1.87), indicating that the model captures more variance than a constant-output predictor. However, because the EDSS distribution is right-skewed, the standard deviation overstates the error of a constant-mean baseline (for skewed distributions the mean absolute deviation of the mean is typically smaller than the standard deviation), so this comparison should be read as indicative rather than as a precise improvement-over-baseline figure. Pearson $r = 0.39$ is statistically significant ($p < 0.001$). The ICC(A,1) = 0.329 falls in the "fair" range by Koo & Li [19] criteria (cutoffs: <0.50 poor/fair; 0.50–0.75 moderate; >0.75 good), compared to ICC = 0.56–0.76 reported by Valsasina et al. [13] using volumetric MRI and lesion features. The gap reflects that brain volumes and lesion load carry more direct disability signal than topological graph metrics. The significant ICC ($p < 0.001$) nevertheless confirms genuine predictive information in multimodal connectivity. Fold 0 is again the most challenging (MAE = 1.66, $r = 0.27$), consistent with the pattern observed across classification experiments.

TABLE 9: EDSS REGRESSION: PER-FOLD AND OOF GLOBAL RESULTS. DATASET SD = 1.87.

	MAE	RMSE	r	ICC
Fold 0	1.664	2.185	0.274	—
Fold 1	1.271	1.699	0.516	—
Fold 2	1.307	1.761	0.452	—
Fold 3	1.121	1.746	0.349	—
Fold 4	1.180	1.771	0.365	—
Mean±SD	1.31±0.19	1.83±0.18	0.391	—
OOF global	1.310	1.842	0.392	**0.329**

8.7 Explainability: Gradient Saliency

To identify which brain regions drive the HC vs. MS decision, gradient saliency was computed over each fold's validation patients. For each patient graph, the gradient of the logit difference ($z_{\text{MS}} - z_{\text{HC}}$) was backpropagated through node features:

$$s(v) = \left\| \frac{\partial (z_{\text{MS}} - z_{\text{HC}})}{\partial \mathbf{x}_v} \right\|_2$$

The logit difference was chosen over $P(\text{MS})$ for more discriminative gradients: the $P(\text{MS})$ variant yielded compressed distributions (all values 0.08–0.12) due to gradient saturation from batch normalisation layers.

Important caveat: saliency reflects the model's sensitivity to perturbations of graph-theoretic node features, not a direct measure of neurobiological alteration. High saliency means the topological metrics of a region are discriminative for the model's decision. These results correspond to a single training run per fold; multi-seed replication was not performed.

Table 10 shows the top-15 regions by global saliency. MS values exceed HC values in virtually all regions (positive Diff), with the top-15 distributed across frontal (5), temporal (5), occipital (2), cingulate (1), insula (1), and subcortical (1) lobes.

TABLE 10: TOP-15 BRAIN REGIONS BY GLOBAL SALIENCY (MEAN OVER 5 FOLDS). DIFF = MS − HC.

Region	HC	MS	Diff
L-superiorfrontal	0.746	1.136	+0.390
R-lateraloccipital	0.927	1.062	+0.136
L-Pallidum	0.739	1.118	+0.379
R-frontalpole	0.751	1.110	+0.360
R-parahippocampal	0.874	1.062	+0.187
R-transversetemporal	0.701	1.116	+0.415
L-pericalcarine	0.768	1.078	+0.310
L-posteriorcingulate	0.716	1.050	+0.333
R-insula	0.678	1.034	+0.356
L-parahippocampal	0.656	1.031	+0.375
R-lateralorbitofrontal	0.757	0.986	+0.230
L-superiortemporal	0.653	1.020	+0.367
R-rostralmiddlefrontal	0.726	0.988	+0.262
R-middletemporal	0.666	1.009	+0.343
L-caudalmiddlefrontal	0.646	1.008	+0.362

A feature-level decomposition reveals that **local efficiency** produces the largest MS–HC saliency differences in absolute terms (up to $+0.21$), consistently exceeding strength and betweenness centrality across all regions. This makes local efficiency the most discriminative topological feature in the model's decisions and is the most robust finding of the explainability analysis, independent of the interpretation of specific regions. The left middle frontal cortex is the only region appearing simultaneously in the top-10 for all three features.

Among regions with prior literature support: the posterior cingulate cortex and insula are central nodes of the Default Mode Network and salience network respectively, with altered functional connectivity documented in MS [10]; the pallidum is a basal ganglia hub with structural alterations in progressive MS [10]; and the parahippocampal gyrus is linked to cognitive deficits in MS. The right transverse temporal cortex shows the largest overall MS–HC difference ($+0.415$) but is not classically prominent in MS connectivity literature and should be treated as exploratory pending replication.

9 DISCUSSION

The pattern of results across architectures and tasks is consistent with dataset size ($n = 270$) being a major constraint on what additional model capacity can deliver. The attention-based variants do not surpass the EdgeTypeAware GCN despite their additional expressivity (EdgeAware GAT: AUC-ROC = 0.787; GATv2 + min3: 0.808; vs. 0.839 for GCN), the 3-class subtype task collapses for PPMS (F1 = 0.276 ± 0.196 with $n = 13$), and the EDSS regression ICC (0.329) remains below the 0.56–0.76 range reported with larger, feature-richer cohorts [13]. A learning-curve analysis varying the training-set size, proposed in the Future Work below, would be needed to confirm this causally.

The one architectural choice that helped consistently across tasks was separating modalities prior to fusion: it underlies the best result in the binary task, in both subtype tasks, and in the regression head, which reuses the same per-modality design unchanged. Since the rs-fMRI subgraph is denser than DTI or GM for every patient (Table 2), this density imbalance is plausibly a structural property of this type of multimodal connectome graph rather than a dataset-specific artefact, though testing this on other multimodal connectivity datasets would be required to generalise the claim.

The overall HC vs. MS result (AUC-ROC = 0.839 ± 0.054) compares favourably with the DTI-only GAT baseline of Subirà-Cribillers et al. [1] (AUC-ROC ≈ 0.82) and shows lower fold-to-fold variance, while also extending to subtype and EDSS tasks not previously addressed with this dataset. Two methodological caveats apply to this and all other binary results. First, as already noted, per-fold threshold calibration is performed on the same validation fold used for evaluation, so the reported sensitivity and specificity values are optimistic upper bounds rather than out-of-sample estimates. Second, the architecture and graph-construction configuration reported as the final result (Table 7) was itself selected by comparing cross-validation performance on the same 270 patients and the same fold split used to obtain that final number. This is a form of model-selection bias common in small-sample studies: the 0.839 figure is best read as the strongest result found for this dataset and split, rather than as an unbiased estimate of performance on entirely unseen data.

10 CONCLUSIONS

This TFG presents a complete GNN pipeline for multimodal brain connectivity analysis in Multiple Sclerosis, covering all phases from dataset characterisation and pipeline construction to systematic optimisation, clinical tasks, and explainability. The main contributions are: (1) the EdgeTypeAware architecture, which addresses the modality density imbalance and consistently outperforms standard GCN across all experimental conditions; (2) a 28-condition ablation establishing optimal graph construction parameters (top-20%, min3 features, all three modalities); (3) a multi-architecture comparison confirming GCN as the most robust model for this dataset size; (4) first results on EDSS regression (ICC = 0.33) and subtype classification (F1-macro = 0.474 for 3-class; AUC-ROC = 0.710 for RRMS vs. Progressive) using this multimodal graph representation; and (5) a gradient saliency analysis identifying local efficiency as the primary discriminative feature and frontal, temporal, cingulate, and insular regions as the most model-salient.

Two aspects of the originally planned scope were not completed as initially stated. Regarding **O1**, GraphSAGE and GATv2 were implemented and compared against the GCN and GAT baselines (Section 8), but GIN and Graph Transformers, included in the initial scope and reviewed in Section 4.1, were not implemented. The decision to redirect effort was informed by the GATv2 result itself: if a strictly more expressive attention mechanism than GAT (GATv2) still failed to improve on GCN at $n = 270$, it was judged unlikely that the even more data-hungry GIN and Graph Transformer architectures would yield informative results, and the remaining time was directed towards the ablation study and the new clinical tasks (O3,

O4) instead. Regarding **O3**, the originally planned 4-class HC/RRMS/SPMS/PPMS classification was replaced by two complementary tasks excluding HC: a 3-class RRMS/SPMS/PPMS classification and a clinically motivated RRMS-vs.-Progressive binary task. This change was made because HC vs. pwMS discrimination is already covered by the primary binary task, and because grouping SPMS and PPMS into a single Progressive class ($n = 44$) was the only way to obtain a workable minority-class size given the extreme rarity of PPMS ($n = 13$) on its own.

The key overarching finding is that HC vs. MS classification is tractable at AUC-ROC ≈ 0.84 with this pipeline, while subtype discrimination is fundamentally harder due to class imbalance and the subtlety of intra-MS connectivity differences. EDSS prediction from connectivity alone shows genuine signal but is limited by the absence of volumetric and lesion features.

Future work could extend this pipeline in several concrete directions: incorporating volumetric MRI and lesion-load features into the regression head, which Valsasina et al. [13] show carry more direct disability signal than topological metrics alone; applying GNNExplainer [18] for edge-level (rather than node-level) interpretability, to identify which inter-regional connections drive the learned per-modality fusion weights; running a learning-curve analysis over training set size to test directly whether the data-size limitation discussed in Section 9 explains the gap between attention-based and convolutional variants; and implementing GIN and Graph Transformer variants within the EdgeTypeAware framework now that the graph-construction configuration is fixed, completing the architecture comparison originally proposed in O1.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Jordi Casas-Roma and the research group at the Computer Vision Center (CVC) and UAB for their continued guidance and for providing access to the dataset and existing codebase.

REFERENCES

[1] Subirà-Cribillers, M. et al. (2025). GNNs for Multimodal Brain Connectivity Analysis in MS. *GbRPR 2025*, LNCS 15727. Springer. https://doi.org/10.1007/978-3-031-94139-9_9

[2] Casas-Roma, J., Lozano-Bagen, T. (2026). Comparing GNNs for Single and Multi-layer Brain Connectivity Analysis in MS. *MDAI 2025*, LNCS 15957. Springer. https://doi.org/10.1007/978-3-032-00891-6_27

[3] Lozano-Bagen, T. et al. (2025). Evaluating topological and graph-theoretical approaches to extract complex multimodal brain connectivity patterns in MS. *Health Inf. Sci. Syst.*, 13:68.

[4] Casas-Roma, J. et al. (2022). Applying multilayer analysis to morphological, structural and functional brain networks. *Network Neuroscience*, 6(3):916–933.

[5] Kipf, T.N., Welling, M. (2017). Semi-Supervised Classification with GCNs. *ICLR 2017*.

[6] Veličković, P. et al. (2018). Graph Attention Networks. *ICLR 2018*.

[7] Hamilton, W., Ying, Z., Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. *NeurIPS 2017*.

[8] Xu, K. et al. (2019). How Powerful are Graph Neural Networks? *ICLR 2019*.

[9] Yun, S. et al. (2019). Graph Transformer Networks. *NeurIPS 2019*.

[10] Kocevar, G. et al. (2016). Graph Theory-Based Brain Connectivity for Automatic Classification of MS Clinical Courses. *Front. Neuroscience*, 10:478.

[11] Marzullo, A. et al. (2019). Classification of single and multiple sclerosis gait alterations using machine learning. *Front. Neuroscience*.

[12] Chen, X. et al. (2023). GNN-based morphological connectivity for MS subtype classification. *Front. Neuroscience*.

[13] Valsasina, P. et al. (2026). Machine learning and MRI-derived brain volumes for MS diagnosis and EDSS prediction. *Front. Artificial Intelligence*.

[14] Brody, S., Alon, U., Yahav, E. (2022). How Attentive are Graph Attention Networks? *ICLR 2022*.

[15] Wein, S. et al. (2021). A GNN framework for causal inference in brain networks. *Scientific Reports*, 11:8061.

[16] Sujitha, R. et al. (2023). Ensemble GNN model for MDD classification from resting-state fMRI. *Front. Psychiatry*.

[17] Ktena, S.I. et al. (2017). Distance Metric Learning using GCNs: Application to Functional Brain Networks. *MICCAI 2017*.

[18] Ying, R. et al. (2019). GNNExplainer: Generating Explanations for GNNs. *NeurIPS 2019*.

[19] Koo, T.K., Li, M.Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163.

APPENDIX

A.1 Real Patient Graph Example

Fig. 3 shows the full multimodal graph (76 nodes, Desikan-Killiany parcellation) for the same real BCN patient (ID 0000) whose adjacency matrices are shown in Fig. 1, after retaining the top-20% of edges per modality. Compared with the dense raw matrices, the thresholded graph is considerably sparser, and the three modalities contribute clearly distinct, complementary edge sets.

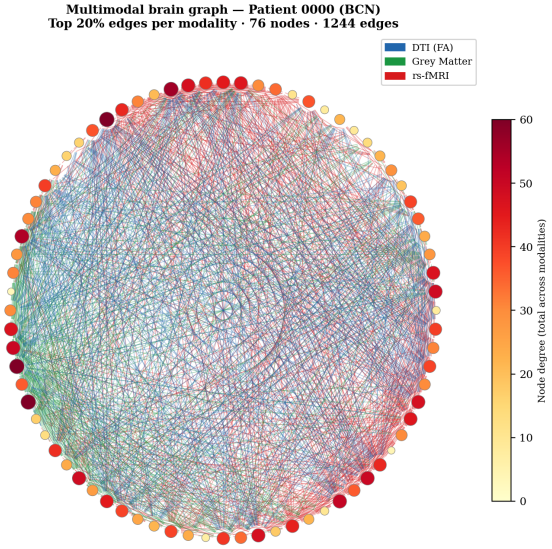

Fig. 3: Multimodal graph for the real BCN patient (ID 0000), top-20% edges retained per modality. The 76 nodes correspond to brain regions from the Desikan-Killiany atlas. Edge colour and line-style conventions as in Fig. 2.

A.2 Complete Phase B Ablation Results

Table 11 reports all 28 conditions of the Phase B feature × modality ablation, ordered by AUC-ROC.

TABLE 11: FULL PHASE B ABLATION (28 CONDITIONS, 5-FOLD CV, TOP-20%).

Features	Mod.	AUC-ROC	MCC
no_degree	DTI+GM+RS	0.833±0.075	0.561±0.122
min3	DTI+GM+RS	0.833±0.086	0.616±0.124
all5	DTI+GM+RS	0.832±0.058	0.551±0.106
min3	DTI+GM	0.821±0.103	0.532±0.142
topologic	DTI+GM+RS	0.819±0.065	0.568±0.099
all5	DTI+GM	0.812±0.103	0.553±0.128
no_degree	DTI+GM	0.810±0.115	0.555±0.148
min3	DTI+RS	0.782±0.081	0.495±0.113
topologic	DTI+GM	0.781±0.090	0.480±0.113
all5	DTI+RS	0.778±0.069	0.506±0.060
no_degree	DTI+RS	0.772±0.069	0.506±0.013
topologic	DTI+RS	0.766±0.060	0.488±0.040
all5	GM+RS	0.757±0.070	0.516±0.104
no_degree	DTI	0.751±0.104	0.467±0.154
topologic	RS	0.751±0.051	0.518±0.100
min3	DTI	0.748±0.124	0.465±0.174
all5	DTI	0.743±0.114	0.442±0.172
no_degree	GM+RS	0.742±0.023	0.508±0.069
topologic	DTI	0.739±0.103	0.420±0.155
topologic	GM+RS	0.732±0.041	0.459±0.060
no_degree	RS	0.725±0.050	0.434±0.081
all5	RS	0.722±0.066	0.463±0.080
min3	GM+RS	0.699±0.045	0.392±0.056
min3	RS	0.694±0.064	0.407±0.062
no_degree	GM	0.629±0.060	0.318±0.088
min3	GM	0.623±0.072	0.285±0.083
all5	GM	0.615±0.060	0.291±0.109
topologic	GM	0.595±0.068	0.276±0.107