UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

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

Gonzalez Armenteras, Miquel. *Synthetic Data Generation for Brain Connectivity Graphs : Augmentation Techniques and Topological Evaluation.* 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/337338

under the terms of the  license.

Synthetic Data Generation for Brain Connectivity Graphs: Augmentation Techniques and Topological Evaluation

Miquel González Armenteras

Abstract– Multiple sclerosis (MS) is a chronic inflammatory and neurodegenerative disease that leads to severe physical and cognitive disability. While machine learning algorithms have demonstrated significant potential in identifying network dysfunction patterns to discriminate between MS patients and healthy subjects, the development of robust, clinical-grade models is heavily bottlenecked by the scarcity of high-quality neuroimaging data. The application of sophisticated data augmentation techniques is therefore critical to enhance MS diagnostic tools. This research investigates generative deep learning models to synthesize brain connectivity graphs. By establishing a rigorous topological and statistical evaluation framework, this project contributes a validated augmentation pipeline that mitigates overfitting, improves model generalization, and ultimately supports the deployment of reliable clinical applications.

Index Terms– Data augmentation, brain networks, graph generative adversarial networks (GAN), multiple sclerosis, topological data analysis, machine learning, validation methods, mathematical methods

Resum– L'esclerosi múltiple (EM) és una malaltia inflamatòria crònica i neurodegenerativa que provoca una discapacitat física i cognitiva greu. Tot i que els algorismes d'aprenentatge automàtic han demostrat un potencial significatiu a l'hora d'identificar patrons de disfunció de la xarxa per discriminar entre pacients amb EM i subjectes sans, el desenvolupament de models robustos i de grau clínic està fortament limitat per l'escassetat de dades de neuroimatge d'alta qualitat. Per això, l'aplicació de tècniques sofisticades d'augment de dades és fonamental per millorar les eines de diagnòstic de l'EM. Aquesta recerca investiga models generatius d'aprenentatge profund per sintetitzar grafs de connectivitat cerebral. En establir un marc d'avaluació topològic i estadístic rigorós, aquest projecte aporta un pipeline d'augment validat que mitiga el sobreajust, millora la generalització del model i, en darrer terme, dóna suport al desplegament d'aplicacions clíniques fiables.

Paraules clau– Augment de dades, xarxes cerebrals, xarxes generatives adversàries de grafs (GAN), esclerosi múltiple, anàlisi topològica de dades, aprenentatge automàtic, mètodes de validació, mètodes matemàtics

1 INTRODUCTION

THE application of artificial intelligence in neuroimaging has opened new avenues for understanding and diagnosing neurodegenerative diseases such as multiple sclerosis (MS). Recently, the representation of the brain as a complex network where anatomical regions act as nodes and the structural or functional connections between them act as edges has proven highly effective. These brain connectivity graphs capture the intricate topology of the human brain, allowing advanced models to identify subtle dysfunction patterns and discriminate between patients and healthy subjects [1], [2].

However, the development of robust predictive models in the medical domain is severely hindered by data scarcity. Acquiring magnetic resonance imaging (MRI) is extremely difficult to get access to for a plethora of reasons, rang-

• Contact: miquel.gonzaleza@autonoma.cat
• Tutor: Jordi Casas Roma (Departament de Ciències de la Computació)
• Class of 2025/2026

ing from there being a low amount of healthy volunteers to compare with ill patients or data privacy between institutions. In traditional machine learning, this lack of data is often mitigated through data augmentation. Standard techniques used in vision and language have no direct analogs for complex, non-Euclidean graph structures [3]. Medical data is extremely sensitive. Randomly adding or removing edges in a brain network can easily destroy its underlying biological and topological properties. This work contributes both a novel framework for brain graph augmentation and quantitative criteria for evaluating the fidelity of synthetic data.

To address this challenge, this project investigates synthetic data generation for brain connectivity graphs using generative models. Given that graph generation differs fundamentally from pixel-based synthesis, a critical component of this research is the rigorous evaluation of the synthesized data. Therefore, this work not only explores augmentation techniques but also defines a comprehensive mathematical and empirical evaluation framework to ensure that the synthetic brain networks retain the statistical and topological fidelity necessary to train robust classification models.

2 Objectives

The goal of this project is to research, implement, and validate synthetic data augmentation techniques for medical graphs, specifically targeting brain connectivity matrices. To achieve this, the project is divided into four main specific objectives:

O1. Exploration of Generative Models for Graphs: Investigate and implement generative architectures (e.g., simple GANs adapted for adjacent matrices or edge prediction methods) capable of generating synthetic non-Euclidean graph data.

O2. Design of a Topological Evaluation Framework: Define and implement objective mathematical metrics to assess the biological realism of the generated graphs. This includes comparing network properties such as local efficiency, node strength, and degree distributions between the real MS datasets and the synthetic ones.

O3. Empirical Validation: Evaluate the utility of the augmented data empirically by training a baseline neural network (or GNN) for patient classification. The objective is to quantify the performance improvement and robustness gained when combining real and different synthetically generated brain networks using different techniques.

O4. Analysis of Synthetic Viability: Determine why certain generative techniques succeed or fail in preserving brain topology.

3 Methodology

This project utilizes brain connectivity graphs (adjacency matrices) extracted from MRI and DTI scans, based on previous clinical studies [1], [2]. The original dataset is limited, comprising exactly 199 people with MS (pwMS) and 71 healthy volunteers (HV). The core methodological strategy of this project is to synthetically multiply this original cohort using generative models.

The development is conducted in Python. Deep learning architectures, including the generative models (e.g., GANs) and the baseline Graph Neural Networks (GNNs) used for empirical validation, are implemented using PyTorch and PyTorch Geometric (PyG). To assess the biological realism of the synthesized matrices (Objective 2), the NetworkX library is employed to compute and compare topological metrics such as local efficiency and node strength.

The project follows an iterative pipeline in which an initial baseline pipeline is built and validated, followed by a second iteration that optimizes the generative process using advanced models to produce higher-quality, realistic synthetic data.

4 Planning

To ensure the robustness of the synthetic data and the validity of the topological metrics, the work is organised as an iterative pipeline: a first iteration builds the whole pipeline and produces initial validation results, and a second iteration returns to the generative step to optimise it and obtain better results. The project is divided into four phases built around these two iterations: (1) **Research and Data Preparation**, covering the literature review, data exploration and data wrangling; (2) **Iteration 1**, the construction and validation of a baseline end-to-end pipeline; (3) **Iteration 2**, the optimisation and advanced generation that builds on that benchmark; and (4) a continuous phase of **Deliverables and Documentation** that runs throughout the project. The full subtask breakdown of every phase is given in App. A.2, and its task-level schedule across the 17 project weeks in the cronogram (Table 10, App. A.1).

5 State of the art

Recent comprehensive surveys on Graph Data Augmentation (GDA) [4], [6] highlight that augmenting graph structures typically falls into two main perspectives: data-level augmentation (modifying features or topology) and algorithm-level augmentation [6].

Given that in brain networks the topology itself carries critical biological information (such as the presence or absence of a neural pathway), this project primarily focuses on topology-level augmentation strategies.

At the algorithm level, GDA is divided into model-free methods, which rely on predefined heuristic rules, and learning-based methods, which utilize parameterized models to learn the underlying data distribution [6].

The challenge, however, remains to perform this topological augmentation without disrupting the underlying clinical significance of the medical data [3][4][5].

Basic topology-level methods based on random edge perturbation (such as DropEdge)[18], while highly effective for social networks, have proven detrimental when applied to biomedical and biochemical graphs, as they randomly destroy essential structural properties [4].

Consequently, to successfully augment multiple sclerosis datasets, recent state-of-the-art approaches propose using advanced generative models such as Generative Adversarial Neural Network Autoencoders (AAE) to synthesize new structural brain networks that preserve topological properties and significantly improve clinical classification tasks [5].

Beyond deep generative models, other mathematical paradigms have also been successfully explored to mitigate data scarcity in neuroimaging. For instance, recent studies have demonstrated the viability of using low-dimensional dynamical systems as a source of data augmentation [7]. By employing whole-brain computational models based on nonlinear oscillators coupled by empirical structural connectivity, researchers can generate synthetic surrogate matrices that yield classification accuracy comparable to empirical data [7].

While dynamical systems offer a robust alternative, the use of parameterized generative models remains a highly versatile and leading approach. This study systematically evaluates such approaches on MS connectome datasets to identify the most effective augmentation strategy.

To identify the most effective augmentation strategy for MS connectome datasets, this project adapts and evaluates several distinct architectures from recent literature, comparing them against established baselines and custom approaches. Specifically, we implement the architecture proposed in BrainNetGAN [8], which is highly relevant to our work as it successfully utilized generative adversarial networks on brain connectivity matrices for dementia classification. We also explore the architecture of a Conditional Autoencoder [9] to understand how latent representations can be conditioned to reconstruct functional and structural networks. Furthermore, we incorporate a Topological Model based on graph diffusion [10], whose original formulation enforces explicit persistent-homology constraints to keep the generated networks topologically plausible; here we adapt its diffusion backbone to our adjacency-matrix format and evaluate empirically how well topology is preserved before imposing those explicit constraints.

Beyond these advanced deep learning architectures, we evaluate simpler and traditional data augmentation approaches. As a fundamental model-free baseline, we implement random edge perturbation techniques, specifically DropEdge [18], which randomly removes a predefined ratio of edges from the adjacency matrix during training. While literature suggests such random modifications may degrade biological networks [4], empirically evaluating this method on our specific MS dataset provides a critical negative benchmark against which the advanced generative models can be compared. Furthermore, recent comprehensive surveys highlight the rapid advancement of synthetic tabular data generation techniques [11]. By treating flattened adjacency matrices as tabular features, we can adapt these methods for our purposes. We also employ the well-established SMOTE (Synthetic Minority Oversampling Technique) [12] as a traditional benchmark. Finally, we design and implement a custom statistical model. This purely statistical approach, developed exclusively for this project, synthesizes new graphs by sampling each connection independently from a truncated Gaussian fitted to the empirical mean and variance of that connection within each stratum, bounded by its observed minimum and maximum values.

Generating synthetic data represents only the first phase of the GDA pipeline; current state-of-the-art literature emphasizes that the critical engineering challenge lies in validating its quality [13]. Because generative models optimize abstract loss functions, their output must be subjected to a rigorous framework to ascertain statistical realism, biological fidelity, and clinical utility.

To ensure synthetic matrices accurately reflect the original distributions without memorizing training data (GDPR compliance), leading frameworks utilize the Kolmogorov-Smirnov (KS) Test for marginal distributions and the Multivariate Hellinger Distance to capture high-dimensional covariance structures [13]. Furthermore, structural distances to the closest real records must be measured to prevent membership inference risks [13].

Statistically realistic graphs must also preserve human brain architecture. Standard graph-theoretical libraries are employed to compute measures highly sensitive to neurological disruption, such as global and local efficiency (measuring integration and segregation) and node strength (preserving critical cognitive "hubs") [8][14]. Additionally, advanced Topological Data Analysis (TDA) using persistent homology is increasingly adopted to guarantee macrostructural isomorphism between real and synthetic cohorts [8][15].

Ultimately, the final validation of any GDA strategy is its downstream clinical utility. The "Train on Synthetic, Test on Real" (TSTR) paradigm has become the gold standard for verifying predictive power [16]. Given the inherent class imbalances in clinical datasets (such as 199 MS patients versus 71 healthy volunteers in this study), relying on simple accuracy is misleading. Consequently, robust evaluation frameworks in the literature prioritize threshold-agnostic metrics like the F1-Score, Cohen's Kappa, and primarily the Matthews Correlation Coefficient (MCC), which is mathematically proven to be the most reliable metric for binary classification on skewed datasets [17].

By systematically implementing these generative architectures alongside this multi-dimensional evaluation framework, this study identifies the most effective augmentation strategy for neuroimaging datasets.

6 Experimental Design

The evaluation follows the leakage-free protocol summarised in Fig. 1. Each stage of this pipeline is detailed in a dedicated section: the cohort, its three channels and the stratified split in Sec. 7; the six generators that produce the synthetic cohort in Sec. 8; the intrinsic audit of that cohort in Sec. 9; and the downstream classifiers in Sec. 10. The remainder of this section describes the protocol that wraps these components together, organised by methodological component.

Cross-validation protocol. Rather than a single split, we use a repeated stratified cross-validation: 5 folds $\times$ 3 seeds

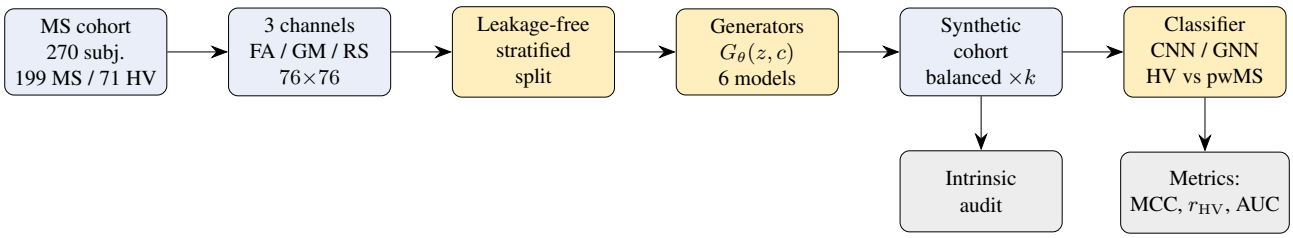


Fig. 1: End-to-end pipeline. Generators are fit on the training split of each cross-validation fold only; the synthetic cohort is audited intrinsically (statistical, topological, TDA) and used to augment the classifier, which is always scored on the held-out real test split.

$= 15$ evaluation units, each stratified jointly by disease status and institutional site (BCN/NAP) so that every fold preserves the cohort's $\approx 74\%$ pwMS prevalence and its site mix. Within each unit the generators are fit exclusively on that fold's ≈ 216 training subjects; neither the generators nor the classifiers ever see the ≈ 54 held-out real subjects, which serve only for scoring. Wrapping the entire generator–classifier pipeline inside the cross-validation makes every reported gain both paired and leakage-free, at the deliberate cost of folding generator variance into the estimate (a conservative choice that widens confidence intervals).

Evaluation regimes. The synthetic graphs of each unit are first audited along the three intrinsic axes of Sec. 9, and then used to train the classifiers under three regimes, all scored on the same held-out real split: TRTR (Train Real, Test Real, the baseline), TSTR (Train Synthetic, Test Real, a strict fidelity probe), and TR+S (Train Real + Synthetic, the augmentation regime of interest).

Two studies. (A) Generator benchmark: all six generators at augmentation factor k=1, matching the real class proportions, to rank them on a level field. **(B) Balanced augmentation sweep:** the synthetic cohort is forced to a 50/50 HV:pwMS balance and its size swept at $k \in \{1, 2, 3\}$, isolating whether balanced augmentation, and increasing amounts of it, relieves the class imbalance.

Statistical analysis. Because the real-only baseline drifts slightly between runs (GPU non-determinism), every augmented cell is reported as its paired per-fold change in MCC (ΔMCC) against that fold's own TRTR baseline, aggregated over the 15 units and tested with a two-sided Wilcoxon signed-rank test ($\alpha = 0.05$); we additionally report the fraction of folds that improve.

TABLE 1: EXPERIMENTAL HYPERPARAMETERS

Parameter	Value
Cross-validation	5 folds $\times$ 3 seeds $= 15$ units
Train/test per fold	≈ 216 / 54 stratified
Augmentation factor k	1 (benchmark); 1, 2, 3 (balanced sweep)
Generator epochs	300
Classifier epochs	150
CNN / GNN batch	32 / 16
TDA filtration	Vietoris–Rips, max dim 1

7 DATA AND PREPROCESSING

The empirical data consists of two anonymised MS cohorts (BCN and NAP) [1], [2], totaling $N_{\text{subj}} = 270$ subjects (71 healthy volunteers, 199 pwMS). Table 2 summarises their demographics.

TABLE 2: DEMOGRAPHICS OF THE TWO COHORTS

	BCN (n=165)		NAP (n=105)	
	HV	pwMS	HV	pwMS
Subjects, n	18	147	53	52
Age (years)	36.6 ± 9.3	47.3 ± 10.1	41.3 ± 11.6	42.5 ± 12.8
Female, %	83%	71%	62%	67%
Disease dur. (yr)	–	16.0 ± 9.0	–	13.1 ± 9.1
EDSS (median)	–	2.0	–	4.5

Following the Desikan–Killiany atlas [19], each subject's cortex is divided into $N = 76$ regions of interest (ROIs). For each subject i, three symmetric adjacency matrices are extracted: Structural FA (DTI), Morphological GM (T1), and Functional RS (rs-fMRI). These form a multi-channel tensor $X_i \in \mathbb{R}^{C \times N \times N}$ ($C = 3$). Per-channel domain constraints (Table 3) are enforced on all synthetic outputs to preserve biological validity.

TABLE 3: PER-CHANNEL STRUCTURAL CONSTRAINTS

Channel	Modality	Diagonal	Range
FA	Structural (DTI)	0	$[0, 1]$
GM	Morphological (T1)	1	$[0, 1]$
RS	Functional (rs-fMRI)	1	$[-1, 1]$

To prevent data leakage, every split is stratified by disease status and institutional site, and the whole pipeline is wrapped in a repeated cross-validation (5 folds $\times$ 3 seeds; Sec. 6). No generator or classifier is ever trained on data reserved for evaluation.

8 GENERATIVE MODELS

We evaluate six conditional generators $G_\theta(z, c)$ (Table 4) capable of producing synthetic tensors matching a target stratum c (disease state $\times$ acquisition site). Because brain connectivity matrices obey strict geometric and clinical constraints, each architecture approaches non-Euclidean generation through a distinct mathematical lens, ranging from model-free topological perturbation to high-dimensional continuous diffusion.

Let $x \in \mathbb{R}^{C \times N \times N}$ denote a (vectorised) real multi-channel connectome, $\tilde{x}$ its synthetic counterpart and c the conditional stratification vector that combines disease state and acquisition site. The operating rule of each generator is as follows:

- **CustomStat**: a non-parametric structural baseline engineered specifically for this study. Rather than

TABLE 4: THE SIX GENERATORS IN OUR STUDY

Model	Family	Origin
CustomStat	Statistical	Ours
DropEdge	Model-free	[18]
SMOTE	Tabular	[12]
BrainNetGAN	Conditional GAN	[8], [24]
AAE	Adversarial AE	[5], [21]
GraphDiffusion	Conditional DDPM	Adapted from [10]

learning a joint feature representation, it estimates the empirical marginal of each individual edge (i, j) within each stratum and samples every connection independently from a localised truncated Gaussian, $\tilde{x}_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2_{ij})$ bounded strictly by the empirical $[\min_{ij}, \max_{ij}]$ so that channel ranges are respected by construction; edges that are always zero in the cohort stay zero, preserving the sparsity pattern.

- **DropEdge [18]**: a model-free negative benchmark operating directly on real graphs. For a randomly chosen real subject of the target stratum, a fixed random fraction p=0.1 of its off-diagonal edges is zeroed, $\tilde{A} = A \odot M$; to maintain inter-channel alignment the same topological mask is broadcast across the structural, morphological and functional channels.
- **SMOTE [12]**: a classic tabular over-sampling technique. The upper triangles of the three channels are concatenated into a flat 8 702-dimensional vector and SMOTE is run independently within each (site, label) stratum: a real sample x is interpolated toward one of its K=5 nearest neighbours, $\tilde{x} = x + \alpha\,(x_{\mathrm{nn}} - x)$ with $\alpha \sim \mathcal{U}(0, 1)$, so synthetic subjects lie on the convex hull of the real ones.
- **BrainNetGAN [8], [24]**: a conditional Wasserstein GAN with gradient penalty (WGAN-GP) [20]. A four-layer feed-forward generator maps a latent Gaussian vector $z \in \mathbb{R}^{64}$ and the one-hot condition to the connectome, while an unconditional critic D built from BrainNetCNN edge-to-edge / edge-to-node kernels [24] is trained with

$$\mathcal{L}_D = \mathbb{E}[D(\tilde{x})] - \mathbb{E}[D(x)] + \lambda\,\mathbb{E}\big[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2\big], \tag{1}$$

  where $\hat{x}$ is sampled uniformly on segments between real and synthetic pairs and $\lambda = 10$. Class conditioning is enforced not through the critic but through an auxiliary BrainNetCNN classifier that steers G toward the requested label.
- **Adversarial Autoencoder (AAE) [5], [21]**: a fully connected encoder/decoder backbone. The encoder maps a real input to a 32-dimensional latent code $z = \mathrm{enc}(x, c)$ that an auxiliary discriminator forces to match a standard Gaussian prior $\mathcal{N}(0, \mathbf{I})$, while the conditional decoder reconstructs the matrix, minimising $\mathcal{L} = \mathbb{E}\|x - \mathrm{dec}(\mathrm{enc}(x, c), c)\|^2 + \mathcal{L}_{\mathrm{adv}}\big(q(z) \,\|\, \mathcal{N}(0, \mathbf{I})\big)$.
- **Graph Diffusion [10], [22]**: a conditional Denoising Diffusion Probabilistic Model (DDPM). The forward process corrupts real connectivity vectors over T=200 steps under a linear variance schedule $q(x_t \mid x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\,x_{t-1}, \beta_t \mathbf{I})$, $\beta_t \in [10^{-4}, 2{\times}10^{-2}]$; a condition-aware MLP $\epsilon_\theta(x_t, t, c)$ is trained to predict the injected noise via $\mathbb{E}\|\epsilon - \epsilon_\theta(x_t, t, c)\|^2$ and synthesises samples through the iterative reverse chain.

9 EVALUATION METRICS

Because generative models optimise abstract loss functions, their output cannot be trusted on the training objective alone. We therefore audit every synthetic cohort along three intrinsic axes before measuring its task utility (Sec. 10). All distances compare the synthetic cohort against the real training split only, so that the held-out subjects remain unseen. Note that element-wise errors such as MSE/RMSE are ill-defined here: synthetic and real graphs are unpaired, so there is no correspondence to subtract. We therefore rely on distributional and topological distances; App. A.7 gives a plain-language primer on each one, with emphasis on the TDA pipeline.

Statistical fidelity. For each of the $D = 8\,702$ upper-triangular edges we compare the real and synthetic marginals with the two-sample Kolmogorov–Smirnov (KS) statistic and the per-edge Hellinger distance $H(p, q) = \frac{1}{\sqrt{2}}\|\sqrt{p} - \sqrt{q}\|_2$, averaged over edges. Joint structure is captured by the kernel Maximum Mean Discrepancy (MMD) with an RBF kernel and by the energy distance. To rule out memorisation of real subjects (a GDPR requirement), we report the privacy ratio

$$\rho_{\mathrm{priv}} = \frac{\mathbb{E}\big[\, d(x_{\mathrm{synth}}, \mathrm{NN}_{\mathrm{real}}(x_{\mathrm{synth}}))\,\big]}{\mathbb{E}\big[\, d(x_{\mathrm{real}}, \mathrm{NN}_{\mathrm{real}}(x_{\mathrm{real}}))\,\big]}, \tag{2}$$

the mean nearest-neighbour distance from synthetic to real samples normalised by the real-to-real baseline; $\rho_{\mathrm{priv}} \ll 1$ flags copied records, whereas $\rho_{\mathrm{priv}} \gtrsim 1$ indicates genuinely novel samples.

Graph-topological measures. Each channel is read as a weighted graph and summarised by classical network descriptors [14]: the node strength $s_i = \sum_j w_{ij}$, the weighted degree, and the local and global efficiency, with $E_{\mathrm{glob}} = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}^{-1}$ where d_{ij} is the shortest weighted path. These quantify, respectively, the functional segregation and integration that neurological damage is known to disrupt. For every descriptor we compare its real and synthetic distributions through the Earth Mover's Distance (EMD), reported per channel (Table 7).

Topological data analysis (TDA). To verify macrostructural isomorphism we build a Vietoris–Rips filtration on each connectome and extract its persistence diagram [15]. Real and synthetic diagrams are compared through the bottleneck distance and the L_1 distance between Betti curves (β_0 counts connected components, β_1 counts loops), and we track the relative deviation in total persistence $\Delta\mathrm{TP}$, which exposes generators that inflate or destroy topological structure (reported in Sec. 11).

10 DOWNSTREAM CLASSIFICATION

The empirical utility of synthetic graphs is validated via binary classification (HV vs. pwMS) with two deliberately contrasting architectures, so that conclusions do not hinge on a single inductive bias. Their layers and training hyperparameters are summarised in Table 5.

CNN: BrainNetClassifier. The $3 \times 76 \times 76$ tensor is first symmetrised ($X \leftarrow \frac{1}{2}(X + X^\top)$ per channel) and read as a 3-channel image. The feature extractor stacks four 3×3 convolutions (stride 1, padding 1, so spatial resolution is preserved) in two blocks of width 32 then 64, each convolution followed by batch normalisation and ReLU; a single 2×2 max-pooling separates the blocks and an adaptive global-average pooling collapses the final 64 feature maps into a 64-dimensional embedding. A two-layer MLP head ($64 \to 64 \to 2$, ReLU, dropout 0.3) outputs the logits. All down-sampling is thus performed by the lone max-pool and the global pool, while the convolutions preserve the 76×76 grid.

GNN: BrainGNNClassifier. Each subject becomes a graph of 76 nodes. The structural FA channel defines the topology: an edge (i, j) is kept where $\mathrm{FA}_{ij} > 0$, with weight FA_{ij} normalised by the subject's maximum. The feature of node i is the concatenation of its FA, GM and RS connectivity rows ($3 \times 76 = 228$ features). Two Graph Convolutional layers propagate node features via the GCN propagation of Kipf and Welling [23]:

$$H^{(l+1)} = \mathrm{ReLU}\big(\hat{D}^{-1/2}\hat{A}\,\hat{D}^{-1/2}\,H^{(l)}W^{(l)}\big), \qquad (3)$$

where $\hat{A} = A + I$ is the self-looped weighted adjacency and $\hat{D}$ its degree matrix (dimensions $228 \to 64 \to 64$, each layer followed by batch normalisation); global mean pooling yields the graph embedding, which is fed to the same $64 \to 64 \to 2$ MLP head.

Training. Both networks are optimised with Adam (learning rate 10^{-3}, weight decay 10^{-4}) under an inverse-frequency weighted cross-entropy that already counters the 199:71 imbalance at the loss level. Predictive quality is summarised by the Matthews Correlation Coefficient (MCC), our headline metric for this skewed task [17], complemented by balanced accuracy and the minority (HV) recall, which together expose whether a gain reflects genuine minority-class discrimination rather than majority bias.

TABLE 5: CLASSIFIER ARCHITECTURES AND TRAINING HYPERPARAMETERS

	CNN	GNN
Input	3×76×76 image	76 nodes, 228-d feat.
Core	4×Conv3×3	2×GCNConv
Channels/dims	32, 64	228→64→64
Conv stride/pad	1 / 1	n/a
Pooling	MaxPool2 + glob. avg	global mean
Embedding	64	64
Head	64→64→2	64→64→2
Dropout	0.3	0.3
Optimiser	Adam, lr 10^{-3}, wd 10^{-4}	
Batch size	32	16
Epochs	150	150
Loss	inv.-frequency weighted CE	

11 VALIDATION RESULTS

Before measuring downstream utility, every synthetic cohort is audited intrinsically along the three axes of Sec. 9: statistical fidelity to the real distribution, the privacy ratio that guards against memorisation, and the graph-topological and persistent-homology descriptors that capture biological plausibility. All distances are computed against the real training split only, so the held-out subjects stay unseen.

Statistical fidelity (Table 6). The per-edge distances expose two opposite extremes. SMOTE attains the best Hellinger, MMD and energy distances and a near-best Kolmogorov–Smirnov, simply because each of its samples is a convex interpolation of two real subjects; this apparent fidelity is an artefact of copying, as its privacy ratio confirms below. At the other end, GraphDiffusion is the joint-distribution outlier on every metric, with an energy distance of 12.79 (about an order of magnitude above any other generator) and a Hellinger of 0.465, so its reverse-diffusion samples drift far from the real marginals. The remaining four generators form a healthy middle band: DropEdge reaches the lowest KS (0.089) because it only perturbs real edges, while CustomStat, BrainNetGAN and AAE keep all four distances within a small factor of one another, the signature of genuinely novel yet distributionally plausible graphs.

TABLE 6: STATISTICAL FIDELITY AGAINST THE REAL TRAINING SPLIT (LOWER IS BETTER; THE PRIVACY RATIO $\rho_{\mathrm{PRIV}} \approx 1$ IS HEALTHY WHEREAS $\ll 1$ SIGNALS MEMORISATION).

Generator	KS	Hell.	MMD	Energy	ρ_{priv}
CustomStat	0.271	0.202	0.027	1.01	0.85
DropEdge	**0.089**	0.175	0.013	0.67	1.89
SMOTE	0.110	**0.121**	**0.009**	**0.40**	0.26
BrainNetGAN	0.325	0.176	0.026	1.35	1.01
AAE	0.271	0.161	0.020	0.84	0.79
GraphDiffusion	0.373	0.465	0.114	12.79	2.81

Memorisation and privacy (Fig. 2). The privacy ratio ρ_{priv} separates true generation from copying: it is the mean synthetic-to-real nearest-neighbour distance normalised by the real-to-real baseline, so values near 1 mark novel samples and values $\ll 1$ mark memorisation. SMOTE collapses to $\rho_{\mathrm{priv}} = 0.26$, its synthetic subjects sitting roughly four times closer to real ones than real subjects sit to each other, a membership-inference risk that disqualifies it on GDPR grounds despite its strong distances. Every other generator is safe, at or above the ideal value (CustomStat 0.85, AAE 0.79, BrainNetGAN 1.01, DropEdge 1.89, GraphDiffusion 2.81), so their gains, where present, cannot come from leaking training data.

Topological fidelity (Table 7). Read as weighted graphs, the synthetic connectomes are compared to the real ones through node strength and local and global efficiency, per channel. The five faithful generators track the real topology within a small factor on every descriptor and channel; GraphDiffusion is again the outlier, deviating by an order of magnitude (for instance a node-strength EMD of 24.97 on the RS channel against ≤ 3.4 for the rest, and a global-efficiency EMD of 0.317 against ≤ 0.026), the structural counterpart of its statistical drift.

Persistent homology (Table 8). The Vietoris–Rips analysis tells the same story at the macro-structural scale. The faithful generators keep the bottleneck and Betti-curve distances small and their total-persistence deviation near zero (between -9.3% and $+15.3\%$), whereas GraphDiffusion inflates total persistence by $+88.9\%$ and shows the largest bottleneck distances, meaning its graphs carry many more

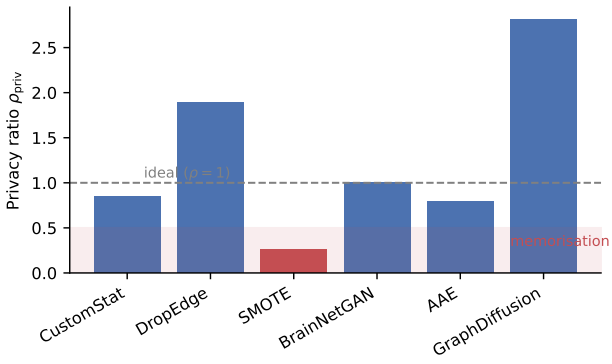


Fig. 2: Privacy ratio ρ_{priv} (mean synthetic-to-real over real-to-real nearest-neighbour distance) per generator. Values near 1 (dashed line) indicate genuinely novel samples; the shaded band ($\rho_{\text{priv}} < 0.5$) flags memorisation. SMOTE (red) copies real subjects, a membership-inference (GDPR) risk despite its strong distance metrics, whereas the genuine generators sit at $\rho_{\text{priv}} \gtrsim 1$.

TABLE 7: TOPOLOGICAL EMD PER CHANNEL (15 CV UNITS; LOWER IS BETTER)

	FA			GM			RS		
Generator	NS	LE	GE	NS	LE	GE	NS	LE	GE
CustomStat	0.70	0.033	0.022	3.77	0.084	0.046	3.39	0.025	0.026
DropEdge	2.11	0.014	0.012	3.59	0.026	0.023	2.12	**0.010**	**0.008**
SMOTE	**0.22**	**0.010**	**0.007**	**1.05**	0.027	**0.004**	2.11	0.028	0.022
BrainNetGAN	0.69	0.025	0.013	1.88	0.050	0.019	**1.88**	0.023	0.019
AAE	0.57	0.028	0.012	2.19	0.061	0.026	2.63	0.026	0.014
GraphDiffusion	5.85	0.203	0.244	8.67	0.326	0.411	24.97	0.285	0.317

and longer-lived loops and connected components than real brains, the fingerprint of spurious connections injected by the sampler.

TABLE 8: TDA FIDELITY ON THE FA CHANNEL (15 CV UNITS)

Generator	B_0	B_1	$L_1^{(0)}$	$L_1^{(1)}$	ΔTP
CustomStat	0.066	**0.064**	2.29	14.4	-9.3%
DropEdge	**0.044**	0.087	**0.17**	**1.69**	-2.9%
SMOTE	0.068	0.073	0.40	2.17	$+0.4\%$
BrainNetGAN	0.217	0.136	8.49	25.3	$+7.5\%$
AAE	0.185	0.099	8.59	28.4	$+15.3\%$
GraphDiffusion	0.369	0.215	22.9	42.6	$+88.9\%$

Fidelity does not predict utility. Taken together these axes rank the generators from most to least faithful as SMOTE (by copying), then the statistical and adversarial models, and finally GraphDiffusion. As Sec. 12 shows, this ranking is essentially inverted at the downstream task: the least faithful generator is the most useful CNN regulariser, while the most faithful one must be discarded for memorisation. High distributional fidelity is therefore neither sufficient nor necessary for downstream usefulness, which is why every augmentation claim in this work is backed by a utility measurement and a privacy check rather than by a distance metric alone.

12 RESULTS

All numbers are aggregated over the 15 cross-validation units (mean $\pm$ std); the headline of every augmented cell is its paired ΔMCC against that fold's own baseline, tested with a Wilcoxon signed-rank test, as defined in Sec. 6.

Generator benchmark (k=1). At factor one, no generator produces a significant positive. The faithful BrainNetGAN now converges to a healthy distribution ($\rho_{\text{priv}} = 1.01$, Table 6) but still degrades the CNN (ΔMCC $= -0.164$, $p = 0.011$), and AAE significantly degrades the GNN (-0.087, $p = 0.012$). The only consistent upward trend is GraphDiffusion on the CNN (ΔMCC $= +0.074$, $11/15$ folds, $p = 0.12$).

Balanced augmentation (Table 9). Forcing a $50/50$ class balance and raising the factor turns that trend significant: GraphDiffusion + CNN at k=3 raises MCC by $+0.074$ ($11/15$ folds, $p = 0.035$), with every secondary metric also improving (accuracy $+1.7\%$, balanced accuracy $+0.040$, HV recall $+0.087$; Fig. 3) and without memorisation ($\rho_{\text{priv}} = 2.81$). The gain scales with the amount of balanced synthetic data (Fig. 4), the signature of a regulariser rather than of added biological signal. A second, equally noteworthy result comes from our own baseline: CustomStat + GNN at k=3 lifts MCC by $+0.054$ ($10/15$ folds). Although it falls short of strict significance ($p = 0.21$), the gain is consistent and, crucially, achieved with a healthy privacy ratio ($\rho_{\text{priv}} = 0.85$): unlike SMOTE, CustomStat synthesises genuinely novel, non-memorised graphs, and its independent per-edge marginals act as a useful inductive prior for the message passing of the GCN. Extending the same balanced sweep to the two deep adversarial generators confirms neither helps: BrainNetGAN significantly degrades both classifiers at every factor (CNN ΔMCC $= -0.197$, $p = 0.018$; GNN -0.098, $p = 0.041$ at k=3) and AAE is statistically neutral on the CNN (-0.008, $p = 0.80$) yet degrades the GNN (-0.128, $p = 0.026$); neither improves with k (App. A.4), leaving the diffusion regulariser and the statistical baseline as the only models that benefit from scaling.

TABLE 9: BALANCED AUGMENTATION AT k=3 (15 CV UNITS). TR+S MCC IS THE MEAN$\pm$STD; ΔMCC THE PAIRED CHANGE VS. EACH FOLD'S OWN REAL-ONLY BASELINE. $^*p < 0.05$ (WILCOXON); † SMOTE MEMORISES.

	CNN		GNN	
Generator	TR+S MCC	ΔMCC	TR+S MCC	ΔMCC
CustomStat	$0.512 \pm .130$	$+0.007$	$0.569 \pm .132$	$+0.054$
DropEdge	$0.456 \pm .180$	$+0.029$	$0.489 \pm .110$	-0.071^*
SMOTE†	$0.368 \pm .182$	-0.084	$0.501 \pm .149$	$+0.028$
GraphDiffusion	$\mathbf{0.549 \pm .125}$	$\mathbf{+0.074}^*$	$0.452 \pm .109$	-0.008
BrainNetGAN	$0.240 \pm .149$	-0.197^*	$0.440 \pm .125$	-0.098^*
AAE	$0.389 \pm .197$	-0.008	$0.436 \pm .117$	-0.128^*

Architecture matters. The two beneficial configurations are architecture-specific: the diffusion regulariser helps only the CNN, which consumes each connectome as a Euclidean image, whereas the independent-edge statistical model helps only the GNN, whose message passing benefits from an uncorrelated edge prior. No single generator improves both classifiers, so the choice of augmenter cannot be separated from the downstream architecture.

13 DISCUSSION AND CONCLUSIONS

Our leakage-free, multi-axis evaluation supports three conclusions. First, balanced diffusion-based augmentation is

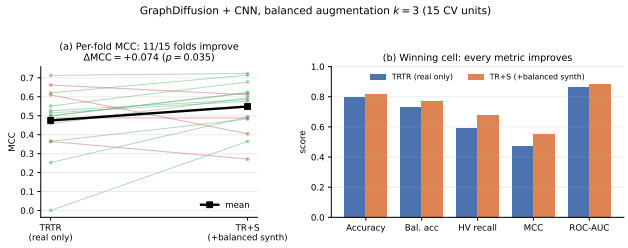


Fig. 3: Anatomy of the single statistically significant augmentation gain (GraphDiffusion + CNN, balanced, k=3; the same 15 CV units as Table 9). **(a)** Per-fold MCC paired between the real-only baseline (TRTR) and real + synthetic (TR+S): $11/15$ folds improve (green), raising the mean MCC by $+0.074$ ($p = 0.035$, Wilcoxon signed-rank). **(b)** The same configuration improves on every headline metric (accuracy, balanced accuracy, HV recall, MCC and ROC-AUC), confirming that the gain is genuine minority-class discrimination rather than a single-metric artefact.

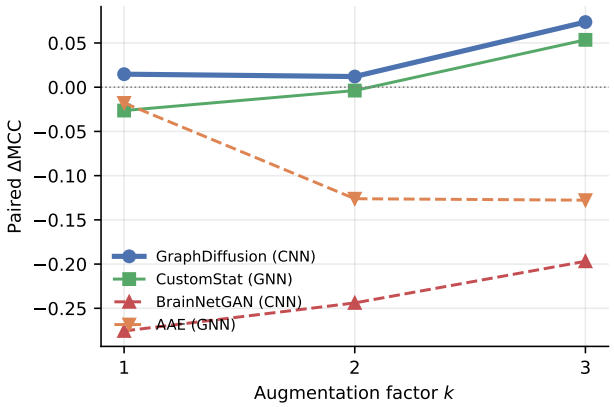


Fig. 4: Paired ΔMCC against each fold's own real-only baseline as the balanced augmentation factor k grows. GraphDiffusion (CNN) and CustomStat (GNN) strengthen with more balanced synthetic data, whereas the deep adversarial generators stay negative at every k. Full per-generator matrix in App. A.4.

the one configuration that yields a statistically significant gain: GraphDiffusion, class-balanced and at factor three, raises the CNN's MCC by $+0.074$ ($p < 0.05$) along with every secondary metric, and the effect grows with the amount of synthetic data (Fig. 4), the signature of a regulariser rather than of added biological information. A second, weaker signal, which we read as a promising positive trend rather than a confirmed benefit, comes from the CustomStat baseline on the GNN (ΔMCC $= +0.054$, $10/15$ folds, $p = 0.21$, $\rho_{\text{priv}} = 0.85$): sampling each edge independently appears to supply a useful inductive prior for message passing, reaching the largest ΔMCC of any generator on that architecture. Independent edge sampling is, in fact, well matched to graph convolutions: it preserves each edge's local weight distribution without injecting the correlated, aberrant community structure that would distort node neighbourhoods and aggravate the over-smoothing of the Kipf–Welling propagation[23] in Eq. 3.

Second, distributional fidelity is decoupled from clinical utility. The faithful BrainNetGAN now converges to a realistic distribution ($\rho_{\text{priv}} \approx 1$) yet still degrades the CNN, whereas the topologically implausible GraphDiffusion is the best CNN regulariser, a paradox with a clear regularisation reading: because the CNN consumes each connectome as a Euclidean image, GraphDiffusion's distorted samples act as structured high-dimensional noise that breaks overfitted local patterns, an effect closer to aggressive dropout or weight decay than to the synthesis of more realistic brains. SMOTE, the most faithful generator, earns that fidelity only by interpolating between (essentially copying) real patients, which is precisely why its privacy ratio collapses to $\rho_{\text{priv}} = 0.26$; its strong fidelity numbers are thus an artefact of memorisation and it must be discarded on GDPR grounds. Augmentation claims must therefore be backed by downstream utility and a privacy check, never by a distance metric alone.

Third, gains are modest by construction: the classifiers already counter the 199:71 imbalance through inverse-frequency weighting (which partly overlaps with data balancing) and 216 real subjects cap the attainable signal; utility is moreover architecture-dependent (diffusion helps the CNN, the statistical model helps the GNN). The balanced sweep, now extended to BrainNetGAN and AAE, closes the comparison of all six generators: neither deep adversarial model beats its own baseline at any factor or classifier ($0/12$ configurations positive), with BrainNetGAN being significantly harmful and AAE neutral-to-harmful, so among the six the only reliable augmenters remain the diffusion regulariser and the statistical model. This reinforces the broader observation that, on cohorts of this size, heavily parameterised adversarial generators are outperformed by simpler statistical and noise-based priors. Furthermore, the high initial dimensionality of the GNN (concatenating $3 \times 76 = 228$ features per node) relative to the small subject pool likely exacerbates overfitting; exploring dimensionality reduction techniques (e.g., PCA on node features) could stabilize GNN performance in future iterations.

Future work will fold persistent-homology constraints [10] into the diffusion sampler, so as to combine its regularising benefit with biological plausibility. In addition, addressing the sample size limitations of adversarial models through transfer learning, such as pre-training BrainNetGAN on massive healthy connectome datasets like the Human Connectome Project (HCP) before fine-tuning on our MS cohort, presents a promising avenue to achieve stable Nash equilibria. Finally, while averaging the Kolmogorov–Smirnov statistic across 8 702 edges provides a useful aggregate of fidelity, future robust statistical evaluations should incorporate multiple testing corrections to isolate true biological signal from random noise. The complete code, the synthetic cohorts and the per-fold metrics behind every figure are released at `https://github.com/Miquel44/MS_MedicalDataAugmentation_TFG`.

REFERENCES

[1] E. Solana, E. Martinez-Heras, J. Casas-Roma, et al., "Modified connectivity of vulnerable brain nodes in multiple sclerosis, their impact on cognition and their discriminative value," *Scientific Reports*, vol. 9, no. 1, p. 20172, 2019.

[2] J. Casas-Roma, E. Martinez-Heras, A. Solé-Ribalta, et al., "Applying multilayer analysis to morphological, structural, and functional brain networks to identify relevant dysfunction patterns," *Network Neuroscience*, vol. 6, no. 3, pp. 916-932, 2022.

[3] T. Zhao, Y. Liu, L. Neves, O. Woodford, M. Jiang, and N. Shah, "Data Augmentation for Graph Neural Networks," in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pp. 11015-11023, 2021.

[4] M. Adjeisah, X. Zhu, H. Xu, and T. A. Ayall, "Towards data augmentation in graph neural network: An overview and evaluation," *Computer Science Review*, vol. 47, p. 100527, 2023.

[5] B. Barile, A. Marzullo, C. Stamile, F. Durand-Dubief, and D. Sappey-Marinier, "Data augmentation using generative adversarial neural networks on brain structural connectivity in multiple sclerosis," *Computer Methods and Programs in Biomedicine*, vol. 206, p. 106113, 2021.

[6] K. Ding et al., "Data Augmentation for Deep Graph Learning: A Survey," *arXiv preprint*, 2022.

[7] Y. Sanz Perl, C. Pallavicini, I. Perez Ipiña, M. Kringelbach, G. Deco, H. Laufs, and E. Tagliazucchi, "Data augmentation based on dynamical systems for the classification of brain states," *Chaos, Solitons and Fractals*, vol. 139, p. 110069, 2020.

[8] C. Li, Y. Wei, X. Chen, and C.-B. Schönlieb, "BrainNet-GAN: Data augmentation of brain connectivity using generative adversarial network for dementia classification," *arXiv preprint arXiv:2103.08494*, 2021.

[9] I. Batta, M. Ajith, and V. D. Calhoun, "Conditioned Graph Reconstruction of Brain Functional Network Connectivity Reveals Interpretable Latent Axes of Sex and Fluid Intelligence," *bioRxiv preprint*, 2024.

[10] J. Park, D. Lee, G. Wu, and W. H. Kim, "Conditional Graph Diffusion with Topological Constraints for Brain Network Generation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2023.

[11] R. Shi, Y. Wang, M. Du, X. Shen, Y. Chang, and X. Wang, "A Comprehensive Survey of Synthetic Tabular Data Generation," *arXiv preprint arXiv:2504.16506*, 2025.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[13] A. M. Alaa, B. van Breugel, E. S. Saveliev, and M. van der Schaar, "How faithfully does synthetic medical data reflect reality?," *npj Digital Medicine*, vol. 5, no. 1, p. 114, 2022.

[14] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage*, vol. 52, no. 3, pp. 1059-1069, 2010.

[15] A. E. Sizemore, J. E. Phillips-Cremins, R. Ghrist, and D. S. Bassett, "The importance of the whole: Topological data analysis for the network neuroscientist," *Network Neuroscience*, vol. 3, no. 3, pp. 656-673, 2019.

[16] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs," *arXiv preprint arXiv:1706.02633*, 2017.

[17] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.

[18] Y. Rong, W. Huang, T. Xu, and J. Huang, "DropEdge: Towards Deep Graph Convolutional Networks on Node Classification," in *International Conference on Learning Representations (ICLR)*, 2020.

[19] R. S. Desikan, F. Ségonne, B. Fischl, et al., "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, 2006.

[20] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved Training of Wasserstein GANs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5767–5777.

[21] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial Autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 6840–6851.

[23] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[24] J. Kawahara, C. J. Brown, S. P. Miller, et al., "BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment," *NeuroImage*, vol. 146, pp. 1038–1049, 2017.

APÈNDIX

A.1 Current cronogram (2 iterations)

Table 10 reports the planned 17-week schedule that drove the iterative pipeline described in Sec. 3 (Methodology) and Sec. 4 (Planning).

TABLE 10: PROJECT CRONOGRAM OVER 17 WEEKS ACROSS THE TWO ITERATIONS

Task	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17
Phase 1 - Research and Data Preparation																	
Literature Review & Setup	■	■															
Data Exploration & Vis.	■	■															
Data Wrangling		■	■														
Phase 2 - Model Design and Synthesis																	
Baseline Model Setup			■	■					■	■							
GAN Architecture Design			■	■					■	■							
Alternative models design				■	■						■	■					
Synthetic Generation					■	■	■	■	■	■	■	■	■				
Phase 3 - Topological and Empirical Validation																	
Validation methods design				■		■											
Topological Metrics							■	■			■	■	■				
Empirical Validation							■	■			■	■	■	■			
Alternative validation							■	■			■	■	■	■			
Phase 4 - Deliverables and Documentation																	
Initial Report (Meeting)		■															
Report 1						■											
Report 2										■							
Final Dossier Writing											■	■	■	■	■	■	
Project Defense																	■

A.2 Detailed work plan

The body (Sec. 4) summarises the four project phases; their full subtask breakdown, which drives the cronogram of Table 10, is detailed below.

Phase 1: Research and Data Preparation. This initial phase focuses on understanding the theoretical background and preparing the empirical data.

- **Literature Review & Setup:** Comprehensive study of state-of-the-art graph augmentation and environment configuration.
- **Data Exploration & Vis.:** Initial analysis and visual representation of the empirical MS and healthy volunteers datasets.
- **Data Wrangling:** Cleaning, normalizing, and preprocessing the adjacency matrices into tensors suitable for neural networks and for our data augmentation models.

Phase 2: Iteration 1 - Baseline Construction & Initial Validation. This phase builds the first end-to-end functional pipeline to establish a performance benchmark.

- **Baseline Model Setup:** Implementation of a simple generative model to create the first synthetic data batch.
- **Validation methods design:** Theoretical formulation of evaluation metrics.
- **Data augmentation Architecture Design:** Design and coding of the different data augmentation models (e.g., GAN).
- **Synthetic Generation:** Execution of the baseline model to generate augmented graphs.
- **Topological Metrics & Empirical Validation:** Testing the baseline synthetic data mathematically (e.g., local efficiency, median deviation) and empirically (GNN classifier) to gather preliminary results.

Phase 3: Iteration 2 - Optimization & Advanced Generation. Based on the benchmark from Iteration 1, this phase focuses on optimizing the generative process to achieve higher biological realism and better results.

- **Alternative models design:** Exploration of secondary generative algorithms to compare against the main GAN.
- **Synthetic Generation (Optimized):** Generating the final, high-quality augmented datasets using the advanced models.
- **Alternative validation:** Applying additional metrics and empirical tests to ensure the robustness of the optimized

results.

Phase 4 (Continuous Phase): Deliverables and Documentation. This phase runs continuously throughout the project to ensure proper communication and academic writing.

- **Reports (Initial, Report 1, Report 2):** Periodic milestones to present preliminary findings and methodology to the tutor.
- **Final Dossier Writing:** Continuous documentation of the experiments culminating in the final bachelor's thesis document.
- **Project Defense:** Preparation of the slides and the final oral defense.

A.3 Full empirical metric panel

Table 11 reports the full metric panel (accuracy, balanced accuracy, HV recall, MCC and ROC-AUC) for the headline balanced configuration at k=3. The TSTR rows make the memorisation/regularisation contrast explicit: SMOTE keeps high TSTR (it reuses real samples), whereas the genuine generators collapse under TSTR yet still help under TR+S. The two deep adversarial generators are now included: both collapse under TSTR (near-zero synthetic-only signal) and fail to recover the baseline under TR+S.

TABLE 11: FULL PANEL FOR THE BALANCED SWEEP AT k=3 (15 CV UNITS, MEAN ± STD). r_{HV} IS THE MINORITY (HV) RECALL.

		CNN					GNN				
Generator	Regime	Acc	bAcc	r_{HV}	MCC	AUC	Acc	bAcc	r_{HV}	MCC	AUC
–	TRTR	.793±.094	.733±.067	.606±.188	.505±.126	.866±.044	.805±.059	.741±.101	.604±.270	.516±.148	.876±.044
CustomStat	TSTR	.504±.097	.625±.062	.879±.074	.239±.108	.692±.067	.751±.058	.687±.074	.555±.163	.373±.145	.782±.062
	TR+S	.795±.058	.752±.084	.654±.229	.512±.130	.880±.032	.828±.050	.775±.073	.657±.175	.569±.132	.889±.041
DropEdge	TSTR	.721±.103	.731±.091	.752±.227	.428±.168	.834±.089	.756±.041	.763±.050	.776±.163	.485±.080	.843±.066
	TR+S	.811±.044	.695±.105	.446±.247	.456±.180	.843±.056	.799±.048	.734±.060	.595±.155	.489±.110	.829±.050
SMOTE†	TSTR	.774±.114	.750±.102	.706±.252	.489±.184	.872±.043	.828±.041	.779±.072	.673±.160	.562±.122	.882±.048
	TR+S	.714±.135	.674±.108	.603±.348	.368±.182	.837±.076	.812±.053	.729±.078	.550±.164	.501±.149	.851±.049
GraphDiffusion	TSTR	.484±.208	.502±.027	.552±.465	.003±.076	.559±.128	.509±.160	.517±.104	.542±.271	.048±.215	.503±.134
	TR+S	**.815**±.052	**.770**±.064	**.678**±.157	**.549**±.125	**.882**±.044	.791±.031	.716±.068	.554±.173	.452±.109	.823±.028
BrainNetGAN	TSTR	.467±.153	.521±.053	.634±.345	.040±.115	.539±.128	.495±.063	.476±.085	.432±.189	−.043±.152	.479±.107
	TR+S	.704±.120	.595±.080	.355±.282	.240±.149	.758±.085	.772±.053	.722±.075	.614±.178	.440±.125	.800±.069
AAE	TSTR	.704±.134	.649±.096	.536±.305	.301±.185	.772±.069	.737±.052	.731±.064	.716±.164	.429±.110	.818±.050
	TR+S	.737±.108	.693±.106	.595±.269	.389±.197	.805±.073	.742±.066	.733±.063	.711±.153	.436±.117	.820±.048

A.4 Full balanced sweep (ΔMCC, all generators)

Table 12 reports the paired ΔMCC against each generator's own per-fold real-only baseline, for every generator, classifier and augmentation factor in the balanced sweep (15 CV units per cell). Out of the 36 configurations only one is a significant positive (GraphDiffusion · CNN at k=3); the two deep adversarial generators are negative throughout, and BrainNetGAN is significantly harmful in five of its six cells.

TABLE 12: BALANCED-SWEEP PAIRED ΔMCC FOR ALL SIX GENERATORS, BOTH CLASSIFIERS, AT $k \in \{1,2,3\}$ (15 CV UNITS). $^*p < 0.05$ (WILCOXON SIGNED-RANK); † SMOTE MEMORISES (ρ_{PRIV}=0.26), SO ITS GAINS ARE DISCOUNTED.

	CNN ΔMCC			GNN ΔMCC		
Generator	k=1	k=2	k=3	k=1	k=2	k=3
CustomStat	+0.072	+0.005	+0.007	−0.026	−0.004	+0.054
DropEdge	+0.013	−0.080	+0.029	+0.043	−0.073 *	−0.071 *
SMOTE†	+0.057	−0.013	−0.084	+0.003	+0.084	+0.028
GraphDiffusion	+0.015	+0.012	**+0.074** *	−0.089 *	−0.046	−0.008
BrainNetGAN	−0.276 *	−0.244 *	−0.197 *	−0.097	−0.164 *	−0.098 *
AAE	−0.007	−0.073	−0.008	−0.018	−0.126 *	−0.128 *

A.5 Statistical fidelity

Fig. 5 visualises the per-edge statistical distances reported in Table 6 (Sec. 11): the Kolmogorov–Smirnov, Hellinger, MMD and energy distances to the real training split, per generator.

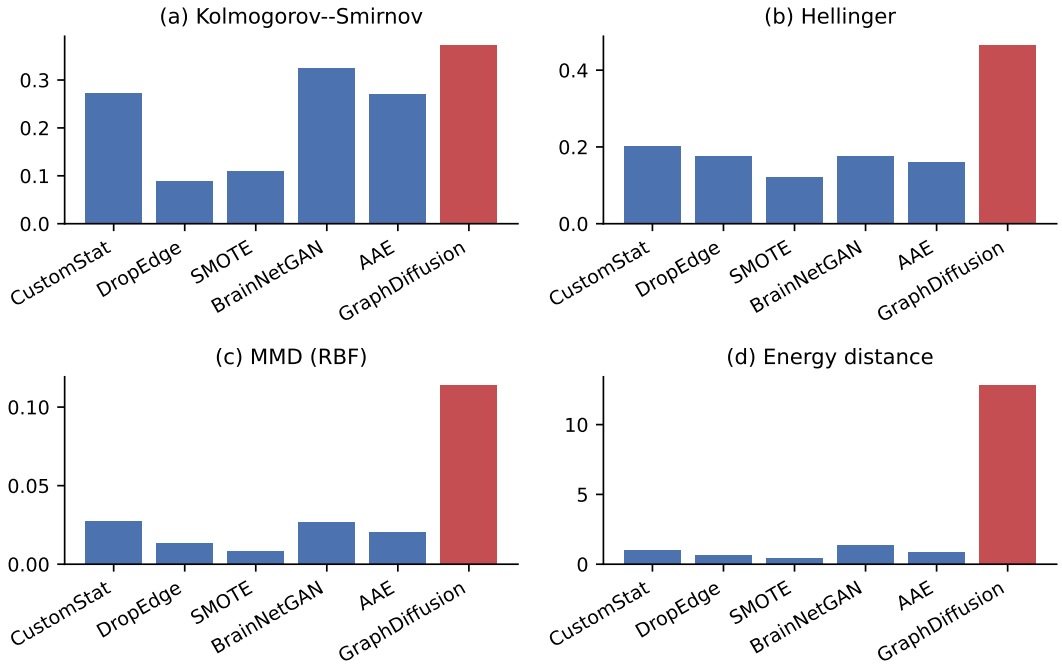


Fig. 5: Statistical-fidelity distances to the real training split, per generator: (a) Kolmogorov–Smirnov, (b) Hellinger, (c) MMD with an RBF kernel, (d) energy distance (all on a linear scale). GraphDiffusion (red) is the joint-distribution outlier, whereas SMOTE attains the smallest distances precisely because it interpolates real subjects (cf. its privacy ratio, Fig. 2).

A.6 Topological fidelity figures

Figures 6 and 7 visualise the per-channel topological distances of Table 7 and the persistent-homology fidelity of Table 8, whose numbers are reported in Sec. 11.

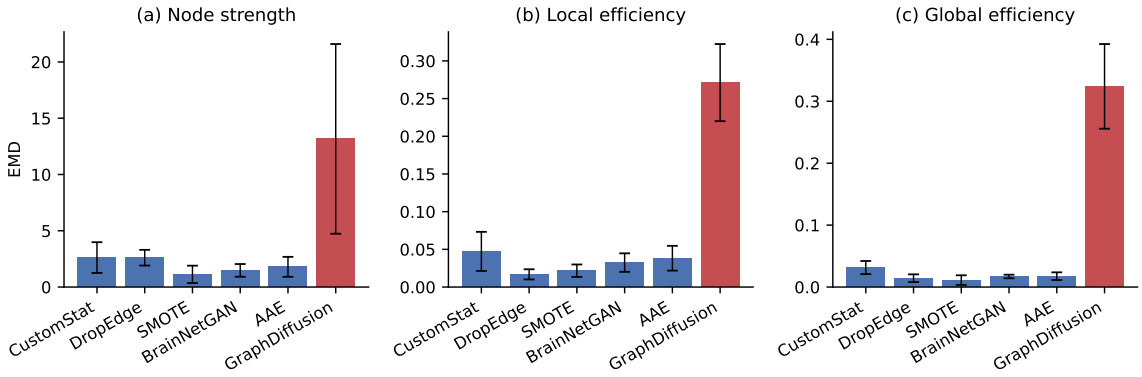


Fig. 6: Topological fidelity (EMD to the real distribution): (a) node strength, (b) local efficiency, (c) global efficiency, per generator (bar = mean over the FA/GM/RS channels, error bar = std across channels). The five faithful generators track the real topology within a small factor; GraphDiffusion (red) deviates by an order of magnitude on every descriptor, yet is the best CNN regulariser (Table 9), so topological realism and downstream utility are decoupled.

A.7 A primer on the fidelity metrics

This appendix expands the intuition behind the metrics of Sec. 9, with emphasis on the topological ones.

Topological data analysis (TDA), step by step. Persistent homology studies the shape of a network across connection scales instead of edge by edge. From a subject's weighted connectome we grow a Vietoris–Rips filtration: edges are switched on from strongest to weakest, producing a nested sequence of graphs in which, as the threshold sweeps, connected components merge and cycles (loops) open and then close. Every such feature is logged by the scale at which it is born and the scale at which it dies, which yields a persistence diagram with one point per feature; points far from the diagonal are long-lived and reflect genuine structure, whereas points near it are noise. We then read three numbers from this diagram. The Betti curves count, at each threshold, the connected components (β_0) and the independent loops (β_1)

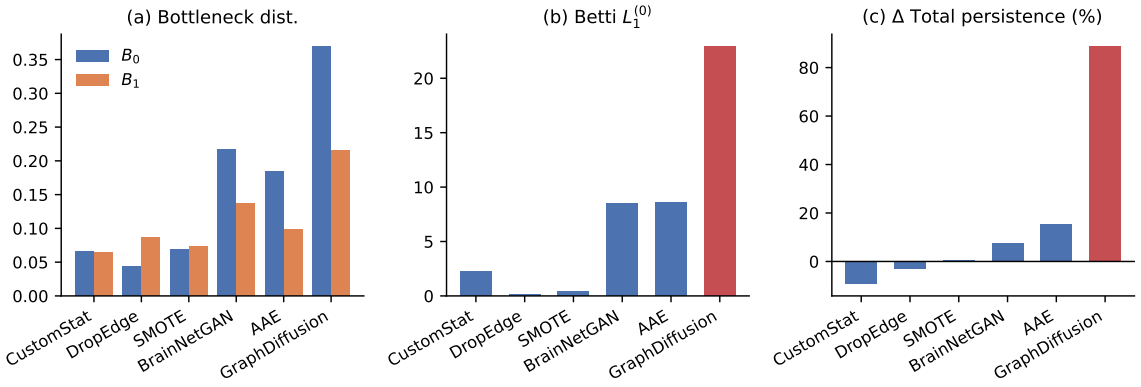


Fig. 7: Persistent-homology fidelity on the FA channel: (a) bottleneck distances B_0, B_1; (b) Betti-curve $L_1^{(0)}$ distance (linear scale); (c) relative deviation in total persistence. The faithful generators stay near zero, whereas GraphDiffusion (red) inflates total persistence by nearly 90%, confirming topologically implausible samples.

present, and the L_1 distance between the real and synthetic curves (the area between them) measures how differently the two cohorts are wired across scales. The bottleneck distances (B_0, B_1) are the standard diagram-to-diagram distances, that is, the smallest radius within which every real feature can be paired with a synthetic one, so small values mean matching topology. Total persistence adds up the lifespan (death minus birth) of all features, i.e. the total amount of topological structure, and $\Delta\mathrm{TP}$ is its relative real-to-synthetic deviation. A large positive $\Delta\mathrm{TP}$ (GraphDiffusion, $+89\%$) therefore means the synthetic graphs carry many more and longer-lived loops and components than real brains, the fingerprint of spurious connections injected by the generator.

Statistical and graph distances, in brief. The Kolmogorov–Smirnov statistic is the largest gap between the real and synthetic cumulative distributions of an edge, and the Hellinger distance a bounded ($[0, 1]$) overlap measure of the same marginals; the kernel MMD and the energy distance instead compare the joint distribution of all edges at once, so they react to inter-edge correlations that per-edge tests miss. Among the graph descriptors, node strength is the total connection weight reaching a region, global efficiency is the mean inverse shortest path (integration) and local efficiency the same restricted to each node's neighbourhood (segregation); each descriptor's real and synthetic distributions are compared with the Earth Mover's Distance, the minimal work needed to reshape one histogram into the other.

A.8 Released artefacts and reproducibility

The full source code, the synthetic cohorts (saved in the same on-disk format as the real data) and the per-fold JSON files behind every metric reported here are released in the project's public repository at `https://github.com/Miquel44/MS_MedicalDataAugmentation_TFG`, whose `README` documents the exact commands that regenerate the factor-1 benchmark and the balanced sweep (a few hours on an RTX 4060 GPU). Each completed study is additionally frozen under `Experiments/Experiment_`n`/` with its own configuration, logs and result tables.