

Received 17 November 2025, accepted 26 November 2025, date of publication 28 November 2025,  
date of current version 11 December 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3638875

## RESEARCH ARTICLE

# Exploiting Graph Convolutional Networks for Insightful Classification and Explanation of Traumatic Brain Injury

**TIZIANA CURRIERI<sup>1</sup>**, **JOAN FALCÓ-ROGET<sup>2</sup>**, **ELHAM ROSTAMI<sup>3,4</sup>**,  
**SALVATORE VITABILE<sup>1</sup>**, **AND ALESSANDRO CRIMI<sup>5</sup>**

<sup>1</sup>Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University of Palermo, 90127 Palermo, Italy

<sup>2</sup>Sano Centre for Computational Medicine, 30-054 Kraków, Poland

<sup>3</sup>Department of Medical Sciences, Neurology, Uppsala University, 753 10 Uppsala, Sweden


<sup>4</sup>Department of Neuroscience, Karolinska Institutet, 171 77 Stockholm, Sweden

<sup>5</sup>AGH Science and Technology, 31-120 Kraków, Poland

Corresponding author: Alessandro Crimi (alecrimi@agh.edu.pl)

This work was supported in part by the Ministry of University and Research (MUR) as part of the FSE REACT-EU-PON 2014–2020 “Research and Innovation” Resources—Green/Innovation Action-DM MUR 1061/2021; in part by European Union-Next Generation EU-Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) 2022, PNRR Prot. P20222MYKE; in part by the Project “A Pilot analysis of behavioral and operational data for detecting socio-emotional precursors of mild cognitive impairments (MCI) and dementia (IRRESPECTIVE)” under Grant B53D23025980001; in part by the Project INNOVA “Italian network of excellence for advanced diagnosis” Italian Complementary National Plan (PNC) to the National Recovery and Resilience Plan (PNRR) under Grant CUP B73C22001770006; in part by the Minister of Science and Higher Education “Support for the Activity of Centers of Excellence Established in Poland under Horizon 2020” under Contract MEiN/2023/DIR/3796; in part by European Union’s Horizon 2020 Research and Innovation Program under Grant 857533; and in part by the Sano Project Carried Out within the International Research Agendas Programme of the Foundation for Polish Science co-financed by European Union under European Regional Development Fund.

**ABSTRACT** Understanding how brain connectivity is disrupted across stages of traumatic brain injury (TBI) is essential for improving diagnosis and treatment. TBI poses major challenges in clinical assessment, requiring advanced neuroimaging and machine learning (ML) for effective patient stratification. This study classifies TBI patients into acute, chronic, and control groups using graph convolutional networks (GCNs) applied to structural connectomes derived from diffusion-weighted imaging (DWI). To enhance interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to identify brain regions contributing to classification. The dataset included 40 participants: 18 acute TBI patients (Glasgow Coma Scale  $\leq 6$ , enrolled after  $\geq 24$  hours of unresponsiveness), 6 chronic patients with persistent disorders of consciousness, and 16 healthy controls. Nine acute patients who regained consciousness were later included in the chronic group to assess longitudinal changes. The GCN was trained on DWI-derived connectomes and evaluated using leave-one-subject-out (subject-wise) cross-validation. It achieved 83.67% accuracy, with precision, recall, and F1-score of 81.6%, 78%, and 79%, respectively, reported as per-fold averages. Grad-CAM identified thalamus, anterior cingulate cortex, and frontal cortex as key regions for group discrimination. Results suggest a shift from widespread neural disruption in acute TBI to more localized impairments in the chronic stage, possibly reflecting compensatory reorganization. Despite the limited sample size, the model’s robustness was supported by conservative regularization and subject-wise validation. Notably, the GCN outperformed classical ML classifiers, offering superior accuracy and greater biological plausibility. These findings support the use of GCN-based pipelines for clinical decision support and highlight their potential to inform interpretable, ML-driven strategies for precision neurorehabilitation.

The associate editor coordinating the review of this manuscript and approving it for publication was Adrian Stern .

**INDEX TERMS** Brain connectivity, brain modelling, diffusion-weighted imaging, explainable AI, grad-CAM, graph-based machine learning, neuroimaging.

## I. INTRODUCTION

Traumatic brain injury (TBI) is a major public health concern, affecting over 60 million people worldwide. While mild and moderate cases are most common, only severe TBI is typically associated with prolonged disorders of consciousness (DoC), such as the minimally conscious state (MCS) and vegetative state (VS) [1], [2]. Chronic and severe TBI cases are of particular clinical interest, as they offer insights into long-term consequences on brain connectivity. However, the classification of patients in different states of consciousness remains a clinical challenge due to the heterogeneity of presentations and underlying neural damage [3]. Traditional neuroimaging methods, including diffusion-weighted imaging (DWI) and functional MRI (fMRI), have revealed critical disconnection patterns related to DoC and neurodegeneration [4], [5]. Yet, these modalities alone are often insufficient for robust diagnostic stratification or outcome prediction. In this context, machine learning (ML) and especially graph-based approaches have emerged as powerful tools to model brain connectivity and aid clinical decision-making. Graph convolutional networks (GCNs) are particularly suited for analyzing structural connectomes, where brain regions form graph nodes and white matter connections define edges. Unlike traditional ML methods that rely on handcrafted features, GCNs operate directly on graph-structured data, preserving topological properties. When combined with attention or explanation mechanisms, GCNs also support interpretability, an essential requirement in clinical contexts. In severe TBI, imaging is not always performed routinely, especially in patients with mild symptoms, due to cost and resource constraints [6]. Nonetheless, imaging may reveal unexpected abnormalities even in low-risk patients [7], and scans often miss subtle connectivity disruptions. Thus, there is growing interest in AI-based tools to support imaging triage and improve the accuracy and efficiency of TBI assessment. The variability in the presentation of TBI in acute and chronic stages further complicates classification and requires computational models capable of discrimination and interpretation. GCNs have shown promise in this area, but most explainability techniques, such as gradient-weighted class activation mapping (Grad-CAM), have been applied in molecular and image analysis rather than in connectomics. Although Grad-CAM is widely used in convolutional neural networks (CNNs), its adaptation to graph neural networks (GNNs) remains in early stages and largely unexplored in neuroimaging applications. To address this gap, we propose a framework combining GCNs with Grad-CAM to classify TBI patients across acute, chronic, and control conditions using structural connectomes from DWI. Our approach identifies brain regions most relevant to classification, providing both diagnostic performance and biological insight. We focus in particular on patients with severe TBI and DoC, with the aim of mapping connectivity alterations across disease stages.

By integrating explainable deep learning with neuroimaging, our study contributes to the development of transparent AI tools in clinical neuroscience. The joint use of GCN and Grad-CAM on structural connectomes offers a novel perspective on brain injury analysis, supporting more interpretable and personalized TBI assessment.

## A. BACKGROUND

Deep learning techniques have significantly advanced neuroimaging by enabling the classification of neurological disorders based on brain connectivity. GCNs have been widely applied in neuroimaging to classify patients with neurological disorders based on brain connectivity patterns [8], [9], [10]. Unlike traditional ML methods, GCNs extend CNNs to graph-structured data, making them particularly suited for analyzing brain networks, where nodes represent brain regions and edges signify the strength of their connections. Several studies have demonstrated the effectiveness of GCNs in clinical applications. References [11] and [12] highlighted their utility in disease prediction, particularly in dementia and autism spectrum disorder, by capturing subtle connectivity patterns that conventional methods might overlook. Similarly, [13] applied a geometric construction based on Euclidean distance to classify Parkinson's disease, achieving high precision, while [14] used advanced GCN architectures to differentiate multiple sclerosis subtypes based on brain morphological connectivity. Reference [15] used GCNs for disease prediction in autism spectrum disorder and Alzheimer's disease, demonstrating their effectiveness in capturing disease-specific patterns in brain connectivity. Despite their success in the classification of neurological diseases, a major limitation of deep learning models in clinical settings is their lack of interpretability. The so-called "black-box" nature of these models hinders their adoption, as clinicians require transparent and interpretable outputs to trust AI-assisted decision-making [16]. Explainable AI (XAI) techniques have been developed to address this issue by providing insights into model decision-making processes. As emphasized by [17], ensuring the transparency of AI models is critical for clinical adoption, particularly in high-stakes applications such as neuroimaging. In the context of TBI, where subtle patterns of brain disconnection can have significant clinical implications, explainability is essential for understanding classification decisions and guiding treatment strategies [18].

## B. RELATED WORKS

While XAI techniques such as Grad-CAM have been widely applied in CNN-based biomedical imaging, their integration with GCNs remains an emerging field. Most prior studies have focused on applications in pharmacology and molecular biology rather than neuroimaging. For instance,

[19] introduced knowledge-embedded message-passing neural networks to improve molecular property prediction, demonstrating how explainability enhances AI-driven drug discovery. Similarly, [20] applied Grad-CAM to molecular interaction prediction, highlighting its relevance in computational chemistry. Despite its established utility in these domains, its potential for analyzing structural brain networks remains largely unexplored. Recently, researchers have adapted Grad-CAM for GCNs, enabling the visualization of how different brain regions (nodes) contribute to model predictions [21], [22]. This approach has been applied in neurological contexts to identify key brain areas involved in motor control, consciousness, and cognitive regulation [23]. However, most applications have focused on neurodegenerative diseases, leaving the explainability of graph-based models in TBI and DoC as an open challenge. Prior research has validated the effectiveness of this approach. For example, [21] extended Grad-CAM to GNNs, demonstrating its efficacy in various graph classification tasks, such as prediction of molecular properties and analysis of social networks. They adapted traditional visualization methods for graph data and showed that gradient-based techniques can highlight important substructures within graphs, providing valuable insights into model predictions. Similarly, [24] applied Grad-CAM to graph-based models to visualize discriminative features across multiple brain imaging modalities. Their work identified significant brain regions and illustrated their contribution to distinguishing between healthy controls, mild cognitive impairment, and Alzheimer's disease, offering valuable insights into the underlying neural mechanisms. The interpretability of deep learning models is particularly crucial in clinical applications, where understanding the rationale behind a model's decision can inform diagnostic processes and guide personalized treatment strategies [25]. If certain brain regions are consistently identified as significant across patients with a specific condition, they may serve as potential targets for therapeutic intervention or further investigation. Thus, the use of XAI techniques such as Grad-CAM bridges the gap between complex deep learning models and practical clinical applications, aligning with the broader movement toward interpretable and trustworthy AI systems in healthcare [26]. The adaptation of Grad-CAM for GCNs in graph-structured neuroimaging data has been explored in several studies, highlighting its relevance to our research. For instance, [27] proposed a convolutional network of spatiotemporal graphs, STGC-GCAM, which applies Grad-CAM to fMRI data to identify biomarkers of functional connectivity in Alzheimer's disease. This model enhances interpretability by highlighting key brain regions and connections, aiding in diagnosis and improving our understanding of disease progression. While STGC-GCAM targets fMRI functional connectivity with a spatiotemporal GCN for Alzheimer's disease, here we investigate diffusion-derived structural connectomes in TBI/DoC using a graph convolutional classifier

with Grad-CAM. This differs in imaging signal, predictive objective, and clinical population, positioning our study as a complementary graph-explainability analysis on structural brain networks in TBI.

To address the existing gap, our study integrates Grad-CAM with GCNs to identify key brain regions associated with different TBI states (acute, chronic, and control), thereby enhancing the interpretability of deep learning models in brain network analysis. This contribution aligns with the broader field of XAI in medical AI, demonstrating how explainability techniques can be adapted for graph-based neuroimaging models. Our findings provide a structured framework that could improve clinical understanding of TBI progression and support future research in personalized treatment strategies.

### C. CONTRIBUTION AND ORGANIZATION

The main contributions are:

- Development of an optimized GCN model for the classification of TBI patients and controls, leveraging structural connectivity features derived from diffusion imaging to distinguish between acute, chronic, and control groups.
- Integration of Grad-CAM explainability into GCNs for TBI classification, enabling the identification of the most relevant brain regions associated with different stages of TBI. This novel combination enhances interpretability in graph-based neuroimaging analysis, addressing a critical gap in XAI for structural connectivity studies.
- Systematic identification of the most affected brain regions, providing a structured framework to quantify trauma-induced neurodegeneration. This allows for a biologically meaningful interpretation of how neural connectivity disruptions evolve, supporting future research in TBI prognosis and rehabilitation strategies.

The subsequent structure of this paper is laid out as follows: Section II provides a comprehensive overview of the materials and methodologies, explaining the pre-processing steps and analytical tools used throughout the study. Section III elaborates on applying these methods specifically to the research objectives. Section IV presents the experimental results, focusing on the classification performance and insights gained from the XAI techniques. Section V discusses these results in a clinical context, emphasizing their implications and relevance. Finally, Section VI concludes the paper by summarizing key findings and suggesting potential future research avenues.

## II. MATERIALS AND METHODS

### A. DATASET

The present study used a publicly available dataset on the OpenNeuro platform (<https://openneuro.org/datasets/ds003367/versions/1.0.0>), comprising 40 participants divided into three distinct groups: acute, chronic, and control.

Scanning protocol and recruitment procedures are fully described in [28], and a brief summary is provided below. Analyses are carried out at the subject level: when a subject is set aside for evaluation, all of that subject's scans are evaluated together and are never used for any fitting step or statistic estimated from the remaining subjects. Details of the training/validation protocol are reported in Sec. III-B.

#### 1) SCANNING PROTOCOL

DWI data were acquired using a Siemens Skyra 3 Tesla scanner equipped with a 32-channel head coil. The imaging protocol employed High Angular Resolution Diffusion Imaging (HARDI) with the following acquisition parameters: voxel size of  $2 \text{ mm}^3$ , field of view (FOV) of 256 mm, repetition time of 3600 ms, and echo time of 95 ms. Sixty diffusion encoding directions were used, supplemented by 10 b0 volumes, with a b-value of  $2000 \text{ s/mm}^2$  [29].

#### 2) ACUTE GROUP

The acute group comprised 18 subjects who had sustained severe TBI. The inclusion criteria required a GCS score of 6 or less and no eye opening for at least 24 hours. Exclusion criteria included pre-existing neurological or psychiatric conditions that could confound neuroimaging analyses. Notably, two subjects succumbed to their injuries during acute hospitalization; however, their data were retained in the analyses to provide a comprehensive understanding of the acute phase of severe TBI. A subgroup of 9 subjects underwent follow-up imaging approximately five months post-initial hospitalization, having regained consciousness and allowing for longitudinal recovery evaluation.

#### 3) CHRONIC GROUP

The chronic group included 6 patients with DoC that persisted for at least five months after injury, and 9 patients who transitioned from the acute cohort. The group of 6 subjects consisted of individuals in VS and MCS, enabling the investigation of structural connectivity disruptions associated with long-term altered consciousness [30]. In particular, 4 patients were diagnosed with MCS and 2 with VS, with their Coma Recovery Scale-Revised (CRS-R) scores reflecting different levels of clinical severity. VS is characterized by wakefulness without conscious awareness, where patients show no voluntary or purposeful reactions to visual, auditory, or tactile stimuli [31]. In contrast, patients with MCS exhibit intentional behaviors but lack effective communication abilities [32]. MCS can also be divided into MCS+ (where patients demonstrate higher-level responses, such as following commands) and MCS- (with lower-level behaviors, including visual tracking or localized responses to pain) [33]. It is important to note that the chronic group, while including patients with persistent DoC, also comprises individuals who have regained consciousness after the acute phase. This distinction is crucial to understanding how the model differentiates between different clinical conditions

and to identify the most relevant brain regions involved in DoC and recovery of consciousness. Our analysis aims to determine whether the regions highlighted by the model correspond to those clinically relevant in TBI pathology, thereby providing an interpretable explanation of the obtained results.

#### 4) CONTROL GROUP

For comparative analysis, a control group of 16 healthy subjects was recruited to establish normative baseline data and mitigate potential confounding demographic variables. Participants were matched in age and sex with the acute TBI group to ensure comparability of the cohort. Across all groups, the average age was approximately 34 years ( $\pm 10 \text{ SD}$ ), and the male-to-female ratio was roughly 2:1, consistent with epidemiological patterns in severe TBI populations [28].

### B. IMAGE PREPROCESSING

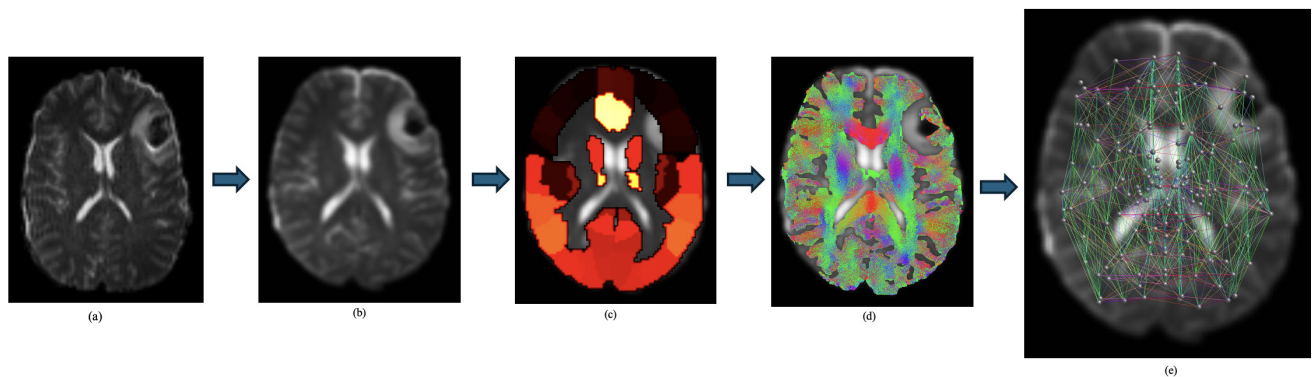
Diffusion-weighted images were preprocessed with FSL, MRtrix3, and DIPY to ensure data reliability. In this study, advanced software tools such as FSL [34], MRtrix3 [35], and DIPY [36] were used to perform the necessary corrections and enhance image quality. Given the absence of T1 and T2-weighted images in the dataset, the b0 volume extracted from the DWI was chosen as the anatomical reference. Initially, skull removal was performed using FSL's BET2 algorithm [37] to eliminate non-brain tissues and improve the delineation of brain structures. Subsequently, specific corrections were applied to mitigate magnetic artifacts and geometric distortions caused by magnetic susceptibilities [38]. To increase the signal-to-noise ratio and optimize overall image quality, denoising based on multivariate principal component analysis (MP-PCA) [39] was implemented, allowing for the reduction of intrinsic thermal noise. Additionally, Gibbs artifacts were corrected to eliminate errors from image discretization during reconstruction [40]. Furthermore, corrections were made to compensate for eddy currents, which are induced magnetic fields generated by rapid gradient changes during diffusion image acquisition. These currents can cause geometric distortions and intensity variations in DWI images, compromising data fidelity. Eliminating eddy currents, along with correcting motion artifacts, is essential to ensure accurate reconstruction of nerve fiber trajectories [41]. Images were then resampled to  $1.5 \text{ mm}^3$  isotropic and bias-field corrected [42], ensuring greater uniformity of signal intensity throughout the entire brain volume [43]. The preprocessing of DWI data involved normalizing the images to the MNI152 1.5 mm template using nonlinear registration algorithms to ensure precise alignment across subjects. The registration primarily utilized the 'mrregister' workflow from MRtrix3, which is specifically designed for DWI data. This method was chosen for its robustness in aligning DWI data with standard brain templates. Anatomical labels were extracted from the

Automated Anatomical Labeling (AAL3v1) atlas [44], which divides the brain into 170 distinct regions. These regions are nodes within the structural network, each marked for detailed neuroanatomical analysis. During registration, a subject-specific affine alignment of AAL3v1 was applied when needed, accommodating anatomical variability and ensuring accurate atlas overlays for precise label extraction. After the atlas alignment, the diffusion gradient matrix (b-matrix) was appropriately reoriented to maintain the accuracy of diffusion direction measurements post-registration [45]. This adjustment of the b-matrix, in line with the affine transformation parameters, ensured the anatomical consistency of the diffusion data, which is vital for maintaining the integrity of diffusion imaging and enhancing the reliability of subsequent fiber tractography analyses. It is important also to note that in the AAL3v1 atlas, the original identifiers for the anterior cingulate cortex (ACC) and the thalamus have been replaced with new subdivisions, leading to the omission of these regions [44]. Consequently, these unassigned areas have been interpreted as isolated nodes within the brain network, lacking interconnections with other regions. This atlas choice entails known coverage limitations for ACC and thalamic territories; we flag the potential impact on node assignment and network topology here and discuss mitigation and future atlas comparisons in the VI-A. A concise network-QC protocol (density, isolates, giant connected component (GCC); plus fixed-density proportional thresholding) was therefore included to contextualise atlas- and scale-related effects.

### C. GENERATION OF THE STRUCTURAL CONNECTOME VIA STREAMLINE TRACTOGRAPHY

Fundamental intermediate steps were performed before creating the tractography and the structural connectome. Specifically, the analysis of streamlines began by placing seeds at strategic positions along the grey matter–white matter boundary. Each streamline was generated starting from a seed, tracing a path through the white matter until terminating in another region of grey matter. Some streamlines could end in anatomically implausible positions, such as at the boundary of cerebral ventricles; such anomalous trajectories were eliminated, ensuring that most connections represented valid links between distant grey matter regions. Creating an accurate boundary between grey matter and white matter was crucial. In the absence of T1- and T2-weighted acquisitions, we used the 5TTgen segmentation with FSL workflow [35] on b0 volumes to delineate grey matter, white matter, and cerebrospinal fluid boundaries. Note that the b0 volume is a T2-weighted image and was used here as a surrogate anatomical reference in the absence of T1/T2. While the b0 is a pragmatic choice in the absence of T1/T2, it offers lower tissue contrast and is more susceptible to residual distortions than T1-weighted anatomy, which can impact parcellation fidelity near orbitofrontal and inferotemporal interfaces. We mitigated these effects with susceptibility/eddy correction

and boundary-based registration within the DWI space. We then manually corrected the segmentation maps by swapping the white matter and grey matter labels, ensuring proper tissue contrast for accurate white matter fiber reconstruction. Subsequently, a single-shell three-tissue constrained spherical deconvolution (SS3T-CSD) [46], [47] was employed using MRtrix3 to estimate the fiber orientation distribution (FOD) in each brain voxel. This technique, adapted for single-shell DWI data, facilitates the differentiation of white matter, grey matter, and cerebrospinal fluid within a single diffusion measurement, enhancing the precision of the FOD estimations. The FOD values were then normalized to equalize intensity variations across different voxels, thereby improving the consistency of the data throughout the entire brain. Anatomically constrained probabilistic tractography (ACT) [48] was used, limiting the analysis to white matter through detailed anatomical segmentations that guide the propagation of streamlines. Backtracking was implemented, allowing trajectories to retrace and recalculate the path if they terminated in implausible regions, improving the accuracy of the generated connections. Seed points were placed at the junction between white matter and grey matter, representing nerve fibers' entry and exit points. The maximum length of streamlines was set to 250 mm to avoid generating excessively long and anatomically improbable paths, and a cutoff threshold set at 0.06 defined the minimum FOD amplitude necessary to continue trajectory propagation. To refine the connectivity model, streamline filtering was implemented using SIFT2 [49], which optimizes the selection of streamlines to ensure a more accurate and representative connectivity matrix. This approach adjusts the contribution of each streamline to match the underlying MR signal, reducing biases in connectivity estimates and improving the anatomical accuracy of the connectome. A total of 10,000,000 streamlines were generated to ensure a detailed and statistically robust representation of brain connections. Using the nerve fiber trajectories a weighted connectivity matrix of size  $170 \times 170$  was constructed, corresponding to the brain regions defined in the AAL3v1 atlas. The entries in this matrix represent the total number of streamlines connecting each pair of brain regions, effectively detailing the network of connections across different areas. The values in the matrix indicate the strength and density of these neural connections, providing a comprehensive map of brain connectivity. All preprocessing, tractography, and connectome construction were executed independently per scan; no parameters estimated on one subject (or on the full dataset) were reused on another subject, and algorithmic hyperparameters (e.g., tracking cutoffs) were fixed a priori rather than fitted from cohort statistics. For quality control of network topology, we quantified per subject the native graph density (non-zero undirected edges over  $N(N-1)/2$ ), the number of isolated nodes (degree = 0 after binarisation), and the size of the GCC. To disentangle scale from topology, the same metrics were recomputed after proportional thresholding at fixed densities of 10% and 15% (retaining the top  $p\%$  of



**FIGURE 1.** Sequential image processing: (a) original diffusion-weighted image, (b) pre-processed image, (c) atlas registration (d) diffusion imaging representing fiber density, (e) connectome imaging with structural brain network mapping.

non-zero weights per subject and binarising), and summaries were derived on the resulting graphs. Zero values were observed in some matrix positions, indicating the absence of direct connections between certain regions. Subsequent analyses revealed that patients suffering from chronic conditions exhibit a higher number of isolated nodes in their brain networks, despite a lower number of subjects in this group than others. This phenomenon suggests a higher level of structural disconnection, potentially indicative of disruptions in neuronal connectivity or alterations in brain communication pathways. These results highlight the importance of deepening the study of structural connectivity in chronic patients to understand their clinical conditions' neurobiological bases better. Fig 1 delineates this analytical workflow utilized for investigating brain connectivity, encapsulating the sequence of methodologies from the initial acquisition of imaging data to the comprehensive network analysis. This illustration succinctly depicts the transition from raw DWI scans, through the enhancement of connectivity patterns, to the ultimate construction of the connectome, providing a systematic visual representation of the methodological approach.

#### D. INTRODUCTION TO GRAPHS

We model each brain as an undirected, weighted graph  $G = (V, E)$ , where  $V$  is the set of  $N$  nodes (anatomical ROIs) and  $E \subseteq V \times V$  the set of edges [50]. Edges carry weights  $w_{ij}$  that quantify structural coupling between regions  $i$  and  $j$  (e.g., streamline-derived strength). The graph is represented by the adjacency matrix  $A \in \mathbb{R}^{N \times N}$  with entries

$$A_{ij} = \begin{cases} w_{ij}, & \text{if nodes } i \text{ and } j \text{ are connected;} \\ 0, & \text{otherwise.} \end{cases}$$

This standard formulation is widely used in computational neuroimaging to analyse topological properties of brain networks and their links to cognition and pathology [51], [52], [53].

#### E. GRAPH CONVOLUTIONAL NETWORKS

GCNs are a class of neural networks designed to operate directly on graph-structured data, extending the principles of CNNs to non-Euclidean domains [54]. GCNs effectively capture the complex relationships and dependencies among nodes in a graph by aggregating and transforming feature information from a node's local neighbourhood.

In traditional CNNs, convolutional operations are defined on regular grids, such as images, where the local neighbourhood of a pixel is well-defined. However, graphs have an irregular structure, making the definition of convolution operations less straightforward. GCNs address this by generalizing the convolution operation to the graph domain, enabling the network to learn representations that consider both the nodes' features and the graph's topology. One widely adopted formulation of GCNs is based on spectral graph theory, where the convolution operation is defined in the spectral domain using the graph Laplacian. Reference [55] proposed a more computationally efficient and scalable approach, which introduced a first-order approximation of spectral graph convolutions.

The propagation rule for a single GCN layer is defined as:

$$H^{(l+1)} = \sigma \left( \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)} \right), \quad (1)$$

where:

- $H^{(l)} \in \mathbb{R}^{N \times F^{(l)}}$  is the matrix of activations at layer  $l$ , with  $F^{(l)}$  features per node.
- $H^{(0)} = X$  is the matrix of initial nodes features.
- $\tilde{A} = A + I$  is the adjacency matrix with added self-loops (where  $I$  is the identity matrix).
- $\tilde{D}$  is the degree matrix of  $\tilde{A}$ , with  $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ .
- $W^{(l)} \in \mathbb{R}^{F^{(l)} \times F^{(l+1)}}$  is the trainable weight matrix for layer  $l$ .
- $\sigma(\cdot)$  is a nonlinear activation function, such as ReLU.

This formulation performs a weighted aggregation of each node's features with those of its neighbours, normalized by the degrees of the nodes. Adding self-loops ensures that

a node's features are included in the aggregation, which is crucial for preserving the information at each node. GCNs iteratively update node representations by aggregating information from their local neighbourhoods, allowing the network to capture local and global structural information. By stacking multiple GCN layers, nodes can incorporate information from nodes several hops away, effectively learning hierarchical feature representations. In the context of brain network analysis, GCNs are compelling tools. Brain connectivity data can naturally be represented as graphs, where nodes correspond to anatomical regions of interest (ROIs) and edges represent structural connections between these regions derived from DWI.

By training the GCNs on these graphs, the network learns to generate node embeddings that capture both local connectivity patterns and global network structures. This method allows us to model complex interactions within the brain network and identify subtle connectivity changes that may indicate TBI or facilitate the discovery of patterns associated with neurological conditions. Full implementation details for the GCN architecture, training hyperparameters are reported in Section III-B.

#### F. EXPLAINABILITY WITH GRAD-CAM

In neural network models, particularly GCNs, the interpretability of model decisions is crucial, especially when applied to sensitive domains like neuroscience and clinical diagnosis. To address this, we employed XAI methods, specifically Grad-CAM, to gain insights into the decision-making processes of our trained GCN model. Although originally developed for CNNs in image classification tasks [56], [57], recent advances have adapted this technique for use with graph neural networks (GNNs), including GCNs, allowing the interpretation of model decisions in graph-structured data [58].

In our study, we employed this adapted version of Grad-CAM to elucidate the inner workings of our GCN model, which is critical given the complex and non-Euclidean nature of brain connectivity graphs. The importance of XAI in our context cannot be overstated, as it allows us to identify specific brain regions (nodes) that significantly influence the model's predictions, thereby facilitating a deeper understanding of the neurobiological underpinnings associated with different clinical conditions. By computing the gradients of the output class scores concerning the node features or embeddings, we generated class activation maps highlighting each node's contribution to the final prediction. Mathematically, for a given graph  $G$  and target class  $c$ , the Grad-CAM importance score for node  $i$  can be computed as:

$$\text{Importance}_i^c = \text{ReLU} \left( \sum_k \alpha_k^c h_i^k \right) \quad (2)$$

where  $h_i^k$  represents the activation of node  $i$  at layer  $k$ , and  $\alpha_k^c$  is the importance weight of feature map  $k$  for class  $c$ ,

calculated by globally averaging the gradients over all nodes:

$$\alpha_k^c = \frac{1}{N} \sum_{i=1}^N \frac{\partial y_c}{\partial h_i^k} \quad (3)$$

Here,  $y_c$  is the output score (logit) for class  $c$ ,  $N$  is the total number of nodes in the graph, and  $\frac{\partial y_c}{\partial h_i^k}$  denotes the gradient of  $y_c$  concerning the activation  $h_i^k$  of node  $i$  at layer  $k$ . The ReLU function ensures that only positive contributions are considered, focusing on features that positively influence the class score.

Integrating Grad-CAM into our analysis enhances the transparency of our GCN model, allowing us to correlate the most influential nodes with established neurological functions and clinical observations. This is essential for building trust in the model's predictions and potentially uncovering novel insights into the pathophysiology of DoC and TBI. The ability to interpret decisions at the node level not only validates the model against existing neuroscientific knowledge but also facilitates the exploration of new hypotheses regarding brain connectivity patterns in different clinical states. Full implementation details for the Grad-CAM computation are reported in Section III-C.

### III. GRAPH ANALYSIS AND NETWORK APPLICATIONS

#### A. GRAPH METRICS USED

In our analysis of brain network connectivity, we employed several key graph metrics to characterize the topological properties of the structural connectomes derived from DWI data. The local features selected include *degree*, *PageRank*, *betweenness centrality*, *local efficiency*, *average neighbour degree*, *weighted clustering coefficient*, *eigenvector centrality*, and *Katz centrality* [59]. Several of these descriptors are scale-invariant or bounded by construction (e.g., PageRank, eigenvector centrality, weighted clustering, local efficiency), whereas others retain the scale of the underlying edge weights. For consistency, we standardized all node-wise features via per-feature z-scoring (zero mean, unit variance) computed on the training split and applied to the corresponding evaluation split. Below, we provide definitions and mathematical formulations for each metric:

##### 1) Degree ( $k_i$ ):

The degree of a node  $i$  represents the total strength of its connections to other nodes in the network. In weighted graphs, it is calculated as:

$$k_i = \sum_{j=1}^N A_{ij}, \quad (4)$$

where  $A_{ij}$  is the weight of the edge between nodes  $i$  and  $j$ , and  $N$  is the total number of nodes [60].

##### 2) PageRank ( $PR_i$ ):

PageRank measures the influence of a node based on the concept of link analysis, originally developed for

ranking web pages. It is defined recursively as:

$$PR_i = \frac{1 - d}{N} + d \sum_{j \in M_i} \frac{PR_j}{k_j^{out}}, \quad (5)$$

where  $d$  is the damping factor (typically set to 0.85),  $M_i$  is the set of nodes linking to node  $i$ , and  $k_j^{out}$  is the out-degree of the node  $j$  [61].

3) Betweenness Centrality ( $BC_i$ ):

Betweenness centrality quantifies the importance of a node in terms of the number of shortest paths passing through it. It is given by:

$$BC_i = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}, \quad (6)$$

where  $\sigma_{st}$  is the total number of shortest paths between nodes  $s$  and  $t$ , and  $\sigma_{st}(i)$  is the number of those paths that pass through node  $i$  [62].

4) Local Efficiency ( $E_{loc}(i)$ ):

Local efficiency reflects how efficiently information is exchanged by the immediate neighbours of a node when it is removed. It is calculated as:

$$E_{loc}(i) = \frac{1}{k_i(k_i - 1)} \sum_{j, h \in N_i} (A_{jh} + A_{hj}) \times l_{jh}^{-1}, \quad (7)$$

where  $N_i$  is the set of neighbors of node  $i$ , and  $l_{jh}$  is the shortest path length between nodes  $j$  and  $h$  [63].

5) Average Neighbor Degree ( $AND_i$ ):

This metric computes the average degree of the neighbours of node  $i$ :

$$AND_i = \frac{1}{k_i} \sum_{j \in N_i} k_j, \quad (8)$$

providing insight into the connectivity of a node's immediate network [64].

6) Weighted Clustering Coefficient ( $C_i$ ):

The weighted clustering coefficient assesses the tendency of a node's neighbours to form tightly knit groups, considering the weights of the connections:

$$C_i = \frac{1}{s_i(k_i - 1)} \sum_{j, h \in N_i} \frac{(w_{ij} + w_{ih})}{2} \times w_{jh}, \quad (9)$$

where  $s_i = \sum_{j \in N_i} w_{ij}$  is the strength of node  $i$ ,  $w_{ij}$  is the weight of the edge between nodes  $i$  and  $j$ , and the sums are over all pairs of neighbors  $j$  and  $h$  of node  $i$  [65].

7) Eigenvector Centrality ( $EC_i$ ):

Eigenvector centrality assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node:

$$EC_i = \frac{1}{\lambda} \sum_{j=1}^N A_{ij} EC_j, \quad (10)$$

where  $\lambda$  is the largest eigenvalue of the adjacency matrix  $A$  [66].

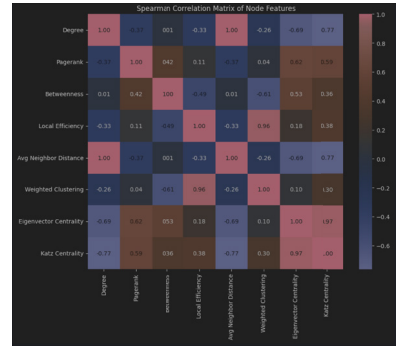


FIGURE 2. Spearman correlation matrix displaying relationships between various node features in a network analysis.

8) Katz Centrality ( $KC_i$ ):

Katz centrality measures the relative influence of a node within a network by considering the total number of walks between nodes, penalized by the length of the walks:

$$KC_i = \alpha \sum_{j=1}^N A_{ij} KC_j + \beta, \quad (11)$$

where  $\alpha$  is a damping factor (must be less than the reciprocal of the largest eigenvalue of  $A$ ), and  $\beta$  is a constant representing the initial centrality [60], [67].

Fig 2 illustrates the Spearman correlation matrix, highlighting the pairwise correlations between various graph metrics employed in our analysis. Despite significant correlations between certain metrics, such as Degree and Average Neighbor Degree, and Weighted Clustering and Local Efficiency, we retained these features in our analysis. The decision to include these metrics stems from their ability to offer complementary insights into different structural and functional aspects of the network. Furthermore, maintaining a diverse set of metrics enhances the robustness of our predictive models, enabling them to capture a broader range of underlying patterns and dynamics within the brain network. Additionally, certain correlated features are identified as statistically significant predictors in our models, providing unique and valuable contributions to our understanding of the network's complexities. While GCNs can learn node representations directly from connectivity matrices, we incorporate predefined graph-theoretic features to enhance interpretability and provide domain-specific insights into brain connectivity alterations. Handcrafted metrics such as betweenness centrality and clustering coefficient capture established topological properties linked to neurological dysfunctions in TBI and disorders of consciousness. Including these features allows the model to integrate data-driven learned representations and expert-defined metrics, improving classification performance and biological interpretability. This approach ensures that the model does not solely rely on implicit patterns learned from

training data but also leverages well-established graph-based biomarkers of neural network alterations.

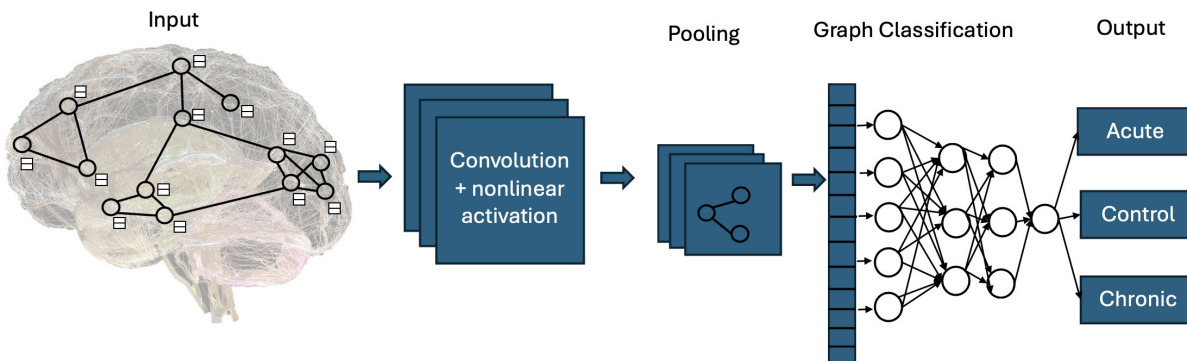
## B. GRAPH CONVOLUTIONAL NETWORK IMPLEMENTATION

Our study utilizes GCNs to analyze the structural connectomes constructed from DWI data. Each node in the graph represents a brain region defined by the AAL3v1 atlas, and edges are weighted by the strength of the structural connections between regions derived from tractography. We associated a feature vector comprising the graph metrics calculated and explained in the previous section for each node within the graph. Specifically, each node was characterized by an 8-dimensional feature vector. By assigning this vector of features to each node, we provided the GCN with rich information capturing both the local properties of nodes and their roles within the global network topology. The model consisted of three graph convolutional layers that progressively extracted high-level representations of node features while considering the graph structure. The specific architecture included a first graph convolutional layer that accepts the 8-dimensional node feature vector as input and produces 32 output features per node. The second graph convolutional layer takes the 32-dimensional node features from the previous layer and outputs 64 features per node. The third graph convolutional layer processes the 64-dimensional node features and outputs 128 features per node, as depicted in the attached figures.

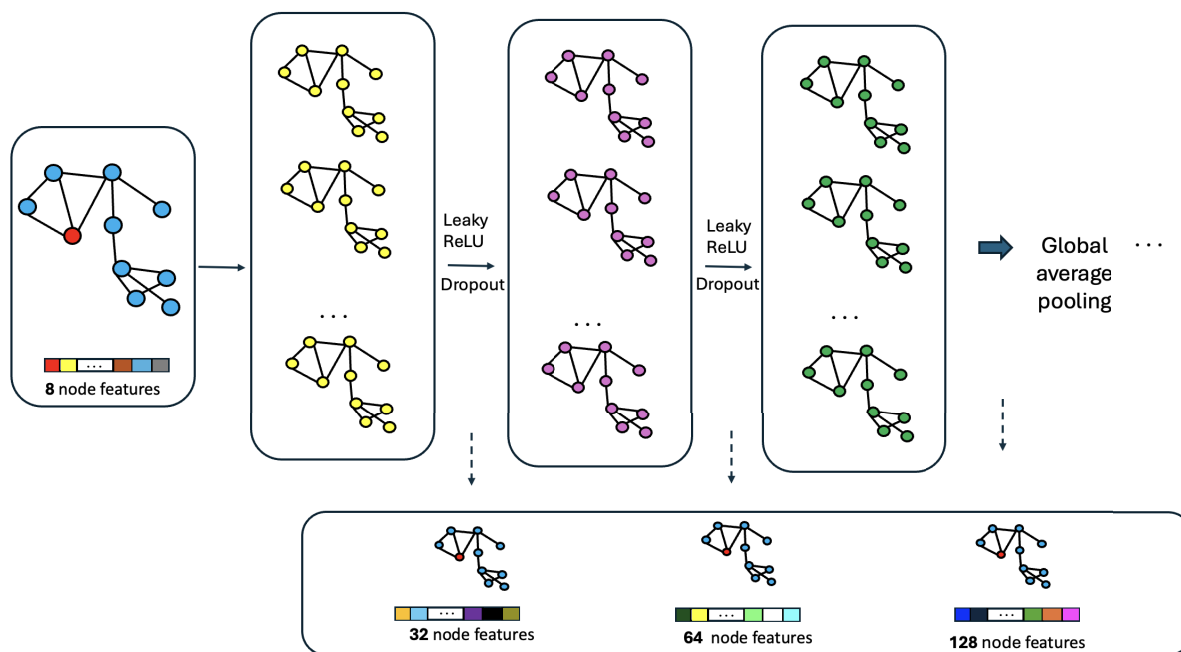
Fig 3 illustrates the overall workflow from the initial graph input through the convolution and pooling stages to the final classification output. Fig 4 provides a detailed view of the convolution and activation process, highlighting the transformation of node features through successive layers with Leaky Rectified Linear Unit (ReLU) activation and dropout, culminating in global average pooling (GAP). After each convolutional layer, the Leaky ReLU activation function with a negative slope of 0.01 was applied to introduce non-linearity. It is an activation function that allows a small, non-zero gradient when the unit is not active (i.e., when the input is negative), which helps mitigate the “dying ReLU” problem where neurons become inactive and stop learning. Batch normalization was performed after each activation to stabilize and accelerate the training process, and dropout layers with a rate of 0.5 were incorporated to prevent overfitting. Empirical results indicate that, in addition to weight decay and dropout, batch normalization and symmetric adjacency normalization ( $\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ ) further stabilize optimization under correlated inputs and heterogeneous feature scales. Following the convolutional layers, a GAP operation was used to aggregate node-level features into a graph-level representation, resulting in a 128-dimensional graph-level feature vector that captures the most prominent signals in the graph. The pooled graph representation was then passed through a series of fully connected layers: the first fully connected layer transformed the 128-dimensional input into 64 units, the second reduced the 64 units to 32 units, the

third further reduced the 32 units to 16 units, and the output layer mapped the 16 units to 3 output units corresponding to the three classes (acute patients, chronic patients, and healthy controls). Importantly, this layer did not include an activation function, outputting raw logits. The model was trained using the Adam optimizer with an initial learning rate of  $1 \times 10^{-5}$  and a weight decay of  $1 \times 10^{-4}$  to enhance performance and prevent overfitting. The loss function used was a cross-entropy loss, suitable for multiclass classification problems which expects raw, unnormalized scores (logits) as input and internally applies the LogSoftmax operation during loss computation. The training process involved 450 epochs, providing sufficient iterations for converging the model. The selection of hyperparameters was determined through systematic testing to balance model performance and generalization. The dropout rate of 0.5 was chosen after empirical validation to prevent over-reliance on specific features while maintaining sufficient learning capacity. Lower dropout values (0.3–0.4) led to slight overfitting, while higher values (above 0.6) caused excessive information loss, reducing classification accuracy. Similarly, weight decay was set at  $1 \times 10^{-4}$  to introduce mild regularization and prevent overly complex feature learning. Experiments with alternative values ( $5 \times 10^{-5}$  and  $5 \times 10^{-4}$ ) confirmed that the selected value provided the best trade-off between model flexibility and overfitting prevention.

To evaluate generalisation, we adopted a subject-wise leave-one-subject-out (LOOCV) protocol: at each fold, one subject is held out and all of that subject’s scans (acute and/or chronic) constitute the test set, while the model is trained on the remaining subjects. Graph construction and node-feature extraction for the held-out subject are performed from that subject’s adjacency alone, with no dataset-wide fitting or reuse of training statistics. During training, batch-normalisation statistics are learned on the training fold only; at test time, the model is set to evaluation mode to prevent any information flow from the test set. No external scaler or dataset-level normalisation is fit across subjects. Out-of-fold (OOF) predictions are obtained by concatenating per-fold predictions for the held-out subject. This subject-wise design maximises patient-level independence between training and test and is preferable for small, heterogeneous cohorts; at each fold, all network weights and batch-normalisation running statistics were freshly re-initialised before training. In addition to LOOCV, we ran subject-wise 5-fold stratified and nested cross-validation (outer=5, inner=3 for model selection) to probe robustness and quantify uncertainty. In small- $n$  settings, these alternatives yielded slightly lower yet concordant estimates with fold-wise variability of approximately 2–3% (numerical results in Table 3). We therefore report LOOCV as the primary estimate, favoured here for its bias, variance characteristics under limited  $n$ , and use k-fold/nested results as conservative bounds [68], [69]. Model performance was evaluated for accuracy, precision, recall and the F1 score. Ninety-five percent confidence intervals were computed via subject-level bootstrap (10 000 resamples) on



**FIGURE 3.** Overview of a graph convolutional network applied to brain connectivity data, illustrating the workflow from input graph through feature extraction and pooling, to final classification into three categories.



**FIGURE 4.** Detailed view of the convolution and nonlinear activation stages in a graph neural network, showing feature transformation from 8 node features to 128 node features through successive layers with Leaky ReLU activation and dropout, culminating in global average pooling.

out-of-fold (OOF)  $(y, \hat{y})$  pairs using percentile CIs, and permutation p-values via 10 000 two-sided label shuffles preserving class priors; permutations were conducted at the *subject* level by shuffling  $y$  to  $y^\pi$  on fixed OOF  $(y, \hat{y})$  pairs, with no model retraining. The computational experiments were conducted using Python 3.9 and PyTorch 2.2. The classification was performed on an iMac equipped with a 3.1 GHz Intel Core i5 processor (6 cores) and an AMD Radeon Pro 5300 graphics card with 4 GB of VRAM.

**C. UNVEILING KEY BRAIN REGIONS WITH GRAD-CAM**

To apply Grad-CAM to our trained GCN model and identify the most significant nodes for each patient and subsequent group, we implemented a method that leverages the gradients

of the output class scores concerning the input node features, as outlined in the previous section. We performed a forward pass through the GCN model for each patient’s brain connectivity graph to obtain the raw output logits for all classes. Focusing on the target class corresponding to the true label of the patient, we computed the gradients of the class score for the input node features by backpropagating through the network. These gradients capture the sensitivity of the model’s predictions to changes in each node’s features, effectively highlighting which nodes have the most significant impact on the classification outcome. We averaged the gradients across all feature dimensions for each node to aggregate this information into a single importance score per node. This process resulted in a scalar value for each

node, representing its overall contribution to the model's prediction for the target class. By applying the ReLU activation function to these scores, we focused on nodes that positively influence the class score, aligning with the original Grad-CAM methodology. After obtaining the importance scores, we ranked all nodes in descending order of their scores to identify the most significant ones. Grad-CAM was computed on the last GCN layer before global average pooling, backpropagating from the pre-softmax class logit  $y_c$ ; channel weights  $\alpha_k^c$  were obtained by global averaging  $\partial y_c / \partial h_{jk}^{(L)}$  across nodes, and node saliency was  $s_i^c = \text{ReLU} \left( \sum_k \alpha_k^c h_{ik}^{(L)} \right)$ , with the model in evaluation mode and maps computed per subject without any cross-subject normalisation.

After calculating the Grad-CAM importance scores, we identified the most significant nodes and edges that influenced our GCN model's predictions for each patient and group. To achieve this, we selected the top 30 nodes and edges, which constitute approximately 17.6% of the total 170 nodes in each brain connectivity graph. This threshold was determined through methodological experimentation and domain expertise. In our exploratory analysis, reducing the number of selected nodes below this threshold failed to provide a clear separation between patients within the same clinical group, hindering the identification of the most significant brain regions. Conversely, selecting more than 30 nodes often included regions of less relevance, which diluted the meaningful patterns and increased the risk of overfitting. Therefore, selecting the top 30 nodes achieved an optimal balance, providing sufficient detail to discern impactful patterns and maintaining the analysis within a manageable scope, optimizing computational efficiency and analytical precision. In fact, from a neuroscientific perspective, focusing on this subset allowed us to relate our findings to established brain networks and functions without oversimplifying the inherently complex connectivity patterns of the brain. This approach is supported by practices in the neuroimaging field, where identifying and examining key nodes with high centrality measures has been shown to enhance the understanding of network dynamics [70], [71]. For example, [70] emphasised the importance of a subset of regions in understanding the modular organisation of the brain, while [71] demonstrated that concentrating on highly interconnected and central nodes or the "rich club", provides valuable insights into the global communication efficiency of the brain. By mapping these significant nodes back to their corresponding anatomical regions using the AAL3v1 atlas, we could correlate the model's predictions with known neuroanatomical and functional properties, thereby providing a deeper understanding of the neural mechanisms underlying DoC and TBI. By strategically selecting and analyzing the top 30 nodes, we enhanced the interpretability of our findings, facilitating the identification of key brain regions that contribute to the classification of clinical conditions, thereby bridging the gap between complex deep learning models

**TABLE 1. Classification Performance Metrics of the GCN Model under subject-wise LOOCV (OOV-aggregate metrics).**

Metric	Value
Accuracy	83.67%
Precision	81.6%
Recall	78%
F1-Score	79%

and meaningful neuroscientific insights. For visualization, we employed NetworkX and Matplotlib to generate graphical representations of the brain connectivity networks. The nodes were coloured according to their importance scores and labels were assigned according to their anatomical regions. This visual representation illustrated the spatial distribution of significant brain regions and their connections, making identifying patterns and differences across patient groups easier.

## IV. RESULTS

### A. NETWORK QC AND DENSITY CONTROL

At native density, isolated nodes were observed across groups (median isolates: Acute 6.0; Chronic 6.0; Controls 4.5). Mean GCC sizes were 163.78 (Acute), 162.53 (Chronic), and 165.25 (Controls), with lower mean native density in Chronic (0.44) than in Acute (0.52) and Controls (0.53). After proportional thresholding at fixed densities, GCC sizes remained high but below full connectivity (e.g., at 10%: Acute 162.89; Chronic 161.13; Controls 164.56), indicating that isolates persist despite density matching. Overall, these summaries are consistent with greater structural disconnection in Chronic relative to Acute and Controls while maintaining group comparability under matched density.

### B. GCN CLASSIFICATION PERFORMANCE

The performance of our GCN model in classifying different patient groups was evaluated using key metrics, including accuracy, precision, recall and F1 score. In all iterations of the LOOCV procedure, the model achieved an average accuracy of 83.67%, demonstrating a high level of correctness in its classification predictions. In addition, the model achieved an average precision of 81.6%, ensuring a low false-positive rate, which is crucial in the clinical setting to avoid misclassifying healthy individuals as patients. The average recall was 78%, reflecting the model's ability to correctly identify true positives, ensuring that a substantial percentage of real patient conditions are accurately recognised. The F1 score, which harmonises precision and recall, was 79%, emphasising the model's balanced performance in identifying true positives and minimising false negatives. Uncertainty and significance on subject-level out-of-fold predictions were: ACC = 83.67% (95% CI: 73.5–93.9%) and macro-F1 (per-fold macro-average) = 79.0% (95% CI: 69.9–92.0%); permutation tests with 10 000 two-sided label shuffles preserving class priors yielded  $p_{\text{ACC}} < 0.001$  and  $p_{\text{F1}} < 0.001$ .

Table 2 reports the out-of-fold (OOV) confusion matrix obtained under subject-wise LOOCV (rows = ground truth;

columns = predicted). Because all scans from the held-out subject are evaluated together, the matrix aggregates 49 scans from 40 subjects across folds; totals reflect 18 acute, 15 chronic (including 9 longitudinal follow-ups), and 16 control scans. Accuracy computed from this aggregated OOF matrix is 83.67%. Small differences with the per-fold macro-averages (precision 81.6%, recall 78.0%, F1 79.0%) are expected because Table 1 reports *OOF-aggregate* metrics, whereas per-fold macro-averages compute metrics within each held-out fold and then average across folds.

**TABLE 2. Aggregated out-of-fold confusion matrix under subject-wise LOOCV (rows = ground truth; columns = predicted). Counts reflect 49 scans from 40 subjects; all scans from the held-out subject are evaluated together in each fold.**

	Controls	Chronic	Acute
Controls	14	2	0
Chronic	4	9	2
Acute	0	0	18

The few classification errors are predominantly concentrated along the Controls–Chronic axis (2 Controls → Chronic; 4 Chronic → Controls), with perfect separability of the Acute class. This pattern is consistent with clinical expectations: in the chronic phase, partial recovery and neuroplastic reorganization can shift global network topology towards more normative configurations, thereby increasing similarity to healthy controls and making boundaries between these two groups inherently softer. Heterogeneity within the chronic cohort (e.g., persistent DoC versus patients who regained consciousness) further broadens the distribution of chronic connectomes, while the acute connectomes remain well separated due to widespread disruption. Residual misclassifications of Chronic as Acute (2 cases) likely reflect subjects with more severe or less compensated disconnection patterns. Together, these observations support the model’s biological plausibility while highlighting the expected overlap between controls and chronic patients after long-term reorganization.

As illustrated in Table 1, the average metrics indicate the robust overall performance of the GCN model in distinguishing between different patient groups. The consistently high average metrics demonstrate the model’s overall reliability and effectiveness in leveraging complex brain connectivity patterns. Furthermore, recognizing the worst-performing fold provides valuable insights into the model’s limitations and areas for potential improvement. Analyzing the specific characteristics of the misclassified case can inform strategies to enhance the model’s robustness, such as incorporating additional data augmentation techniques, refining feature engineering processes, or exploring architectural modifications to handle diverse patient profiles better.

For completeness, we also assessed cross-protocol robustness using subject-wise 5-fold stratified and nested cross-validation (outer 5 folds with an inner 3-fold selection loop). As summarised in Table 3, both alternatives produce slightly

lower but concordant estimates across Accuracy, Balanced Accuracy, and Macro-F1 (fold-to-fold SD  $\approx 2\text{--}3\%$ ). The expected ordering LOOCV > 5-fold > nested holds for all metrics, consistent with reduced training data per fold and the greater conservativeness of nested evaluation. These results indicate that the LOOCV figures are the upper end of a stable range rather than outliers, supporting the robustness of the reported conclusions.

### 1) ABLATION: GCN DEPTH AND CONTRIBUTION OF HANDCRAFTED FEATURES

To assess the robustness of our architectural choices, we performed a subject-wise LOOCV ablation comparing the primary model (3-layer GCN with handcrafted topological features) against: (i) a structure-only variant without handcrafted features, (ii) a shallower 2-layer GCN, and (iii) a deeper 4-layer GCN. All metrics are macro-averaged across classes and computed on out-of-fold predictions under subject-wise LOOCV (each subject held out with all of its scans).

The 3-layer GCN with handcrafted features (primary model) achieves the best overall balance (ACC 0.84, macro-F1 0.79). Removing handcrafted descriptors substantially degrades performance (ACC 0.63, F1 0.52), indicating that graph-theoretic features provide complementary signal beyond raw connectivity. Increasing depth to 4 layers boosts recall (0.75) but harms precision (0.48) and F1 (0.59), consistent with over-smoothing on a small cohort; the 2-layer variant underfits (F1 0.47). Overall, the chosen 3-layer depth with handcrafted features offers the most favorable bias–variance trade-off under subject-wise evaluation (Table 4).

### 2) COMPARATIVE EVALUATION OF CLASSICAL MACHINE LEARNING MODELS

To evaluate the added value of the proposed GCN model, we conducted a comparative analysis using a range of classical ML classifiers. These models were trained on the same set of handcrafted topological features previously used in the GCN, thus ensuring a fair comparison across methods. Since the sample size was relatively small, we applied dimensionality reduction through two feature selection techniques: a filter method based on mutual information and a wrapper method using recursive feature elimination (RFE) [72]. The first method, implemented with SelectKBest [73], allowed us to retain the five most informative features; the second approach, based on RFE with logistic regression as the estimator, selected the most predictive subset through iterative elimination. Both strategies were chosen to mitigate overfitting and emphasize the most salient topological descriptors. All feature vectors were standardized before training. Each classifier was optimized via grid search combined with stratified 5-fold cross-validation, using fixed hyperparameter settings determined through empirical testing. For Random Forest, we set `n_estimators=100`,

**TABLE 3. Robustness checks (subject-wise; mean  $\pm$  SD across folds).**

Protocol	Accuracy	Balanced Acc.	Macro-F1
5-fold Stratified	78.0% ( $\pm$ 2.4%)	76.8% ( $\pm$ 2.2%)	75.8% ( $\pm$ 2.7%)
Nested (5 $\times$ 3, outer)	74.4% ( $\pm$ 2.8%)	73.2% ( $\pm$ 2.5%)	72.6% ( $\pm$ 3.0%)

**TABLE 4. Subject-wise LOOCV ablation on GCN depth and use of handcrafted features.**

Configuration	Accuracy	Precision	Recall	F1
GCN (3 layers) + handcrafted ( <i>primary</i> )	0.837	0.816	0.780	0.790
GCN (3 layers) <i>structure-only</i>	0.631	0.475	0.573	0.519
GCN (2 layers) + handcrafted	0.583	0.417	0.533	0.467
GCN (4 layers) + handcrafted	0.685	0.482	0.752	0.585

max\_depth=None, and min\_samples\_split=5. The SVM model used an RBF kernel with  $c=100$  and  $\gamma=0.001$ . Logistic Regression was configured with  $C=10$  and the `liblinear` solver. For k-Nearest Neighbors, we used `n_neighbors=3`, `weights=uniform`, and the Euclidean distance metric. XGBoost was trained with default parameters and `eval_metric` set to `mlogloss`. These configurations were selected based on their robust performance across validation folds. Ensemble models (Voting and Stacking classifiers) were constructed by combining the four base learners, with logistic regression as the meta-learner for stacking. Table 5 reports the average classification performance of each model. Among all tested configurations, the best result from classical models was obtained using XGBoost trained on features selected via mutual information, achieving 57.1% accuracy and an F1-score of 56.9%. Ensemble methods such as the Voting Classifier and Stacking Classifier yielded similar performances (55.1% accuracy), indicating that combining multiple base learners can help partially compensate for the limitations of individual models. The use of RFE generally resulted in reduced performance, except for logistic regression, which achieved its best results (55.1%) with wrapper-based selection. Despite adopting optimized pipelines and tailored feature selection, all classical methods consistently underperformed compared to the proposed GCN. As shown in Table 5, the GCN achieved significantly higher values across all metrics, including an accuracy of 83.7% and an F1-score of 79%. This marked improvement highlights the limitations of relying solely on global topological descriptors and supports the hypothesis that learning from the full connectivity structure of the brain graph provides superior discriminative power. Unlike classical models, which treat features as independent inputs, the GCN effectively captures spatial dependencies and higher-order patterns through its convolutional operations on the graph topology.

These results confirm that classical approaches, even when supported by feature selection and ensemble strategies, fail to match the performance of deep learning models designed to exploit the structural complexity of brain networks. The comparative evaluation not only reinforces the effectiveness of the GCN in classifying patients with TBI but also illustrates the importance of modeling the relational structure of connectomic data in a principled and expressive manner. In addition to classical ML baselines, we also evaluated a 3-layer GraphSAGE under the identical subject-wise LOOCV protocol, graphs, and node featureisation used for GCN (mean aggregator; 32–64–128 channels with batch normalisation, LeakyReLU and dropout; global mean pooling; MLP 64–32–16; Adam  $10^{-5}$ , weight decay  $10^{-4}$ , 450 epochs). GraphSAGE reached ACC = 55.0%, macro-Precision = 52.7%, macro-Recall = 54.6%, and macro-F1 = 53.3%, confirming that neighborhood aggregation alone provided weaker discriminative power on this cohort than the proposed GCN. This gap is consistent with the small- $n$  regime and with the need to capture higher-order spectral/topological cues beyond first-hop averaging; in practice, GCN plus handcrafted graph descriptors yielded a more favourable bias–variance trade-off than GraphSAGE on the same inputs. We also preliminarily evaluated a Graph Attention Network (GAT) using various combinations of attention heads and hidden dimensions; results were unstable and suboptimal, with accuracies between 46.15% and 53.85%, with several configurations failing to converge. Due to strong performance fluctuations across validation folds and lack of convergence in several configurations, we therefore did not report precision, recall, or F1-score for this model. These outcomes further underscore the robustness and effectiveness of the GCN architecture, which consistently outperformed both classical and alternative graph-based models in our setting. To avoid leakage, standardisation and feature selection (SelectKBest or RFE) were fit within training folds only and then applied to the corresponding test fold; all network weights and batch-normalisation statistics were re-initialised at each fold; the held-out subject’s scans never contributed to any fitting step, and evaluation was done in `eval` mode.

### C. GRAD-CAM HIGHLIGHTS: SIGNIFICANT BRAIN REGIONS BY PATIENT GROUP

Our Grad-CAM analysis was performed to identify the most significant nodes and edges in the brain networks of patients, focusing on how specific brain regions contribute to the classification of TBI. The analysis primarily focused on identifying each patient’s top 30 most influential nodes, rather than their corresponding edges. This approach was chosen

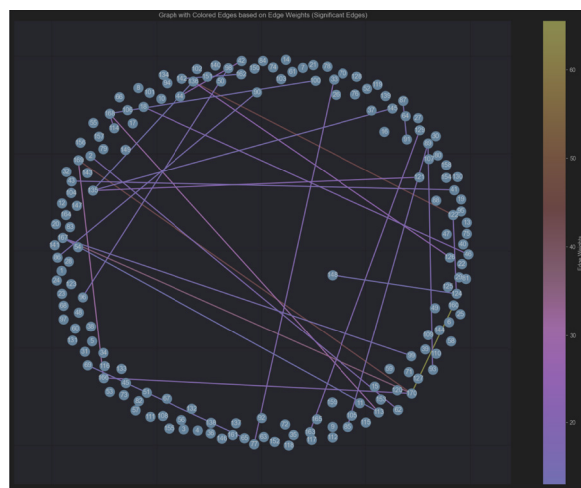
**TABLE 5.** Classification performance comparison between GCN and classical machine learning models using topological features.

Model	Accuracy	Precision	Recall	F1-score
<b>GCN (proposed)</b>	<b>0.837</b>	<b>0.816</b>	<b>0.780</b>	<b>0.790</b>
Random Forest (SelectKBest)	0.510	0.509	0.509	0.508
SVM (SelectKBest)	0.531	0.625	0.534	0.526
Logistic Regression (RFE)	0.551	0.550	0.560	0.551
XGBoost (SelectKBest)	0.571	0.570	0.570	0.569
Voting Classifier	0.551	0.574	0.548	0.555
Stacking Classifier	0.551	0.551	0.550	0.550
GAT (preliminary)	0.461–0.538	n/a	n/a	n/a
GraphSAGE	0.550	0.527	0.546	0.533

to emphasize the most critical brain regions, which are more directly relevant to understanding neural dysfunction in TBI and DoC.

Although the analysis primarily focuses on identifying the most significant nodes, Fig 5 provides a complementary visualization, showing the most important edges connecting these 170 brain regions for the acute patient depicted in Fig 1. The edges, colour-coded by weight, highlight the strength of the connections between these regions. Warmer colours indicate stronger connections, reflecting higher importance as determined by Equation 3. Notably, there is a strong edge between Raphe\_M (node 170) and VTA\_R (node 160), a connection between two critical areas involved in modulating arousal and autonomic functions, which are often disrupted in DoC [74]. Similarly, the edge between Thal\_AV\_R (node 122) and Thal\_MDm\_R (node 136) reflects robust communication within thalamic nuclei essential for sensory integration and consciousness regulation. Additionally, a prominent connection exists between Raphe\_D (node 169) and Raphe\_M (node 170), further emphasizing the role of brainstem structures in maintaining wakefulness and alertness [75], [76]. This visualization illustrates how the key regions identified in the analysis interact within the broader network, offering insights into the neural pathways that may be involved in consciousness disorders. While this study focuses on nodes, the role of these critical edges will be explored in future work.

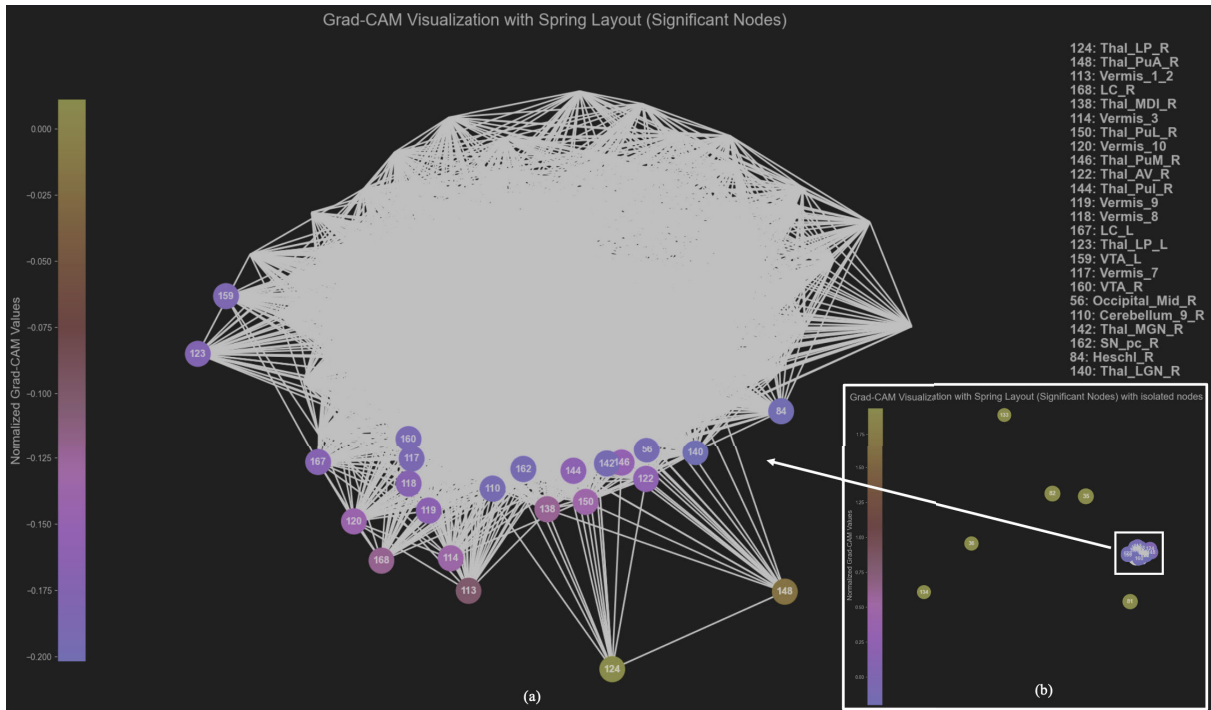
Fig 6(a) and (b) provide an example of a detailed visualization of Grad-CAM analysis applied to the same acute patient, showing the significant nodes in the network. Panel (a) displays a focused representation of the network, excluding the isolated nodes, thus emphasizing the broader connectivity and highlighting the most influential regions within a well-defined anatomical context. Each node is labelled using the AAL3v1 atlas, showcasing regions consistently identified across multiple analyses as pivotal in understanding the neuropathology associated with TBI and DoC. Conversely, panel (b) presents a general visualization that includes isolated nodes. Notably, nodes such as Cingulate\_Ant\_L (35), Cingulate\_Ant\_R (36), Thalamus\_L (81), and Thalamus\_R (82), though highly significant according to the Grad-CAM results, were not labelled by the atlas but were divided into many other areas [44]. These nodes appeared as isolated regions in the network but had a significant impact on



**FIGURE 5.** Graph showing all 170 nodes (representing brain regions) and the most significant edges for the acute patient shown in Fig 1. The edges are colour-coded based on weight, with stronger connections displayed in warmer colours. This visualization emphasizes the key connections between brain regions identified through Grad-CAM analysis.

the model's predictions, suggesting that while these nodes played a role in classification, they do not correspond to anatomically recognized areas. These isolated nodes are excluded from functional analyses due to their lack of anatomical labelling but highlight areas where the model identified critical connections. Additionally, isolated nodes such as Thal\_Re\_R (node 134) and Thal\_Re\_L (node 133) show elevated Grad-CAM values, further indicating their relevance in the network for this acute patient. The Reuniens nuclei in the thalamus coordinate communication between the prefrontal cortex and the hippocampus, which are critical for memory consolidation and cognitive functions, often disrupted in acute brain injuries [77]. These isolated thalamic nodes suggest that disruptions in these pathways could play a key role in the patient's altered consciousness and cognitive deficits. The inclusion of these isolated nodes highlights their disproportionate influence on the network, suggesting potential disruptions in critical communication pathways that could contribute to the patient's clinical symptoms.

By going into detail, in Fig 6(a) notable nodes are observed such as Thal\_LP\_R (node 124), Thal\_PuA\_R (node 148), Vermis\_1\_2 (node 113), LC\_R (node 168), Thal\_MDI\_R (node 138), VTA\_R (node 160) and Vermis\_3



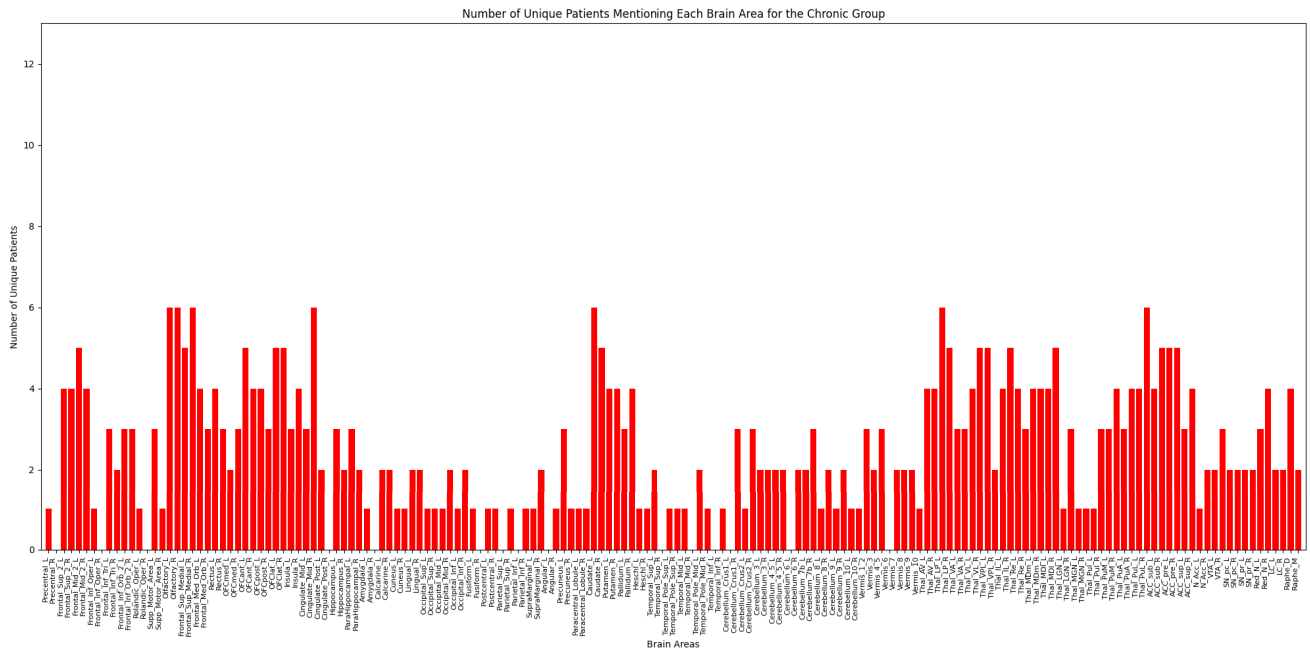
**FIGURE 6.** Panel (a) presents a focused view of the brain connectivity network, highlighting significant nodes identified using Grad-CAM, with isolated nodes excluded to emphasize the core functional interactions. Panel (b) includes the isolated nodes, marked by their strong influence on network dynamics, illustrating their critical roles despite being anatomically uncharacterized.

(node 114). These regions are critically involved in sensory processing, motor control, and the regulation of arousal, all of which are frequently disrupted in patients with severe TBI. From a clinical perspective, the thalamic nuclei, particularly Thal\_LP\_R and Thal\_PuA\_R, are crucial for relaying sensory information and maintaining attentional processes. Damage or dysfunction in these areas often leads to deficits in sensory integration and attention regulation [78], [79]. The Thal\_MDI\_R region, involved in thalamocortical signalling, is critical for higher cognitive functions, and its disruption can contribute to impaired consciousness and cognitive engagement [80]. The involvement of Vermis\_1\_2 and Vermis\_3, regions in the cerebellum, is significant due to their role in motor coordination and balance. Damage to these regions is often associated with motor impairments, such as uncoordinated movements and posture difficulties, frequently seen in patients with TBI. The cerebellum also plays a role in cognitive functions, and its dysfunction may contribute to cognitive deficits observed in these patients [81]. Additionally, LC\_R, part of the locus coeruleus, and VTA\_R, involved in the ventral tegmental area, exhibit high Grad-CAM values, indicating their importance in regulating arousal, autonomic functions, and the sleep-wake cycle. These structures are essential to maintain wakefulness, and damage to these areas can result in decreased levels of arousal or even prolonged states of unconsciousness, as commonly seen in cases of severe TBI [75], [82]. The exclusion of isolated nodes from this visualization enables a more focused examination of

the most significant interconnected regions, providing crucial insight into the disrupted neural pathways contributing to consciousness and motor impairments.

To visually represent our findings, we analyzed histograms depicting the distribution of node mentions between the three groups of patients (acute, chronic, and controls), illustrating how specific brain regions vary in their frequency of mention. These histograms show clear differentiation in the regions highlighted by each group, providing a deeper understanding of the neural patterns specific to each clinical condition. In Fig 7, we present the histogram for the acute group, showcasing the number of unique patients mentioning each brain region. The distribution shows that many regions, particularly in the frontal cortex, temporal pole, and cerebellum, are frequently mentioned across multiple patients, with several areas highlighted by more than five patients. This suggests a concentration of significant neural activity in these regions, potentially reflecting their involvement in acute phases of TBI. The variability in the number of mentions across different regions also indicates that certain areas are less frequently involved, possibly reflecting a more specialized or secondary role in the neural network dynamics of these patients. Fig 8 presents the histogram for the control group, displaying a concentrated distribution with a higher number of mentions in a smaller set of regions, particularly within the frontal cortex. The peak, around 12 mentions, suggests that these regions are crucial for maintaining stable neural function in healthy individuals.





**FIGURE 9.** Histogram showing the distribution of significant brain regions in the chronic patient group. Although the overall distribution is wider than in the controls, the mentions are less concentrated than in the acute group, indicating a potential for network stabilization and adaptation over time.

**TABLE 6.** Grouping of significant brain regions in acute patients. Region names are followed by their corresponding atlas identifiers (ID) to ensure unambiguous mapping to the AAL3 atlas.

Anatomical Area	Regions
Frontal Cortex	Frontal_Sup_2_R (ID 4), Frontal_Mid_2_R (ID 6), Frontal_Inf_Orb_2_L (ID 11), Frontal_Inf_Orb_2_R (ID 12), Frontal_Med_Orb_L (ID 21), Frontal_Med_Orb_R (ID 22), Rectus_L (ID 23), Rectus_R (ID 24), OFCmed_L (ID 25), OFCmed_R (ID 26), OFCant_L (ID 27), OFCpost_L (ID 29), OFClat_L (ID 31), OFClat_R (ID 32)
Olfactory	Olfactory_L (ID 17), Olfactory_R (ID 18)
Insula	Insula_R (ID 34)
Basal Ganglia	Caudate_L (ID 75), Putamen_R (ID 78)
Temporal Pole	Temporal_Pole_Sup_L (ID 87), Temporal_Pole_Mid_L (ID 91), Temporal_Pole_Mid_R (ID 92)
Cerebellum	Cerebellum_7b_L (ID 105), Cerebellum_7b_R (ID 106), Cerebellum_10_L (ID 111), Vermis_1_2 (ID 113), Vermis_9 (ID 119), Vermis_10 (ID 120)
Thalamus	Thal_LP_R (ID 124), Thal_Re_L (ID 133), Thal_Re_R (ID 134), Thal_MDI_L (ID 137)
Anterior Cingulate Cortex	ACC_sub_L (ID 151), ACC_sub_R (ID 152), ACC_pre_L (ID 153), ACC_pre_R (ID 154), ACC_sup_R (ID 156)
Midbrain	VTA_L (ID 159), LC_L (ID 167), LC_R (ID 168)

neural disruption has potentially diminished over time, likely due to compensatory mechanisms or neural adaptations. Furthermore, the chronic group shows greater variability in regions such as the thalamus and ACC, indicating ongoing network reorganisation that differentiates it from the control and acute groups. This is also confirmed by the observation that the chronic group includes patients with DoC and patients with recovered consciousness. By examining the nodes that Grad-CAM highlighted as significant, we aimed to identify brain regions most strongly associated with the conditions under investigation. After running the analysis across all patients, we compiled the results for each clinical group, focusing on nodes that appeared significant across multiple patients. To quantify this, we established a threshold: any node found to be important in more than five patients within the same group was considered to have clinical relevance. This threshold was selected based on its ability to

highlight nodes that played a consistent role in classification, striking a balance between capturing important regions and the small number of subjects.

As part of our comprehensive analysis, Fig. 10(A-F) offers detailed DWI visualizations of the significant brain regions identified across different patient groups. These figures complement the respective tables, illustrating the spatial distribution of these regions and how they relate to the clinical features observed in each group. Fig. 10(A) corresponds to the acute patient group, represented in Table 6, where green nodes indicate the most significant regions. In acute TBI cases, disruptions in regions like the thalamus, frontal cortex, and brainstem are critical, given their roles in motor control, consciousness, and cognition. These alterations are central to the acute manifestations of TBI, reflecting widespread network disruption and offering deeper insights into the immediate neural impacts of injury. The frontal cortex, for

**TABLE 7. Grouping of significant brain regions in the control group. Region names are followed by their AAL3 IDs.**

Anatomical Area	Regions
Frontal Cortex	Frontal_Sup_2_L (ID 3), Frontal_Sup_2_R (ID 4), Frontal_Mid_2_L (ID 5), Frontal_Inf_Tri_L (ID 9), Frontal_Inf_Orb_2_L (ID 11), Frontal_Inf_Orb_2_R (ID 12), Frontal_Sup_Medial_L (ID 19), Frontal_Sup_Medial_R (ID 20), Frontal_Med_Orb_L (ID 21), Frontal_Med_Orb_R (ID 22), Rectus_L (ID 23), Rectus_R (ID 24), OFCmed_L (ID 25), OFCmed_R (ID 26), OFCant_L (ID 27), OFCant_R (ID 28), OFCpost_L (ID 29), OFCpost_R (ID 30), OFClat_L (ID 31)
Olfactory	Olfactory_L (ID 17), Olfactory_R (ID 18)
Basal Ganglia	Putamen_L (ID 77)
Anterior Cingulate Cortex	ACC_sub_L (ID 151), ACC_sub_R (ID 152), ACC_pre_L (ID 153), ACC_pre_R (ID 154), ACC_sup_L (ID 155), ACC_sup_R (ID 156)
Nucleus Accumbens	N_Acc_L (ID 157), N_Acc_R (ID 158)

**TABLE 8. Grouping of significant brain regions in chronic patients (all). Region names are followed by their AAL3 IDs.**

Anatomical Area	Regions
Frontal Cortex	Frontal_Mid_2_L (ID 5), Frontal_Sup_Medial_L (ID 19), Frontal_Sup_Medial_R (ID 20)
Olfactory	Olfactory_L (ID 17), Olfactory_R (ID 18)
Orbitofrontal Cortex	OFCant_L (ID 27), OFClat_L (ID 31), OFClat_R (ID 32)
Cingulate Cortex	Cingulate_Mid_R (ID 38)
Basal Ganglia	Caudate_L (ID 75), Caudate_R (ID 76)
Thalamus	Thal_LP_L (ID 123), Thal_LP_R (ID 124), Thal_VL_R (ID 128), Thal_VPL_L (ID 129), Thal_IL_R (ID 132), Thal_MDL_R (ID 138), Thal_PuM_R (ID 146), Thal_PuL_R (ID 150)
Anterior Cingulate Cortex	ACC_sub_R (ID 152), ACC_pre_L (ID 153), ACC_pre_R (ID 154)

**TABLE 9. Grouping of significant brain regions in chronic patients (Minimally Conscious State). Region names are followed by their AAL3 IDs.**

Anatomical Area	Regions
Frontal Cortex	Frontal_Sup_2_R (ID 4), Frontal_Mid_2_R (ID 6), Frontal_Sup_Medial_R (ID 20)
Olfactory	Olfactory_L (ID 17), Olfactory_R (ID 18)
Orbitofrontal Cortex	OFClat_L (ID 31)
Insula	Insula_R (ID 34)
Cingulate Cortex	Cingulate_Mid_R (ID 38)
Basal Ganglia	Caudate_L (ID 75), Caudate_R (ID 76)
Thalamus	Thal_AV_L (ID 121), Thal_AV_R (ID 122), Thal_LP_L (ID 123), Thal_LP_R (ID 124), Thal_VA_R (ID 126), Thal_VL_R (ID 128), Thal_IL_L (ID 131), Thal_IL_R (ID 132), Thal_Re_L (ID 133), Thal_Re_R (ID 134), Thal_MDm_R (ID 136), Thal_MDI_L (ID 137), Thal_MDL_R (ID 138), Thal_LGN_R (ID 140), Thal_PuM_R (ID 146), Thal_PuA_R (ID 148), Thal_PuL_R (ID 150)
Anterior Cingulate Cortex	ACC_sub_R (ID 152), ACC_pre_R (ID 154)
Red Nucleus	Red_N_L (ID 165), Red_N_R (ID 166)
Raphe	Raphe_D (ID 169), Raphe_M (ID 170)

example, is heavily involved in executive functions and emotional regulation, while the basal ganglia and thalamus are crucial for motor control and consciousness regulation. The cerebellum, though traditionally associated with motor coordination, also plays a role in cognitive processing, further emphasizing the complexity of disruptions observed in acute TBI cases [80], [83]. In contrast, Fig. 10(B), corresponding to Table 7, presents the control group’s brain networks, where blue nodes represent the most significant regions. These areas, primarily located within the frontal cortex and olfactory regions, reflect a stable neural network, crucial for maintaining normal cognitive, motor, and sensory functions. The high density of significant nodes in the frontal cortex underscores its role in higher-order cognitive functions, such as attention, decision-making, and emotional regulation. The

visualization of this stable network contrasts sharply with the more disrupted and widespread network impairments seen in TBI patients. Additionally, fewer isolated regions and stronger cohesion in the control group emphasize the integrity of a healthy brain network, providing a baseline for comparison with both acute and chronic TBI patients. Fig. 10(C) visualizes the significant brain regions identified in chronic patients, corresponding to Table 8, without subgroup distinction (including those who regained consciousness, as well as patients in MCS and VS). The red nodes highlight key areas that remain affected in the chronic phase of TBI, including the frontal cortex, thalamus, and brainstem. The chronic patient group presents a more concentrated pattern of significant regions, suggesting a stabilization or reorganization of neural networks over time. However, persistent

**TABLE 10. Grouping of significant brain regions in chronic patients (Vegetative State). Region names are followed by their AAL3 IDs.**

Anatomical Area	Regions
Hippocampus	Hippocampus_L (ID 41)
Basal Ganglia	Caudate_L (ID 75), Putamen_L (ID 77), Pallidum_L (ID 79)
Thalamus	Thal_VL_L (ID 127), Thal_MDm_L (ID 135)

**TABLE 11. Grouping of significant brain regions in chronic patients who regained consciousness. Region names are followed by their AAL3 IDs.**

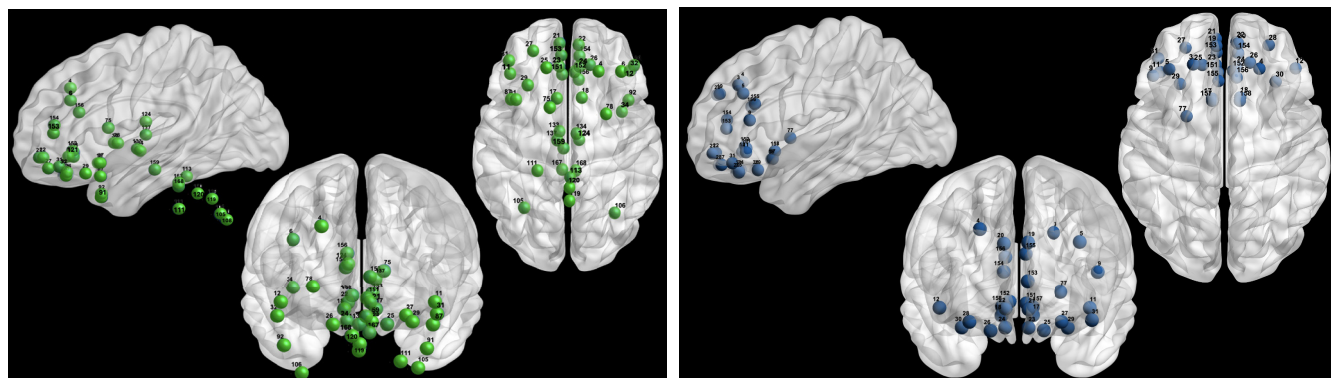
Anatomical Area	Regions
Frontal Cortex	Frontal_Mid_2_L (ID 5), Frontal_Sup_Medial_L (ID 19), Frontal_Sup_Medial_R (ID 20)
Olfactory	Olfactory_L (ID 17), Olfactory_R (ID 18)
Orbitofrontal Cortex	OFCant_L (ID 27), OFClat_L (ID 31), OFClat_R (ID 32)
Thalamus	Thal_VL_R (ID 128), Thal_VPL_L (ID 129), Thal_IL_R (ID 132), Thal_PuL_R (ID 150)
Anterior Cingulate Cortex	ACC_sub_R (ID 152), ACC_pre_L (ID 153), ACC_pre_R (ID 154)

disruptions in areas such as the ACC and thalamus continue to reflect ongoing functional impairments in consciousness and cognitive processing [84]. The visualization highlights the long-term effects of TBI, where despite recovery efforts, essential networks involved in motor control, cognition, and sensory integration remain affected. In the VS subgroup, Fig. 10(E) corresponds to Table 10, where yellow nodes mark the most significant brain regions. The visualization reveals a reduced set of significant areas, particularly concentrated in the hippocampus, basal ganglia, and thalamus, underscoring severe deficits in-memory processing, motor coordination, and sensory integration. The limited number of significant nodes reflects the widespread neural damage characteristic of VS patients, where the capacity for basic motor and sensory functions is heavily compromised. This aligns with clinical observations of profound consciousness deficits in patients in a vegetative state. For the MCS subgroup, Fig. 10(D) (pink nodes) represents the significant regions listed in Table 9. MCS patients exhibit a broader network engagement than VS patients, with key regions such as the thalamic nuclei, insula, and red nucleus showing notable involvement. The engagement of these regions suggests that despite severe impairments, MCS patients retain minimal awareness, consistent with their limited cognitive and sensory functions [85]. The broader engagement of thalamic nuclei reflects the partial retention of connectivity in critical regions associated with sensory integration and awareness, providing insights into the clinical presentation of MCS patients [86]. Lastly, Fig. 10(F) visualizes the significant brain regions for chronic patients who have regained consciousness after TBI, corresponding to Table 11, with orange nodes highlighting key regions. The visualization demonstrates the ongoing recovery of neural function, with significant areas concentrated in the frontal cortex, thalamus, and ACC. These regions are vital for executive function, sensory processing, and attention regulation, and their re-engagement suggests the brain's capacity for plasticity and recovery, though residual challenges persist [87]. The ability of these patients to regain some level of consciousness reflects

a partial restoration of network integrity, particularly in regions governing higher-order cognition and awareness. In all these visualizations, calculating the most significant nodes varies specifically for the VS, MCS, and conscious subgroups, reflecting their distinct patient characteristics. For these subgroups, regions with mentions in more than two patients per node for VS and MCS patients, and more than three patients for those who regained consciousness, were considered significant. However, for the other groups, the threshold remained as regions mentioned in more than five patients, ensuring consistency in identifying key nodes across larger patient populations. This adjusted threshold allowed for a more nuanced analysis, particularly given the smaller subgroup sizes (4 MCS patients and 2 VS patients), compared to the stringent threshold of five patients used in Table 11. The comparison of these visualizations reveals the evolution of neural network disruptions from acute to chronic stages, with acute patients showing widespread network impairment, and chronic patients displaying more localized but persistent disruptions. Meanwhile, VS patients exhibit severe and widespread damage, while MCS patients show a broader but less severe pattern of engagement, reflecting minimal awareness. The network appears to re-engage patients who regained consciousness, highlighting the ongoing restoration of brain function, albeit incomplete.

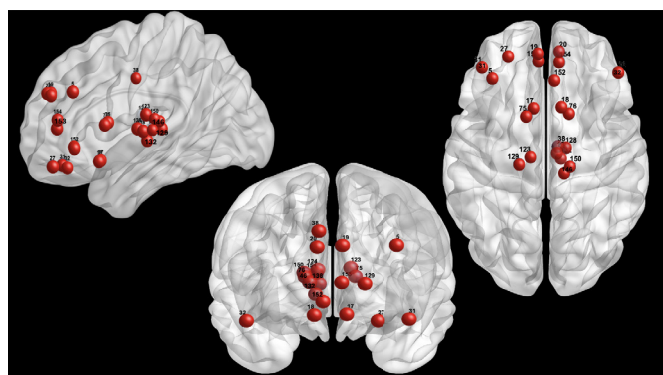
#### 1) OBJECTIVE VALIDATION OF GRAD-CAM AND CLINICAL CORRELATIONS

As a post-hoc sanity check against random saliency, Grad-CAM group-level findings were benchmarked against null maps obtained by within-subject permutation of node saliencies (1,000 draws); the observed overlaps/enrichments reported below exceed the random baselines where indicated by the permutation  $p$ -values. To move beyond visual inspection, we quantified (i) the internal consistency of Grad-CAM-derived salient-node lists within each clinical group, (ii) the enrichment of a priori regions of interest (ROIs; thalamus, ACC, frontal cortex) in the Grad-CAM mass, and (iii) the association between Grad-CAM ROI mass/fraction

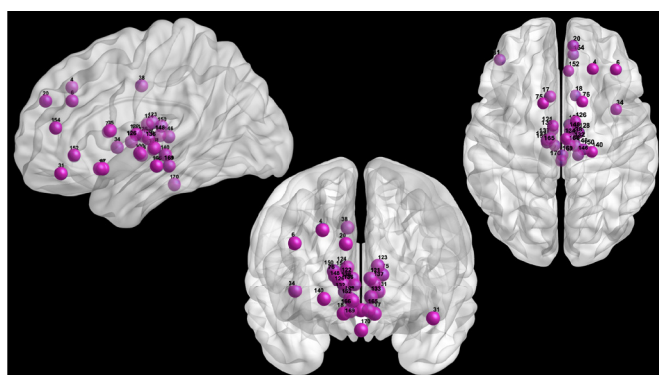


(A) Acute. Some nodes appear outside the cortical sheet because several acute ROIs are subcortical/cerebellar; positions are true MNI centroids. See Table 6.

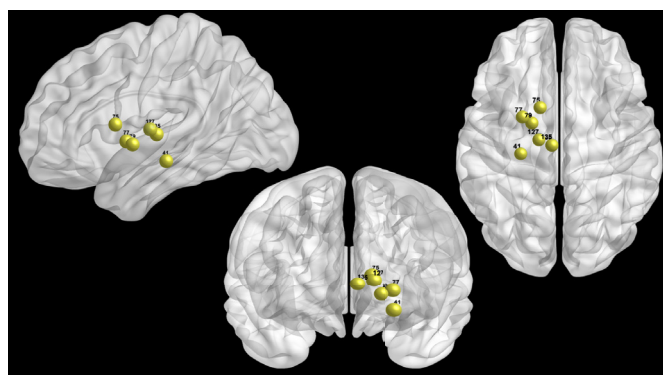
(B) Controls. See Table 7.



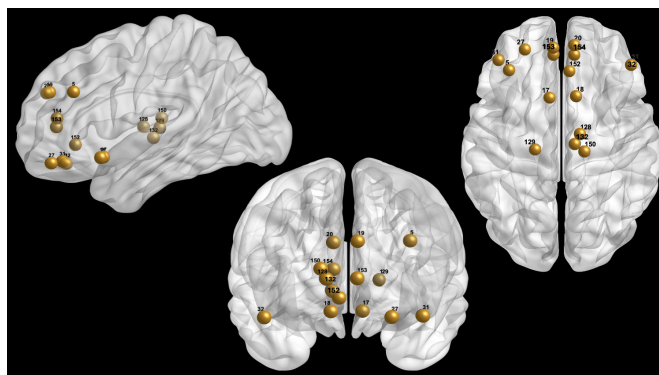
(C) Chronic (all). See Table 8.



(D) Minimally Conscious State (MCS). See Table 9.



(E) Vegetative State (VS). See Table 10.



(F) Recovered consciousness. See Table 11.

**FIGURE 10.** Composite DWI-based visualization of the top-30 network nodes for each cohort, rendered with BrainNet Viewer [88]. Panels show a consistent 3-view layout (left sagittal, coronal cut, axial view). Spheres are unweighted (same radius); colours encode cohorts (*Acute*=green; *Controls*=blue; *Chronic-all*=red; *VS*=yellow; *MCS*=pink; *Recovered*=orange). Black numeric labels on each sphere are the atlas ROI IDs and match exactly the entries reported in the corresponding tables. Apparent differences in sphere shading/transparency reflect depth and lighting effects in the 3D rendering, not weighting. The cortical surface used for display is the ICBM152 cortex; subcortical and cerebellar ROIs are nevertheless plotted at their true MNI coordinates. For this reason, some nodes, particularly in the *Acute* panel, may appear “outside” the cortical sheet: these are subcortical/cerebellar loci correctly positioned in 3D space but not intersecting the cortical mesh.

and clinical scores where available. Intra-group consistency, estimated by pairwise overlap among subjects’ top-30 nodes using the Jaccard index and Rank-Biased Overlap (RBO,  $p = 0.9$ ), showed moderate stability within groups (*Acute*: Jaccard 0.162, RBO 0.281; *Chronic*: 0.129, 0.148; *Controls*: 0.237, 0.153; Table 13). A permutation-based enrichment analysis compared the observed Grad-CAM mass fraction

within each ROI against a null model; observed mean ROI fractions and one-sided permutation  $p$ -values are reported in Table 14. In the chronic cohort with available CRS-R ( $n = 6$ ), Spearman correlations between Grad-CAM ROI metrics and CRS-R showed a positive association for ACC mass ( $\rho = 0.759, p = 0.080$ ) and positive trends for frontal mass/fraction ( $\rho = 0.647, p = 0.165$ ), whereas

**TABLE 12.** Spearman correlations between Grad-CAM ROI metrics and CRS-R in chronic patients ( $n = 6$ ).

Grad-CAM metric	$\rho$	$p$ -value
ACC mass	0.759	0.080
Frontal fraction	0.647	0.165
Frontal mass	0.647	0.165
Thalamus fraction	-0.395	0.439
ACC fraction	0.395	0.439
Thalamus mass	-0.395	0.439

**TABLE 13.** Within-group stability of top-30 Grad-CAM nodes (pairwise mean Jaccard and RBO).

Group	Mean Jaccard	Mean RBO
Acute	0.162	0.281
Chronic	0.129	0.148
Controls	0.237	0.153

thalamic mass/fraction displayed negative associations ( $\rho = -0.395$ ,  $p = 0.439$ ) (Table 12). Clinically, greater ACC and prefrontal weighting is compatible with better command-following, attention allocation, and goal-directed behaviour captured by the CRS-R, as these territories anchor salience/executive control mechanisms that support recovery of purposeful responsiveness. Conversely, higher thalamic weighting in lower CRS-R cases is consistent with persistent thalamo-cortical dysregulation typical of severe disorders of consciousness, in which impaired relay/association nuclei limit effective cortical integration. Taken together, these effect sizes are directionally in line with expected meso-circuit dynamics in recovery from severe TBI: progressive re-engagement of anterior cingulo-frontal systems in patients with higher behavioural responsiveness, and a relative predominance of subcortical drivers where responsiveness remains limited. While the cohort with clinical scores is necessarily small, the associations provide a quantitative bridge between model saliency and bedside metrics, complementing the enrichment and stability analyses and supporting the biological plausibility of the Grad-CAM maps without relying on visual inspection alone.

## V. DISCUSSION

### A. CLINICAL AND METHODOLOGICAL IMPLICATIONS

The results of the GCN model, combined with the insights derived from the Grad-CAM analysis, provide a comprehensive understanding of the neural mechanisms underlying TBI. The high classification accuracy of 83.67%, along with the identification of significant brain regions, demonstrates the potential of graph-based deep learning for distinguishing between different patient groups and offers meaningful clinical insights into the pathophysiology of these disorders. In particular, the model's strong precision and recall scores highlight its reliability in classifying patients based on their brain connectivity patterns, an essential feature for accurate diagnosis and patient stratification in clinical settings. This has important implications for clinical practice,

**TABLE 14.** Permutation-based enrichment of Grad-CAM mass in *a priori* ROIs. *obs\_mean\_pct*: observed mean fraction of Grad-CAM mass in ROI; *expected\_node\_prevalence*: proportion of atlas nodes in that ROI. One-sided  $p$  tests enrichment over null.

Group	ROI	obs_mean_pct	$p$ (perm.)	expected
Acute	Thalamus	0.402	0.565	0.188
Chronic	Thalamus	0.393	0.551	0.188
Controls	Thalamus	0.179	0.931	0.188
Acute	ACC	0.197	0.998	0.047
Chronic	ACC	0.175	0.494	0.047
Controls	ACC	0.201	0.663	0.047
Acute	Frontal	0.180	0.695	0.129
Chronic	Frontal	0.085	0.754	0.129
Controls	Frontal	0.234	0.568	0.129

where accurate diagnosis and stratification of patients with TBI are critical for determining treatment pathways. These results reinforce the potential of the GCN model as a robust classification tool for distinguishing between TBI patient groups based on brain connectivity patterns. The high accuracy and precision suggest that such an approach could contribute to the development of reliable diagnostic tools, personalized treatment strategies, and improved patient monitoring. Additionally, the balanced recall ensures that a substantial proportion of true positive cases are correctly identified, which is particularly crucial in clinical settings where misclassification could impact treatment decisions.

To further assess the model's generalization capability and rule out potential overfitting, we examined the training and validation loss curves for 450 epochs. The loss curves showed consistent convergence, with no significant divergence between training and validation loss, suggesting that the model does not suffer from memorization effects. Additionally, the LOOCV strategy ensured that each test sample remained entirely unseen during training, providing a robust evaluation framework that inherently prevents overfitting. Moreover, the worst-performing fold achieved an accuracy of 20%, indicating that the model encountered challenging cases but did not systematically overfit to training data. If the model had been overfitting, it would be expected to have near-perfect training accuracy coupled with consistently poor validation performance, which was not observed. Instead, the average validation accuracy remained stable, reinforcing the model's ability to generalize across different subjects. Furthermore, the inclusion of dropout layers and weight decay regularization, as described in III-B, effectively mitigated the risks of overfitting by preventing the model from relying too heavily on specific features. This is further supported by the high F1 score of the model, which balances precision and recall, indicating that the classifier makes correct predictions, and avoids excessive bias towards any single class. These findings underscore the importance of model interpretability, demonstrating that AI-driven approaches can maintain high accuracy and robustness across diverse patient profiles. The model not only captures meaningful connectivity patterns but does so with consistent reliability across different patient populations, making it a viable tool for clinical

use. The robustness of the model suggests its applicability as a clinical decision-support tool, potentially aiding in the development of diagnostic frameworks, personalized treatment strategies, and patient monitoring systems. The precision of 81.6% is noteworthy, especially in a clinical context where distinguishing accurately among the three patient classes is crucial. This precision indicates that the model effectively minimizes the number of false positives, guaranteeing consistent classification performance across patient groups. Similarly, the recall rate of 78% suggests that the model successfully identifies true positive cases, which is key to ensuring that patients with impaired consciousness are correctly diagnosed and managed. Moreover, the ability to quantify disconnection patterns in brain networks provides an objective measure of structural damage progression, supporting clinical decision-making. By mapping structural connectivity disruptions, the model could assist in tracking disease progression and tailoring rehabilitation strategies. In fact, the pipeline is conceived to operate at two clinically relevant junctures. First, in the acute/sub-acute window, the DWI-derived connectome feeds a graph classifier that yields an interpretable subject-level score, together with Grad-CAM maps of thalamo-fronto-cingulate circuitry, intended to support early prognosis (e.g., risk of persisting disorders of consciousness at 3-6 months) alongside standard clinical assessment. Second, during rehabilitation, the same workflow is reapplied at follow-up, producing longitudinal trajectories of simple network indices (native density, number of isolates, giant-component size) and of the model score, thereby offering a consistent proxy of structural re-integration that can be read in parallel with CRS-R and therapy milestones.

However, the variability in performance of the model underscores the complexity and heterogeneity of TBI and DoC. Misclassified cases may correspond to patients with atypical connectivity profiles, highlighting the need to further refine the model to account for individual variability. In clinical settings, such variability may reflect diverse injury severities or compensatory mechanisms, emphasizing the importance of personalized treatment strategies based on specific neural profiles. Moreover, the model's strong generalization across different subjects suggests that it is not prone to overfitting, further supporting its clinical applicability. The LOOCV strategy ensured that each test sample remained entirely unseen during training, providing a robust evaluation framework. This performance stability suggests that GCN-based models can be used in clinical settings with diverse patient profiles while maintaining a high level of accuracy. These attributes establish the GCN model as a valuable tool for clinical applications, particularly in TBI patient management and rehabilitation. With regard to data requirements for clinically robust deployment, the envisaged extension centres on prospective, multi-centre cohorts pairing DWI with structural T1 and resting-state fMRI, plus prospectively collected clinical metadata (e.g., CRS-R, time-to-command following) at acute, sub-acute

(4-8 weeks) and chronic (3-6 months) time-points. Such cohorts enable parsimonious dual-stream or late-fusion GCNs, structural and functional encoders combined at the graph level, while incorporating clinical covariates as graph-level attributes. This design preserves transparent, reportable outputs (risk score, stability summaries, saliency overlays) yet allows explicit tests of structure-function coupling and its evolution under rehabilitation. A comparative analysis with classical ML models confirmed the superior performance of the proposed GCN approach. Despite using optimized pipelines and dedicated feature selection techniques (e.g., SelectKBest and RFE), classical classifiers such as Random Forest, SVM, and XGBoost failed to exceed 57.1% accuracy. This performance gap underscores the limitations of relying solely on global topological descriptors for patient classification. Furthermore, a preliminary evaluation using a GAT architecture yielded unstable results, with accuracies fluctuating between 46.15% and 53.85%. These findings demonstrate that the GCN model not only outperforms classical and alternative graph-based approaches, but also offers more robust and consistent predictions across folds. By capturing higher-order spatial dependencies through convolutional operations on the brain graph, the GCN proved to be particularly effective in modeling the complex structural alterations associated with TBI.

## **B. INTERPRETATION OF BRAIN REGIONS IDENTIFIED BY GRAD-CAM**

This section discusses the biological significance of the brain regions identified through Grad-CAM, relating the model's predictions to known neural circuits involved in consciousness, sensory integration, and executive function, to enhance the clinical interpretability of the results. Grad-CAM analysis improves the interpretability of these classification results by highlighting the most influential brain regions that contribute to the model's decisions. In acute TBI patients, key regions such as the thalamus, brainstem, and frontal cortex emerged as critical connectivity hubs, aligning with known disruptions in motor control, consciousness, and cognitive regulation. The involvement of the thalamus in acute TBI reflects the widespread sensory and cognitive deficits observed in these patients, while disruptions in the frontal cortex are aligned with impairments in executive functions and emotional regulation. Identification of the locus coeruleus (LC) and the ventral tegmental area (VTA) underscores the role of brain stem structures in maintaining wakefulness and arousal, functions critically affected in TBI [89]. These findings suggest that interventions targeting these regions, such as neurostimulation, could play a role in modulating recovery trajectories in patients with severe TBI. In chronic TBI patients, the Grad-CAM analysis revealed a more localized pattern of disruption, with persistent impairments in the anterior cingulate cortex (ACC) and thalamus. This shift from widespread acute damage to localized chronic impairments may indicate compensatory reorganization over time, but also suggests ongoing deficits in higher-order

cognitive functions and sensory processing. Clinically, this highlights the potential for long-term interventions to support neuroplasticity and promote recovery in these critical regions. The comparison between patients with MCS and VS provides further insight into the mechanisms of consciousness. Patients with MCS exhibited a greater involvement of the thalamic and insular regions, suggesting preserved, but impaired, connectivity in key sensory and cognitive circuits. In contrast, the more restricted set of significant regions in VS patients, particularly within the hippocampus, basal ganglia, and thalamus, underscores the extensive neural damage in this group. These findings align with clinical observations of VS patients, whose capacity for voluntary motor and sensory functions is severely compromised [31]. The identification of significant brain regions in patients who regained consciousness further underscores the role of neuroplasticity. Key areas, including the frontal cortex, thalamus, and ACC, were consistently highlighted as crucial for recovery, suggesting that their re-engagement is a biomarker of neural improvement. These findings reinforce the clinical potential of using structural brain connectivity as an objective measure of recovery trajectories, allowing for stratified rehabilitation approaches. Notably, the positive ACC/prefrontal association with CRS-R is consistent with the restoration of executive/salience control during recovery. Insights from the control group provide a baseline for comparison with patient groups. The significant brain regions identified in the controls, particularly in the frontal cortex, olfactory regions, and ACC, represent typical connectivity patterns associated with normal cognitive, motor, and sensory functions. The high density of significant nodes in the frontal cortex highlights its role in executive functions, attention, and emotional regulation. In contrast, patients with TBI exhibit a more fragmented and less cohesive network, reflecting widespread neural disintegration. These findings emphasize the relevance of structural brain connectivity in distinguishing patient groups and highlight key neural regions affected by TBI progression. Beyond demonstrating strong classification performance, our model provides practical applications for clinical decision-making. By differentiating between acute, chronic, and control groups based on structural connectivity, this approach offers an additional tool for patient stratification. In clinical practice, such a model could assist in identifying patients at higher risk of prolonged disorders of consciousness or those most likely to benefit from rehabilitation. Moreover, integrating Grad-CAM enables the visualization of key brain regions contributing to the classification, offering clinicians an interpretable output that aligns with established neuroimaging findings. This transparency is critical for fostering trust in AI-assisted diagnostics, as it allows clinicians to understand and validate the model's reasoning. The ability to map specific brain regions involved in recovery trajectories suggests potential applications in personalized medicine, where treatment strategies could be adapted based on individual brain network alterations.

Overall, the integration of classification performance and Grad-CAM provides a reliable tool for distinguishing patient groups while offering clinically interpretable insights. Identifying key brain regions deepens our understanding of TBI-related disruptions and opens avenues for developing targeted interventions such as neurostimulation and rehabilitation strategies. Furthermore, the automatic quantification of disconnection patterns offers a novel approach for tracking neurodegeneration over time, allowing for objective monitoring of disease progression and therapeutic response. AI-driven brain connectivity analysis strengthens the role of ML in individualized care, ensuring that therapeutic strategies are tailored to each patient's specific neural profile. Although the limited sample size represents a constraint inherent to studies involving complex clinical populations, the methodological precautions taken, including systematic hyperparameter optimization, strong regularization, subject-wise LOOCV, and explainable AI analysis, substantiate the robustness and generalizability of our approach. We believe that this work provides a solid foundation for future investigations and practical integration into clinical neuroinformatics pipelines.

## VI. CONCLUSION

This study demonstrates the potential of GCNs combined with Grad-CAM for classifying TBI patients, with a particular focus on individuals with DoC. The model successfully differentiates between acute, chronic, and control groups, highlighting how disruptions in structural connectivity contribute to impaired consciousness and cognitive deficits. The identification of key regions such as the thalamus, frontal cortex, and brainstem reinforces their central role in TBI pathology and underscores the potential of graph-based AI models in patient stratification and diagnostic decision-making. Beyond achieving high classification accuracy, this work provides clinically interpretable insights into the neuroanatomical substrates affected by trauma, bridging the gap between ML and neuroimaging. The integration of explainability techniques ensures that the model's predictions align with established neuroscientific findings, offering greater transparency in AI-driven TBI analysis. The ability to pinpoint critical brain regions involved in recovery trajectories also suggests potential applications in personalized medicine, where connectome-based biomarkers could assist in treatment planning.

### A. LIMITATIONS AND FUTURE WORKS

Despite the promising results, several limitations must be acknowledged. The primary constraint of this study is the sample size, which remains relatively small. As with any small-sample study, model evaluation is inherently more variable; accordingly, we adopted conservative regularization and subject-wise splits to reduce, but not entirely eliminate, the generic risk of overfitting. Given also the high heterogeneity of TBI presentations, a larger and more diverse dataset would enhance the robustness of the model and improve

generalizability. Future research should aim to incorporate additional datasets from different clinical sources, allowing the model to account for a wider range of patient profiles and increasing its applicability in various clinical settings. In this study we did not perform multi-seed repetitions; future work will include a seed-sensitivity analysis (multiple random initializations under identical splits and budgets) to quantify between-seed variability. To more rigorously assess generalisation and severity grading, a multicentre expansion including mild-to-moderate TBI is envisaged, using subject-level splits stratified by site, basic cross-site harmonisation, and both pooled and per-site reporting. A further consideration is the inclusion of longitudinally overlapping participants (acute scans with later chronic follow-ups), which can introduce dependence between group distributions and potential bias in group-level summaries. As a robustness check, future extensions will provide complementary analyses that exclude follow-up scans and adopt per-subject aggregation to assess the stability of the reported findings.

A methodological consideration concerns the choice of parcellation. We adopt AAL3v1 as an anatomically grounded atlas with comprehensive cortical-subcortical coverage that is widely used in structural-connectome studies and supports stable node definitions and literature comparability. By contrast, many higher-resolution parcellations (e.g., Schaefer-200/400 [90], Desikan-Killiany [91]) are functionally derived and optimised for fMRI, prioritising functional homogeneity and yielding numerous small parcels whose boundaries do not necessarily align with white-matter tract topology; when ported to diffusion graphs, this can amplify isolate counts and increase the sensitivity of degree- and path-based metrics to local partitioning. Within this structural-connectivity setting, our primary analyses therefore privilege an anatomically based atlas at a moderate resolution, while ensuring full transparency by reporting exact AAL3 region identifiers (IDs). Another crucial aspect is the heterogeneity of TBI cases, which introduces variability in brain network disruptions and recovery trajectories. Although our model successfully identifies discriminative regions in acute, chronic, and control groups, further refinements are needed to ensure that these classifications generalise across broader populations, possibly grouping by location of the lesion. A promising avenue for future research is the integration of multimodal data. While this study focuses on structural connectivity, incorporating fMRI and clinical data could enhance classification performance and provide a more comprehensive understanding of brain network disruptions. To this end, we will investigate multimodal GCNs that jointly leverage structural edges and fMRI-derived functional connectivity, using simple dual-stream or late-fusion encoders, to probe structure-function coupling alongside key clinical covariates. The inclusion of cognitive scores, electrophysiological measures, or clinical biomarkers could help refine patient stratification and improve the interpretability

of results. Related clinical decision-support work based on spirometry and surface EMG shows how algorithmic analyses of physiological signals can be operationalized into actionable reports, offering a practical template for translating connectome-derived markers into workflow-ready outputs [92]. Additionally, longitudinal studies represent a key direction for further exploration. Although this work provides insights into patients in both acute and chronic phases, tracking how neural connectivity evolves would improve our understanding of neuroplasticity and recovery mechanisms. We will therefore consider dynamic graph models that treat each subject's follow-up as a sequence of connectomes, enabling time-resolved analysis of connectivity changes and their relationship with recovery trajectories. The subset of patients who regained consciousness after five months highlights the value of follow-up imaging in identifying recovery biomarkers. Future studies should investigate how connectivity changes over time, aiming to develop predictive models capable of forecasting patient outcomes and guiding personalized rehabilitation strategies. As a possible future extension, hyperparameter tuning may rely on reproducible random/Bayesian search with nested subject-level cross-validation and stability-oriented group feature selection; additionally, a lightweight bandit-style reinforcement-learning scheme could be explored on a compact search space to automate configurations while controlling model complexity. We did not adopt reinforcement learning in the present study because the modest sample size and the need for a fixed, pre-specified search space and reward function would make bandit policies difficult to validate reproducibly and prone to overfitting; we therefore prioritised transparent, data-efficient baselines. From a methodological perspective, while Grad-CAM was effective in identifying the most influential brain regions, further research should explore alternative explainability techniques. Methods such as Layer-wise Relevance Propagation and SHAP values could complement Grad-CAM by providing a more detailed interpretation of model decisions at both the node and network levels. In parallel, we will consider edge-aware explanations to capture node-edge interactions at the graph level, complementing node-wise Grad-CAM with simple edge-importance visualisations and basic stability checks across folds. Beyond these approaches, a promising direction for enhancing trust in AI-driven neuroimaging is the adoption of counterfactual explanations, which allow for a more transparent exploration of model decisions. Counterfactual methods provide hypothetical input conditions under which a classification outcome would change, offering an intuitive understanding of model behaviour. This technique is particularly relevant in clinical decision-making, where the ability to simulate alternative neural connectivity patterns could help clinicians assess potential recovery trajectories. Recent studies have emphasized the importance of counterfactual explanations in building trustworthy AI models [93], while graph-based approaches such as CLARUS have

demonstrated their applicability in analyzing neural networks with explainability in mind [94]. Future research should explore counterfactual reasoning within GCN-based models, allowing for a more interactive and interpretable AI system that aligns with the needs of clinicians and neuroscientists.

In conclusion, while this study presents promising advances in the classification of TBI and DoC patients, addressing limitations related to dataset size, patient heterogeneity, and the lack of multimodal integration will be essential for improving model robustness. In addition, longitudinal analysis and advanced explainability techniques can further enhance the clinical relevance of the model, contributing to a deeper understanding of the neural mechanisms that underlie brain injury and recovery. Future research should focus on these directions to refine the model and ensure its applicability in real-world clinical scenarios.

## REFERENCES

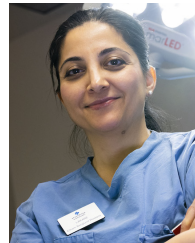
- [1] M. C. Dewan, A. Rattani, S. Gupta, R. E. Baticulon, Y.-C. Hung, M. Punchak, A. Agrawal, A. O. Adeleye, M. G. Shrima, A. M. Rubiano, J. V. Rosenfeld, and K. B. Park, "Estimating the global incidence of traumatic brain injury," *J. Neurosurgery*, vol. 130, no. 4, pp. 1080–1097, Apr. 2019.
- [2] A. I. R. Maas et al., "Traumatic brain injury: Progress and challenges in prevention, clinical care, and research," *Lancet Neurol.*, vol. 21, no. 11, pp. 1004–1060, 2022.
- [3] S. O'Brien, K. Metcalf, and J. Batchelor, "An examination of the heterogeneity of cognitive outcome following severe to extremely severe traumatic brain injury," *Clin. Neuropsychologist*, vol. 34, no. 1, pp. 120–139, Jan. 2020.
- [4] A. L. Lee, "Advanced imaging of traumatic brain injury," *Korean J. Neurotrauma*, vol. 16, no. 1, p. 3, 2020.
- [5] Y. Osmanhoğlu, D. Parker, J. A. Alappatt, J. J. Gugger, R. Diaz-Arrastia, J. Whyte, J. Kim, and R. Verma, "Connectomic assessment of injury burden and longitudinal structural network alterations in moderate-to-severe traumatic brain injury," *Human brain mapping*, vol. 43, no. 13, pp. 3944–3957, 2022.
- [6] K. K. Nagy, K. T. Joseph, S. M. Krosner, R. R. Roberts, C. L. Leslie, K. Dufty, R. F. Smith, and J. Barrett, "The utility of head computed tomography after minimal head injury," *J. Trauma: Injury, Infection, Crit. Care*, vol. 46, no. 2, pp. 268–270, Feb. 1999.
- [7] K. S. Quayle, D. M. Jaffe, N. Kuppermann, B. A. Kaufman, B. C. P. Lee, T. S. Park, and W. H. McAlister, "Diagnostic testing for acute head injury in children: When are head computed tomography and skull radiographs indicated?" *Pediatrics*, vol. 99, no. 5, p. e11, May 1997.
- [8] S. Han, Z. Sun, K. Zhao, F. Duan, C. F. Caiafa, Y. Zhang, and J. Solé-Casals, "Early prediction of dementia using fMRI data with a graph convolutional network approach," *J. Neural Eng.*, vol. 21, no. 1, Feb. 2024, Art. no. 016013.
- [9] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2008.
- [10] S. Mazurek, R. Blanco, J. Falcó-Roget, and A. Crimi, "Explainable graph neural networks for EEG classification and seizure detection in epileptic patients," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, 2024.
- [11] D. Coluzzi, V. Bordin, M. Walter Rivolta, I. Fortel, L. Zhang, A. Leow, and G. Baselli, "Biomarker investigation using multiple brain measures from MRI through XAI in Alzheimer's disease classification," 2023, *arXiv:2305.03056*.
- [12] X. Li, N. C. Dvornek, Y. Zhou, J. Zhuang, P. Ventola, and J. S. Duncan, "Graph neural network for interpreting task-fMRI biomarkers," in *Proc. MICCAI*, vol. 11768, Oct. 2019, pp. 485–493.
- [13] U. Saha, I. U. Ahamed, I. U. Ahamed, and A.-A. Hossain, "Graph convolutional network-based approach for Parkinson's disease classification using Euclidean distance graphs," in *Proc. 7th Int. Conf. Informat. Comput. Sci. (ICICoS)*, Jul. 2024, pp. 532–537. [Online]. Available: <https://api.semanticscholar.org/CorpusID:271935661>
- [14] E. Chen, B. Barile, F. Durand-Dubief, T. Grenier, and D. Sappey-Mariniere, "Multiple sclerosis clinical forms classification with graph convolutional networks based on brain morphological connectivity," *Frontiers Neurosci.*, vol. 17, Jan. 2024, Art. no. 1268860.
- [15] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, "Disease prediction using graph convolutional networks: Application to autism spectrum disorder and Alzheimer's disease," *Med. Image Anal.*, vol. 48, pp. 117–130, Aug. 2018.
- [16] F. Prinzi, T. Currier, S. Gaglio, and S. Vitabile, "Shallow and deep learning classifiers in medical image analysis," *Eur. Radiol. Experim.*, vol. 8, no. 1, p. 26, Mar. 2024.
- [17] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, and J. H. Holmes, "A manifesto on explainability for artificial intelligence in medicine," *Artif. Intell. Med.*, vol. 133, Nov. 2022, Art. no. 102423.
- [18] H. Zhou, L. He, B. Y. Chen, L. Shen, and Y. Zhang, "Multi-modal diagnosis of Alzheimer's disease using interpretable graph convolutional networks," *IEEE Trans. Med. Imag.*, vol. 44, no. 1, pp. 142–153, Jan. 2024.
- [19] T. Hasebe, "Knowledge-embedded message-passing neural networks: Improving molecular property prediction with human knowledge," *ACS Omega*, vol. 6, no. 42, pp. 27955–27967, Oct. 2021.
- [20] J. Shin, Y. Piao, D. Bang, S. Kim, and K. Jo, "DRPreter: Interpretable anticancer drug response prediction using knowledge-guided graph neural networks and transformer," *Int. J. Mol. Sci.*, vol. 23, no. 22, p. 13919, Nov. 2022.
- [21] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10764–10773.
- [22] F. Ferri, M. Cannariato, L. Pallante, E. A. Zizzi, and M. A. Deriu, "Explainable machine learning and deep learning models for predicting TAS2R-bitter molecule interactions," 2024, *arXiv:2406.15039*.
- [23] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.
- [24] H. Zhou, L. He, Y. Zhang, L. Shen, and B. Chen, "Interpretable graph convolutional network of multi-modality brain imaging for Alzheimer's disease diagnosis," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5. [Online]. Available: <https://api.semanticscholar.org/CorpusID:248407604>
- [25] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," *Diagnostics*, vol. 12, no. 2, p. 237, Jan. 2022.
- [26] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" 2017, *arXiv:1712.09923*.
- [27] Y. Zhang et al., "A novel spatiotemporal graph convolutional network framework for functional connectivity biomarkers identification of Alzheimer's disease," *Alzheimer's Res. Therapy*, vol. 16, no. 1, p. 60, 2024.
- [28] S. B. Snider, Y. G. Bodien, A. Frau-Pascual, M. Bianciardi, A. S. Foulkes, and B. L. Edlow, "Ascending arousal network connectivity during recovery from traumatic coma," *NeuroImage, Clin.*, vol. 28, 2020, Art. no. 102503.
- [29] S. B. Snider, Y. G. Bodien, M. Bianciardi, E. N. Brown, O. Wu, and B. L. Edlow, "Disruption of the ascending arousal network in acute traumatic disorders of consciousness," *Neurology*, vol. 93, no. 13, pp. e1281–e1287, Sep. 2019.
- [30] D. Kondziella, A. Bender, K. Diserens, W. van Erp, A. Estraneo, R. Formisano, S. Laureys, L. Naccache, S. Ozturk, B. Rohaut, J. D. Sitt, J. Stender, M. Tiainen, A. O. Rossetti, O. Gosseries, and C. Chatelle, "European academy of neurology guideline on the diagnosis of coma and other disorders of consciousness," *Eur. J. Neurol.*, vol. 27, no. 5, pp. 741–756, May 2020.
- [31] A. Magliacano, P. Liuzzi, R. Formisano, A. Grippo, E. Angelakis, A. Thibaut, O. Gosseries, G. Lambertini, E. Noé, S. Bagnato, B. L. Edlow, N. Lejeune, V. Veeramuthu, L. Trojano, N. Zasler, C. Schnakers, M. Bartolo, A. Mannini, and A. Estraneo, "Predicting long-term recovery of consciousness in prolonged disorders of consciousness based on coma recovery scale-revised subscores: Validation of a machine learning-based prognostic index," *Brain Sci.*, vol. 13, no. 1, p. 51, Dec. 2022.
- [32] E. Landsness, M.-A. Bruno, Q. Noirhomme, B. Riedner, O. Gosseries, C. Schnakers, M. Massimini, S. Laureys, G. Tononi, and M. Boly, "Electrophysiological correlates of behavioural changes in vigilance in vegetative state and minimally conscious state," *Brain*, vol. 134, no. 8, pp. 2222–2232, Aug. 2011.

- [33] M.-A. Bruno, A. Vanhauzenhuyse, A. Thibaut, G. Moonen, and S. Laureys, "From unresponsive wakefulness to minimally conscious PLUS and functional locked-in syndromes: Recent advances in our understanding of disorders of consciousness," *J. Neurol.*, vol. 258, no. 7, pp. 1373–1384, Jul. 2011.
- [34] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. J. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J. M. Brady, and P. M. Matthews, "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, pp. S208–S219, Jan. 2004.
- [35] J.-D. Tournier, R. Smith, D. Raffelt, R. Tabbara, T. Dhollander, M. Pietsch, D. Christiaens, B. Jeurissen, C.-H. Yeh, and A. Connelly, "MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation," *NeuroImage*, vol. 202, Nov. 2019, Art. no. 116137.
- [36] E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. van der Walt, M. Descoteaux, I. Nimmo-Smith, and D. Contributors, "Dipy, a library for the analysis of diffusion MRI data," *Frontiers Neuroinform.*, vol. 8, p. 8, Feb. 2014.
- [37] M. Jenkinson, "Bet2: Mr-based estimation of brain, skull and scalp surfaces," in *Proc. 11th Annu. Meeting Org. Human Brain Mapping*, vol. 17, 2005, p. 167.
- [38] J. L. R. Andersson, S. Skare, and J. Ashburner, "How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging," *NeuroImage*, vol. 20, no. 2, pp. 870–888, Oct. 2003.
- [39] J. Veraart, E. Fieremans, and D. S. Novikov, "Diffusion MRI noise mapping using random matrix theory," *Magn. Reson. Med.*, vol. 76, no. 5, pp. 1582–1593, Nov. 2016.
- [40] E. Kellner, B. Dhital, V. G. Kiselev, and M. Reiser, "Gibbs-ringing artifact removal based on local subvoxel-shifts," *Magn. Reson. Med.*, vol. 76, no. 5, pp. 1574–1581, Nov. 2016.
- [41] J. L. R. Andersson and S. N. Sotiropoulos, "An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging," *NeuroImage*, vol. 125, pp. 1063–1078, Jan. 2016.
- [42] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: Improved N3 bias correction," *IEEE Trans. Med. Imag.*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010.
- [43] J. Falcó-Roget, A. Cacciola, F. Sambataro, and A. Crimi, "Functional and structural reorganization in brain tumors: A machine learning approach using desynchronized functional oscillations," *Commun. Biol.*, vol. 7, no. 1, p. 419, Apr. 2024.
- [44] E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot, "Automated anatomical labelling atlas 3," *NeuroImage*, vol. 206, Feb. 2020, Art. no. 116189. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1053811919307803>
- [45] A. Leemans and D. K. Jones, "The B-matrix must be rotated when correcting for subject motion in DTI data," *Magn. Reson. Med.*, vol. 61, no. 6, pp. 1336–1349, Jun. 2009.
- [46] T. Dhollander, D. Raffelt, and A. Connelly, "Unsupervised 3-tissue response function estimation from single-shell or multi-shell diffusion MR data without a co-registered t1 image," in *Proc. ISMRM Workshop Breaking Barriers Diffusion MRI*, vol. 5, 2016.
- [47] T. Dhollander, R. Mito, and A. Connelly, "3-tissue compositional data analysis of developing HCP (dHCP) diffusion MRI data," *Hum. Brain Mapp.*, vol. 25, p. 498, 2019.
- [48] R. E. Smith, J.-D. Tournier, F. Calamante, and A. Connelly, "Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information," *NeuroImage*, vol. 62, no. 3, pp. 1924–1938, Sep. 2012.
- [49] R. E. Smith, J.-D. Tournier, F. Calamante, and A. Connelly, "SIIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography," *NeuroImage*, vol. 119, pp. 338–351, Oct. 2015.
- [50] J. A. Bondy and U. S. R. Murty, *Graph Theory*. Cham, Switzerland: Springer, 2008.
- [51] E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems," *Nature Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, Mar. 2009.
- [52] O. Sporns, "Structure and function of complex brain networks," *Dialogues Clin. Neurosci.*, vol. 15, no. 3, pp. 247–262, Sep. 2013.
- [53] A. Crimi, L. Giancardo, F. Sambataro, A. Gozzi, V. Murino, and D. Sona, "MultiLink analysis: Brain network comparison via sparse connectivity analysis," *Sci. Rep.*, vol. 9, no. 1, p. 65, Jan. 2019.
- [54] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond Euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.
- [55] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
- [56] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
- [57] E. Dührkoop, L. Malih, C. Erfurt-Berge, G. Heidemann, M. Przysucha, D. Busch, and U. Hübner, "Automatic classification of wound images showing healing complications: Towards an optimised approach for detecting maceration," *Stud. Health Technol. Informat.*, vol. 317, pp. 347–355, Aug. 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:272398996>
- [58] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.
- [59] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, Sep. 2010. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S105381190901074X>
- [60] M. Newman, *Networks: An Introduction*. London, U.K.: Oxford Univ. Press, 2010, doi: [10.1093/acprof:oso/9780199206650.001.0001](https://doi.org/10.1093/acprof:oso/9780199206650.001.0001).
- [61] O. Tanglay, I. M. Young, N. B. Dadio, H. M. Taylor, P. J. Nicholas, S. Doyen, and M. E. Sughrue, "Eigenvector PageRank difference as a measure to reveal topological characteristics of the brain connectome for neurosurgery," *J. Neuro-Oncology*, vol. 157, no. 1, pp. 49–61, Mar. 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:246492842>
- [62] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, p. 35, Mar. 1977.
- [63] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Phys. Rev. Lett.*, vol. 87, no. 19, Oct. 2001, Art. no. 198701.
- [64] D. Chen and H. Su, "Identification of influential nodes in complex networks with degree and average neighbor degree," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 13, no. 3, pp. 734–742, Sep. 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259825335>
- [65] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 11, pp. 3747–3752, Mar. 2004.
- [66] P. Bonacich, "Power and centrality: A family of measures," *Amer. J. Sociology*, vol. 92, no. 5, pp. 1170–1182, Mar. 1987. [Online]. Available: <https://api.semanticscholar.org/CorpusID:145392072>
- [67] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, Mar. 1953.
- [68] M. Esterman, B. J. Tamber-Rosenau, Y.-C. Chiu, and S. Yantis, "Avoiding non-independence in fMRI data analysis: Leave one subject out," *NeuroImage*, vol. 50, no. 2, pp. 572–576, Apr. 2010.
- [69] G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion, "Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines," *NeuroImage*, vol. 145, pp. 166–179, Jan. 2017.
- [70] D. S. Bassett and E. T. Bullmore, "Human brain networks in health and disease," *Current Opinion Neurol.*, vol. 22, no. 4, pp. 340–347, 2009.
- [71] M. P. van den Heuvel and O. Sporns, "Rich-club organization of the human connectome," *J. Neurosci.*, vol. 31, no. 44, pp. 15775–15786, Nov. 2011.
- [72] A. M. Priyatno and T. Widiyaningtyas, "A systematic literature review: Recursive feature elimination algorithms," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 9, no. 2, pp. 196–207, Feb. 2024.
- [73] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, Jan. 2014.
- [74] R. P. Vertes and S. B. Linley, "Efferent and afferent connections of the dorsal and median raphe nuclei in the rat," in *Serotonin and Sleep: Molecular, Functional and Clinical Aspects*, 2008, pp. 69–102.
- [75] S. J. Sara and S. Bouret, "Orienting and reorienting: The locus coeruleus mediates cognition through arousal," *Neuron*, vol. 76, no. 1, pp. 130–141, Oct. 2012.

- [76] B. L. Jacobs and E. C. Azmitia, "Structure and function of the brain serotonin system," *Physiological Rev.*, vol. 72, no. 1, pp. 165–229, Jan. 1992.
- [77] C. Varela, S. Kumar, J. Y. Yang, and M. A. Wilson, "Anatomical substrates for direct interactions between hippocampus, medial prefrontal cortex, and the thalamic nucleus reuniens," *Brain Struct. Function*, vol. 219, no. 3, pp. 911–929, May 2014.
- [78] S. M. Sherman, "Thalamus plays a central role in ongoing cortical functioning," *Nature Neurosci.*, vol. 19, no. 4, pp. 533–541, Apr. 2016.
- [79] Y. B. Saalmann, M. A. Pinsk, L. Wang, X. Li, and S. Kastner, "The pulvinar regulates information transmission between cortical areas based on attention demands," *Science*, vol. 337, no. 6095, pp. 753–756, Aug. 2012.
- [80] A. S. Mitchell and S. Chakraborty, "What does the mediodorsal thalamus do?" *Frontiers Syst. Neurosci.*, vol. 7, p. 37, Aug. 2013.
- [81] H. Baillicux, H. J. D. Smet, P. F. Paquier, P. P. De Deyn, and P. Mariën, "Cerebellar neurocognition: Insights into the bottom of the brain," *Clin. Neurol. Neurosurgery*, vol. 110, no. 8, pp. 763–773, Sep. 2008.
- [82] C. B. Saper, T. E. Scammell, and J. Lu, "Hypothalamic regulation of sleep and circadian rhythms," *Nature*, vol. 437, no. 7063, pp. 1257–1263, Oct. 2005.
- [83] E. K. Miller and J. D. Cohen, "An integrative theory of prefrontal cortex function," *Annu. Rev. Neurosci.*, vol. 24, no. 1, pp. 167–202, Mar. 2001.
- [84] E. D. Bigler, "Anterior and middle cranial fossa in traumatic brain injury: Relevant neuroanatomy and neuropathology in the study of neuropsychological outcome," *Neuropsychology*, vol. 21, no. 5, pp. 515–531, 2007.
- [85] D. Coleman, "The minimally conscious state: Definition and diagnostic criteria," *Neurology*, vol. 58, no. 3, pp. 506–507, Feb. 2002.
- [86] R. E. Cranford, "Dialogue on end-of-life decision making: What is a minimally conscious state?" *Western J. Med.*, vol. 176, no. 2, p. 129, 2002.
- [87] B. L. Edlow, J. Claassen, N. D. Schiff, and D. M. Greer, "Recovery from disorders of consciousness: Mechanisms, prognosis and emerging therapies," *Nature Rev. Neurol.*, vol. 17, no. 3, pp. 135–156, Mar. 2021.
- [88] M. Xia, J. Wang, and Y. He, "BrainNet viewer: A network visualization tool for human brain connectomics," *PLoS ONE*, vol. 8, no. 7, Jul. 2013, Art. no. e68910.
- [89] B. L. Edlow et al., "Sustaining wakefulness: Brainstem connectivity in human consciousness," *BioRxiv*, 2023.
- [90] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. T. Yeo, "Local–global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI," *Cerebral Cortex*, vol. 28, no. 9, pp. 3095–3114, Sep. 2018.
- [91] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany, "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, Jul. 2006.
- [92] A. K. Kumar, M. H. Assaf, V. Z. Groza, and E. M. Petriu, "Spirometer and sEMG respiratory patterns for clinical decision support system," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2023, pp. 1–6.
- [93] J. D. Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, A. Saranti, and A. Holzinger, "On generating trustworthy counterfactual explanations," *Inf. Sci.*, vol. 655, Jan. 2024, Art. no. 119898.
- [94] J. M. Metsch, A. Saranti, A. Angerschmid, B. Pfeifer, V. Klemm, A. Holzinger, and A.-C. Hauschild, "CLARUS: An interactive explainable AI platform for manual counterfactuals in graph neural networks," *J. Biomed. Informat.*, vol. 150, Feb. 2024, Art. no. 104600.



**JOAN FALCÓ-ROGET** received the degree in physics and biophysics with the University of Barcelona and the Autonomous University of Madrid. He is currently pursuing the Ph.D. degree with the Sano Centre for Computational doing research in functional, diffusion, and effective connectivity in clinically oriented frameworks. He is also active in the field of theoretical neuroscience and has done research in normative models of animal behaviour and in quantum computing solutions applied to computational neuroscience as well as medicine.



**ELHAM ROSTAMI** currently pursuing the Ph.D. degree. She is a Neurosurgeon and an Associate Professor with Uppsala University and Karolinska Institute. She specializes in traumatic brain injury (TBI) with a focus on understanding the underlying pathology to develop personalized treatment approaches. She has extensive experience in both clinical TBI management and research, contributing significantly to clinical guidelines and innovative therapeutic strategies. Her work spans from clinical trials to advanced neuroimaging techniques, with the aim of improving patient outcomes and bridging gaps in TBI Care. She is also a Key Figure in international collaborations addressing brain injury research.



**SALVATORE VITABILE** is currently a Full Professor with the Department of Biomedicine, Neuroscience, and Advanced Diagnostics, University of Palermo, Italy. He is the co-author of more than 200 scientific papers in referred journals and conferences. He has chaired, organized, and served as member of the organizing committee of several international conferences and workshops. His research interests include medical data processing and analysis, clinical decision support systems, specialized architecture design and prototyping, and machine and deep learning applications. He is an Associate Editor of *Human-Centric Computing and Information Sciences* journal and an Editorial Board Member of *Electronics*.



**TIZIANA CURRIERI** received the B.S. and M.S. degrees in computer engineering from the University of Palermo, Italy, in 2021, and the Ph.D. degree in biomedicine, neuroscience and advanced diagnostics with the BiND Department, in 2025. She was a Visiting Ph.D. Student with the Brain and More Laboratory, Sano Centre for Computational Medicine, Kraków, Poland. Her research interests include the development and use of models in the field of machine learning and explainable artificial intelligence, with a special focus on medical image analysis methods. The main research activities are devoted to the analysis and processing of biomedical images to support clinicians' decision-making activities.



**ALESSANDRO CRIMI** received the degree in engineering from the University of Palermo, the Ph.D. degree in machine learning applied for medical imaging from the University of Copenhagen, and the M.B.A. degree in healthcare management from the University of Basel. He was a Postdoctoral Researcher with French Institute for Research in Computer Science (INRIA), Technical School of Switzerland (ETH-Zurich), Italian Institute for Technology (IIT), and University Hospital of Zurich. He is currently a Professor with the AGH, Kraków.

...