Contrastive Pretraining for Ensemble Graph Neural Networks in Alzheimer's Disease Classification

Walter Mangione*, Alberto Gaetano Valerio*, Gabriella Casalino*, Miguel Ángel Ferrer†, Giovanna Castellano*, Moises Diaz†, Gennaro Vessio*

* Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
Email: {w.mangione, a.valerio31}@phd.uniba.it,
{gennaro.vessio, gabriella.casalino, giovanna.castellano}@uniba.it
† Institute for Technological Development and Innovation in Communications,
University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
Email: {miguelangel.ferrer, moises.diaz}@ulpgc.es

***Abstract*—Alzheimer's disease (AD) is a progressive neurodegenerative disorder that necessitates accurate early diagnosis. We propose a novel training strategy using an ensemble graph neural network (GNN) with contrastive pretraining for AD classification, utilizing atlas-based graphs from the AAL atlas to ensure anatomically meaningful node semantics. This ensemble combines various GNN architectures to capture diverse connectivity patterns, while contrastive pretraining enhances robust latent representation before supervised training. Evaluated on the publicly available ADNI dataset, our approach outperforms state-of-the-art methods across multiple metrics. Additionally, integrating PGExplainer improves model transparency by identifying disease-relevant subgraphs and contributions from specific brain regions. Our framework enhances both classification accuracy and explainability in graph-based AD diagnosis.**

***Index Terms*—Alzheimer's disease, graph neural networks, contrastive pretraining, explainable AI**

I. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder marked by cognitive decline, memory impairment, and widespread brain alterations. As the leading cause of dementia, it represents a major global health challenge, making early and accurate diagnosis critical.

Graph-based modeling has emerged as an effective paradigm for neuroimaging by representing the brain as a network of interacting regions [1]. In this context, GNNs show strong potential for AD classification [2], as they encode both regional features and connectivity patterns, outperforming vector-based approaches in capturing disease-related interactions [3]. However, despite these advances, two main challenges remain. First, single GNN models may struggle to capture the multi-scale complexity of AD progression. While ensemble learning can improve robustness and generalization, it remains underexplored in graph-based neuroimaging [4]. Second, the reliance on supervised training limits performance when labeled data are scarce. Recent work addresses this via contrastive learning, which enhances representations through self-supervised pretraining of invariant and discriminative features [5], [6].

Moreover, a key requirement for clinical adoption of AI-based decision support systems for AD is model explainability. While GNNs are highly expressive, their decisions are often opaque [7], [8]. Explainability methods like PGExplainer [9] highlight important subgraphs and node contributions, linking predictions to neurobiological factors. Providing clinicians with anatomically grounded explanations of model outputs supports informed judgment, fosters trust, and enables interactive exploration of disease mechanisms, promoting human-centered interaction in clinical decision-making.

Motivated by these considerations, we propose a contrastively pretrained ensemble of GNNs for AD classification. Brain connectivity is modeled using atlas-based graphs, where each node corresponds to a region of interest defined by the AAL atlas [10], enabling anatomically grounded and semantically interpretable representations. Unlike conventional ensembles, each GNN in the proposed framework is trained on a different and randomly masked view of the input graph, encouraging the learning of complementary feature representations. This design, combined with the integration of multiple GNN architectures, allows the ensemble to capture diverse structural and feature-based patterns, while contrastive pretraining further enhances robustness and discriminative power prior to supervised fine-tuning. Finally, graph-based explainability is incorporated through PGExplainer [9] to identify disease-relevant substructures and ROI-level contributions, aligning the proposed framework with human-centered AI principles by promoting transparency, explainability, and clinical trust.

The rest of the paper is structured as follows. Section II discusses related works. Section III introduces our proposed methodology. Section IV reports on experimental results and evaluation. Section V concludes the paper.

II. Related Work

A. Graph Neural Networks for Alzheimer's Disease

Graph neural networks have been widely adopted across multiple domains (e.g., [11]–[13]) due to their ability to model structured and relational data. In AD classification,

they are increasingly used to capture complex inter-regional interactions in brain connectivity.

Among these methods, BrainGNN introduced ROI-aware convolution and pooling to highlight salient brain regions and improve explainability [14]. Subsequent advances include attention-based models that weight brain regions and connections differently, such as the Brain Graph Attention Network (BGAN), which enhances classification performance through graph attention mechanisms [15]. More recent work extends these approaches to multimodal GNNs that integrate structural and functional data via parallel branches and late fusion [16].

Although these methods enhance representational power through complex multimodal fusion, our approach uses a diverse ensemble of GNN architectures to capture complementary structural and feature patterns without depending on additional modalities.

B. Self-Supervised and Contrastive Learning for Brain Graphs

The reliance on supervised training remains a key limitation for many GNN-based AD classification methods, especially given the scarcity of labeled neuroimaging data. To mitigate this, recent work has explored self-supervised contrastive learning to derive brain graph representations, including diffusion-based frameworks that learn robust embeddings from functional connectivity prior to classification [6].

However, these approaches are typically evaluated with a single GNN backbone, limiting the exploitation of complementary inductive biases [6]. In contrast, our method integrates contrastive pretraining within an ensemble framework, enabling multiple graph encoders to jointly benefit from self-supervised learning. This improves robustness and generalization while encouraging complementary latent representations, an aspect not explicitly explored before. Additionally, contrastive pretraining promotes invariant representations across graph views, which is crucial when brain graphs are affected by noise or missing information [5], [17].

C. Ensemble Learning and Explainability in Graph-Based Models

Ensemble learning improves stability and predictive performance in biomedical graph analysis by aggregating complementary decision patterns [18], [19]. However, its application to GNNs in neuroimaging and AD remains limited, with existing work focusing mainly on performance rather than understanding individual model contributions.

Explainability is essential for clinical adoption [20]. While intrinsically interpretable GNNs have been proposed for dementia prediction [21], they are often tied to specific architectures, limiting their ability to capture diverse patterns. In contrast, our framework combines a complementary GNN ensemble with post-hoc explainability via PGExplainer [9], avoiding retraining and reducing computational cost compared to perturbation or gradient-based methods, identifying disease-relevant edge connections by leveraging information across all models.

III. Materials and Methods

A. Dataset and Graph Representation

We used T1-weighted structural magnetic resonance imaging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [22], a large, publicly available cohort that provides longitudinal neuroimaging and clinical data. T1-weighted MRI offers high-resolution anatomical contrast, enabling accurate delineation of brain structures and gray–white matter boundaries, and is therefore well suited for morphometric and radiomic analysis.

The cohort includes subjects from three diagnostic categories: cognitively normal (CN), mild cognitive impairment (MCI), and AD. To manage computational complexity while preserving the original class distribution, we randomly selected a subset of 600 subjects, including 206 CN, 270 MCI, and 124 AD cases, with standard preprocessing procedures, including spatial normalization and motion correction.

We further parcellated each brain volume into 116 ROIs according to the AAL atlas [10]. Brain connectivity is modeled by defining edges based on physical adjacency between ROIs, determined by shared anatomical boundaries or direct voxel-level contact. Node features were extracted from structural MRI using the PyRadiomics library [23], following the configuration proposed in [24]. Specifically, 85 radiomic features were computed for each ROI, including first-order intensity statistics and texture descriptors derived from gray-level co-occurrence matrices (GLCM), gray-level run-length matrices (GLRLM), gray-level size zone matrices (GLSZM), and gray-level dependence matrices (GLDM), which characterize intensity patterns and textural heterogeneity within each region.

Based on this graph construction and feature extraction process, each subject is formally represented as a graph $G = (V, E, X)$, where V denotes the set of ROI nodes, E the edges based on anatomical adjacency, and $X \in \mathbb{R}^{|V| \times 85}$ the node feature matrix. Each graph was associated with a diagnostic label (AD, MCI, or CN), enabling supervised multi-class classification after contrastive pretraining. This representation supports learning discriminative, semantically interpretable graph embeddings by jointly modeling structural relationships and ROI-level radiomic information. An overview of the overall pipeline is shown in Fig. 1.

B. Ensemble Architecture and Contrastive Pretraining

Our framework employs an ensemble of three distinct GNN encoders to capture complementary structural patterns in brain connectivity graphs. Specifically, the ensemble comprises a Graph Convolutional Network (GCN) [25], which models local connectivity patterns, a Graph Attention Network (GAT) [26], which assigns adaptive importance to neighboring regions to emphasize informative connections, and a GraphSAGE encoder [27], which enables inductive learning through flexible neighborhood aggregation. To further promote model diversity and robustness, each encoder was trained on a different randomly masked view of the input graph, with 50% of nodes and 70% of edges masked. This stochastic

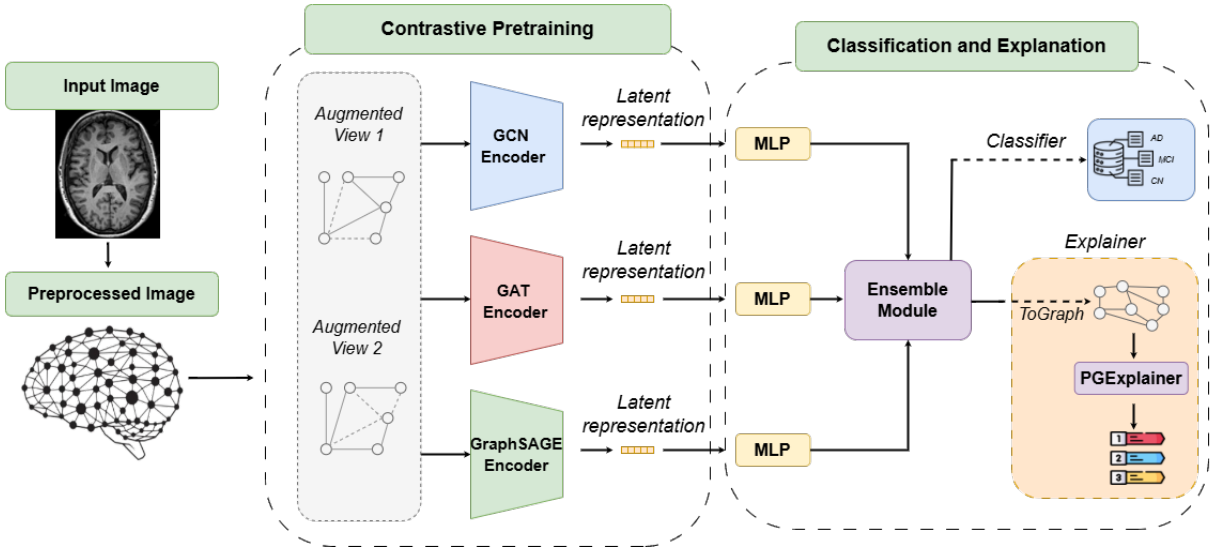

Fig. 1. Overview of the proposed pipeline for Alzheimer's disease classification. Structural MRI data are first transformed into atlas-based brain graphs using the AAL parcellation, yielding anatomically grounded node representations. During contrastive pretraining, two augmented views of each brain graph are generated and used to learn robust and invariant graph-level representations in a self-supervised manner. These representations are then processed by a complementary ensemble of GNN encoders, whose predictions are fused for supervised diagnosis. Finally, post-hoc graph explainability via PGExplainer identifies disease-relevant subgraphs and region-of-interest contributions, enabling anatomically meaningful interpretation of model decisions.

masking strategy encourages each model to focus on different substructures of the graph, reduces co-adaptation among ensemble members, and enhances the learning of complementary representations.

For a given encoder, the node embeddings are computed as

$$H^{(l+1)} = \mathrm{GNNLayer}^{(l)}(H^{(l)}, E),$$

where $H^{(0)} = X$ and $\mathrm{GNNLayer}^{(l)}$ denotes the graph convolution (GCN, GAT, or GraphSAGE) at layer l. Each layer is followed by batch normalization, ELU activation, and dropout to stabilize training. After the final layer, a global pooling operation aggregates node embeddings into a graph-level representation:

$$h_G = \mathrm{LayerNorm}\big([\ \mathrm{mean}(H^{(L)}) \,\|\, \max(H^{(L)})\ ]\big),$$

where $H^{(L)}$ denotes the output of the last graph convolutional layer, $\mathrm{mean}(\cdot)$ and $\max(\cdot)$ perform node-wise mean and max pooling, respectively, $[\cdot\|\cdot]$ denotes feature-wise concatenation, and $\mathrm{LayerNorm}(\cdot)$ applies layer normalization. Then, an attention mechanism is applied:

$$\tilde{h}_G = h_G \odot \sigma(W_{\mathrm{att}}\, h_G)\,,$$

where σ denotes the sigmoid function, W_{att} is a learnable attention weight matrix, and $\odot$ denotes element-wise multiplication. Attention allows the model to adaptively weight the graph-level representation.

The pooled and attention-weighted embedding $\tilde{h}_G$ is passed through a classifier:

$$\hat{y} = f_\theta\Big(\tilde{h}_G\Big) = W_2\,\mathrm{ELU}\Big(\mathrm{Dropout}\Big(W_1\,\tilde{h}_G\Big)\Big)\,,$$

where W_1 and W_2 are learnable weight matrices, ELU is the activation function, and dropout is applied for regularization.

To address class imbalance in our dataset, the classifier was trained with a focal loss that downweights easy examples and emphasizes hard-to-classify samples. For multi-class classification with C classes, the focal loss is defined as:

$$\mathcal{L}_{\mathrm{focal}} = -\sum_{c=1}^{C} y_c\,(1-p_c)^{\gamma}\log(p_c),$$

where $y_c \in \{0,1\}$ is the one-hot encoded ground-truth label for class c, $p_c \in [0,1]$ is the predicted probability for class c obtained via softmax, and γ is the focusing parameter. Larger values of γ emphasize hard-to-classify examples while reducing the contribution of easy ones, improving model robustness on imbalanced datasets.

For encoder training, augmented graph views were generated via random node-feature dropout, edge removal, and Gaussian noise, thereby forming positive pairs for contrastive learning. These transformations produce diverse yet semantically consistent views of the same graph, encouraging the encoders to learn robust and invariant graph representations under small perturbations. Then, each encoder was pretrained using a mini-batch contrastive learning procedure. Contrastive pretraining decouples representation learning from diagnostic supervision, enabling graph encoders to capture intrinsic structural and feature-level regularities of brain connectivity before task-specific fine-tuning.

Given a batch of graphs, two augmented views of each graph are generated using stochastic graph augmentations, including edge dropout and node-feature masking with varying augmentation strengths. Let $G_i^{(1)}$ and $G_i^{(2)}$ denote two augmented views of the same graph G_i. Both augmented graphs are passed through the same encoder f_θ to obtain graph-level

embeddings:

$$h_i^{(1)} = f_\theta\left(G_i^{(1)}\right), \quad h_i^{(2)} = f_\theta\left(G_i^{(2)}\right).$$

The encoder parameters θ are optimized minimizing a normalized temperature-scaled cross-entropy loss [28]:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp\left(\text{sim}\left(h_i^{(1)}, h_i^{(2)}\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(h_i^{(1)}, h_j^{(2)}\right)/\tau\right)},$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature hyperparameter. Embeddings from different graphs within the same batch serve as negative samples. This pretraining objective encourages the encoder to learn graph representations that are invariant to stochastic perturbations of node features and graph structure, while remaining discriminative across different graphs.

Finally, predictions from the ensemble of pretrained models are aggregated by averaging to obtain the final output:

$$\hat{y}_{\text{ensemble}} = \frac{1}{M}\sum_{m=1}^{M}\hat{y}_m,$$

where M denotes the number of models in the ensemble, which was set to three in this study. This aggregation strategy reduces prediction variance and exploits the complementary strengths of the individual models.

This architecture enables the model to leverage node-specific anatomical data and structural relationships between adjacent brain regions. At the same time, contrastive pretraining and the ensemble framework enhance its discriminative power and robustness.

IV. Experimental Evaluation

A. Baselines and Experimental Setup

All GNN encoders were trained using the Adam optimizer with a learning rate of 1×10^{-3}. Each encoder consists of three graph convolutional layers with 128 hidden units, followed by batch normalization, ELU activation, and dropout. Dropout rates were set to 0.55 for GCN and 0.3 for both GAT and GraphSAGE, with the GAT encoder using [2, 4, 1] attention heads across layers. Each encoder was pretrained for 100 epochs using the contrastive learning objective prior to supervised fine-tuning. The downstream classifier is a two-layer multilayer perceptron with 64 hidden units, ELU activation, and a dropout rate of 0.35. Optimization employed a focal loss with focusing parameter $\gamma = 2.5$, and input features were normalized to improve training stability.

Model performance was evaluated using accuracy (ACC), macro-averaged F1 score (MACRO-F1), macro-averaged AUC (MACRO-AUC), macro-averaged AUPR (MACRO-AUPR), and macro-averaged recall (MACRO-REC). All experiments were conducted using 5-fold stratified cross-validation.

The proposed framework was benchmarked against two state-of-the-art graph-based models for AD analysis: BrainGNN [14] and BGAN [15]. To ensure a fair and consistent comparison, both baselines were reimplemented and trained under the same experimental protocol.

B. Quantitative Results

Table I reports the classification performance of the proposed method in comparison with the graph-based baselines across multiple diagnostic tasks, including binary classification settings (AD vs. MCI, CN vs. MCI, and CN vs. AD) and a multi-class classification setting (CN, MCI, and AD). Overall, the proposed ensemble outperforms the competing methods across all binary classification settings, as measured by all considered evaluation metrics.

The improvements are most evident in the CN vs. MCI task, where subtle neuroanatomical patterns complicate differentiation between healthy aging and early cognitive impairment. This distinction is crucial for early diagnosis and intervention. The proposed model demonstrates significant gains across all evaluation metrics, effectively capturing disease-related connectivity patterns.

In the multi-class classification task, the proposed approach achieves the highest accuracy, macro-AUC, and macro-recall among all evaluated methods. Although BrainGNN achieves slightly higher macro-F1 and macro-AUPR scores in this context, the proposed model still demonstrates strong performance and excellent stability, as evidenced by its significantly low standard deviation.

Overall, these results suggest that the proposed ensemble improves predictive performance and yields more robust, reliable predictions across different diagnostic scenarios, thereby supporting the effectiveness of combining complementary GNN architectures with contrastive pretraining.

C. Ablation Study

Table I also reports the results of an ablation study designed to quantify the individual and combined contributions of ensemble learning and contrastive pretraining to the overall performance of the proposed framework. Specifically, we progressively evaluated (*i*) the individual GNN architectures (GCN, GAT, and GraphSAGE), (*ii*) the proposed ensemble trained without contrastive pretraining, and (*iii*) the full model integrating both ensemble learning and contrastive pretraining.

Across all binary classification tasks, the single GNN models exhibit limited and highly variable performance, indicating that relying on a single architectural bias is insufficient to capture the heterogeneous connectivity patterns associated with AD. Introducing the ensemble architecture without contrastive pretraining consistently yields substantial improvements across all evaluation metrics, demonstrating the benefit of combining complementary GNN representations. Notably, this configuration improves both discriminative performance and stability.

Further gains are observed when contrastive pretraining is incorporated. The full model achieves the best performance across all considered metrics and tasks. Similar trends are observed in the multi-class classification setting, where the ensemble substantially outperforms individual GNNs and contrastive pretraining further enhances robustness and discriminative capability. These results confirm that ensemble learning and contrastive pretraining provide complementary benefits,

TABLE I

PERFORMANCE COMPARISON AND ABLATION STUDY ACROSS DIFFERENT CLASSIFICATION TASKS, EVALUATING THE IMPACT OF ENSEMBLE LEARNING AND CONTRASTIVE PRETRAINING (CP) BY COMPARING THE PROPOSED METHOD WITH STATE-OF-THE-ART BASELINE METHODS, SINGLE GNN MODELS (GCN, GAT, GRAPHSAGE), AND AN ENSEMBLE VARIANT WITHOUT CONTRASTIVE PRETRAINING. BEST RESULTS ARE REPORTED IN **BOLD**.

Task	Model	ACC	MACRO-F1	MACRO-AUC	MACRO-AUPR	MACRO-REC
AD vs MCI	BrainGNN [14]	0.714 ± 0.081	0.696 ± 0.123	0.683 ± 0.092	0.672 ± 0.114	0.685 ± 0.078
AD vs MCI	BGAN [15]	0.621 ± 0.152	0.526 ± 0.202	0.697 ± 0.109	0.577 ± 0.137	0.605 ± 0.134
AD vs MCI	GAT	0.540 ± 0.168	0.484 ± 0.130	0.533 ± 0.099	0.587 ± 0.103	0.516 ± 0.192
AD vs MCI	GCN	0.558 ± 0.090	0.544 ± 0.105	0.581 ± 0.104	0.565 ± 0.099	0.588 ± 0.078
AD vs MCI	GraphSAGE	0.591 ± 0.094	0.567 ± 0.113	0.601 ± 0.097	0.582 ± 0.134	0.574 ± 0.136
AD vs MCI	Ours (w/o CP)	0.662 ± 0.081	0.653 ± 0.102	0.639 ± 0.118	0.617 ± 0.086	0.621 ± 0.126
AD vs MCI	**Ours**	$\mathbf{0.740 \pm 0.066}$	$\mathbf{0.724 \pm 0.099}$	$\mathbf{0.725 \pm 0.113}$	$\mathbf{0.703 \pm 0.117}$	$\mathbf{0.716 \pm 0.084}$
CN vs MCI	BrainGNN [14]	0.674 ± 0.093	0.616 ± 0.088	0.636 ± 0.148	0.637 ± 0.099	0.639 ± 0.125
CN vs MCI	BGAN [15]	0.691 ± 0.090	0.640 ± 0.082	0.689 ± 0.113	0.664 ± 0.093	0.659 ± 0.077
CN vs MCI	GAT	0.577 ± 0.087	0.563 ± 0.158	0.581 ± 0.100	0.557 ± 0.120	0.600 ± 0.079
CN vs MCI	GCN	0.570 ± 0.114	0.556 ± 0.077	0.566 ± 0.138	0.589 ± 0.135	0.582 ± 0.139
CN vs MCI	GraphSAGE	0.581 ± 0.167	0.537 ± 0.069	0.562 ± 0.164	0.550 ± 0.133	0.519 ± 0.116
CN vs MCI	Ours (w/o CP)	0.717 ± 0.138	0.627 ± 0.095	0.603 ± 0.105	0.678 ± 0.125	0.592 ± 0.091
CN vs MCI	**Ours**	$\mathbf{0.753 \pm 0.102}$	$\mathbf{0.728 \pm 0.129}$	$\mathbf{0.716 \pm 0.093}$	$\mathbf{0.735 \pm 0.089}$	$\mathbf{0.702 \pm 0.128}$
CN vs AD	BrainGNN [14]	0.666 ± 0.117	0.630 ± 0.137	0.657 ± 0.069	0.686 ± 0.119	0.673 ± 0.101
CN vs AD	BGAN [15]	0.645 ± 0.075	0.584 ± 0.120	0.625 ± 0.057	0.643 ± 0.119	0.578 ± 0.047
CN vs AD	GAT	0.551 ± 0.116	0.549 ± 0.125	0.583 ± 0.114	0.563 ± 0.082	0.566 ± 0.158
CN vs AD	GCN	0.603 ± 0.102	0.598 ± 0.089	0.588 ± 0.096	0.567 ± 0.071	0.592 ± 0.051
CN vs AD	GraphSAGE	0.599 ± 0.098	0.570 ± 0.120	0.589 ± 0.086	0.534 ± 0.123	0.566 ± 0.167
CN vs AD	Ours (w/o CP)	0.697 ± 0.118	0.616 ± 0.124	0.634 ± 0.106	0.656 ± 0.077	0.578 ± 0.083
CN vs AD	**Ours**	$\mathbf{0.772 \pm 0.082}$	$\mathbf{0.748 \pm 0.072}$	$\mathbf{0.712 \pm 0.086}$	$\mathbf{0.725 \pm 0.051}$	$\mathbf{0.704 \pm 0.045}$
CN vs MCI vs AD	BrainGNN [14]	0.629 ± 0.128	$\mathbf{0.627 \pm 0.090}$	0.634 ± 0.062	$\mathbf{0.666 \pm 0.053}$	0.636 ± 0.127
CN vs MCI vs AD	BGAN [15]	0.595 ± 0.075	0.567 ± 0.099	0.518 ± 0.101	0.575 ± 0.121	0.605 ± 0.067
CN vs MCI vs AD	GAT	0.536 ± 0.144	0.512 ± 0.068	0.504 ± 0.108	0.578 ± 0.060	0.526 ± 0.101
CN vs MCI vs AD	GCN	0.571 ± 0.146	0.520 ± 0.158	0.498 ± 0.142	0.531 ± 0.109	0.530 ± 0.116
CN vs MCI vs AD	GraphSAGE	0.501 ± 0.097	0.561 ± 0.136	0.553 ± 0.115	0.543 ± 0.093	0.552 ± 0.156
CN vs MCI vs AD	Ours (w/o CP)	0.576 ± 0.106	0.596 ± 0.133	0.564 ± 0.114	0.596 ± 0.079	0.618 ± 0.138
CN vs MCI vs AD	**Ours**	$\mathbf{0.635 \pm 0.044}$	0.613 ± 0.055	$\mathbf{0.680 \pm 0.078}$	0.632 ± 0.058	$\mathbf{0.644 \pm 0.074}$

TABLE II

MOST FREQUENTLY OCCURRING NODE–NODE CONNECTIONS IDENTIFIED BY PGEXPLAINER FOR EACH DIAGNOSTIC GROUP (AD, MCI, AND CN). FREQUENCIES REPRESENT THE PERCENTAGE OF SAMPLES IN WHICH A GIVEN EDGE WAS SELECTED AMONG THE MOST RELEVANT CONNECTIONS.

AD			MCI			CN		
Node i	**Node j**	**Freq. (%)**	**Node i**	**Node j**	**Freq. (%)**	**Node i**	**Node j**	**Freq. (%)**
Hippocampus L	ParaHippocampal L	19.6	Hippocampus L	ParaHippocampal L	6.2	Frontal Sup L	Frontal Sup Medial L	2.6
Cingulum Ant L	Cingulum Mid L	13.9	Hippocampus R	ParaHippocampal R	4.7	Cerebelum Crus1 L	Cerebelum Crus2 L	1.1
Parietal Sup R	Precuneus R	9.4	Cingulum Mid L	Cingulum Post L	2.1	Temporal Inf L	Fusiform L	0.8
Parietal Sup L	Parietal Inf L	5.5	Precuneus L	Parietal Sup L	1.9	Temporal Sup R	Heschl R	0.5
Hippocampus R	ParaHippocampal R	3.4	Temporal Sup L	Heschl L	1.3	Hippocampus L	ParaHippocampal L	0.2

jointly improving the accuracy, generalization, and stability of the proposed approach.

D. Qualitative Results

To qualitatively assess the explainability of the proposed framework, we analyze the class-specific connectivity patterns identified by PGExplainer [9]. For each test subject, edge importance scores were computed and aggregated across the ensemble to obtain stable relevance estimates. The most influential edges were then selected per graph and analyzed for frequency of occurrence within each diagnostic group.

Table II summarizes the most frequently occurring node–node connections for AD, MCI, and CN subjects. In the AD group, the identified connections are predominantly concentrated in medial temporal and limbic regions, including bilateral hippocampal–parahippocampal and cingulate connections, which are well known to be strongly affected by AD pathology [29], [30]. The MCI group exhibits similar but less frequent patterns, reflecting an intermediate stage of structural disruption. In contrast, the CN group shows sparse, low-frequency connections among frontal, temporal, and cerebellar regions, consistent with preserved, distributed connectivity.

The qualitative results indicate that the proposed ensemble-based explainability framework recovers anatomically meaningful and disease-relevant connectivity patterns while also capturing progressive differences across diagnostic stages.

V. CONCLUSION

We presented a contrastively pretrained ensemble of graph neural networks for Alzheimer's disease classification, utilizing complementary GNN architectures with brain graph representations. Our framework employs contrastive pretraining to develop robust embeddings, while ensemble learning enhances predictive performance across diagnostic tasks. The inclusion of PGExplainer allows for the identification of disease-relevant subgraphs and ROI-level contributions, offering insights into

AD-related connectivity. Results on the ADNI dataset show that our approach outperforms state-of-the-art graph-based baselines in accuracy and robustness.

Future work will extend the framework to longitudinal and multimodal data, strengthen the link between contrastive pretraining and explainability, and incorporate human-in-the-loop evaluation for clinically meaningful refinement. Additionally, the model may be adapted to a fully prototype-based, unsupervised, distance-driven approach.

ACKNOWLEDGMENTS

This work was supported by the FAIR – Future Artificial Intelligence Research (PNRR "Partenariati Estesi") funding initiative (Project H97G22000210007), with contributions from Gi.Ca. A Ph.D. fellowship funds A.G.V.'s research within the Italian "D.M. n. 630, April 24, 2024" – under the NRRP, Mission 4, Component 2, Investment 3.3 – Ph.D. project "Explainability of Artificial Intelligence systems for applications in the medical field", co-supported by Bristol-Myers Squibb s.r.l. (CUP B91I24000160007).

We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

Generative AI tools, specifically OpenAI's ChatGPT and Grammarly, were used exclusively for grammar correction and language refinement. The authors conceived, wrote, and validated all content.

REFERENCES

[1] B. Khemani, S. Patil, K. Kotecha, and S. Tanwar, "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions," *Journal of Big Data*, vol. 11, no. 1, p. 18, 2024.

[2] S. Ali, M. Piana, M. Pardini, and S. Garbarino, "Graph neural networks in Alzheimer's disease diagnosis: a review of unimodal and multimodal advances," *Frontiers in Neuroscience*, vol. 19, 2025.

[3] L. Zhang, Y. Zhao, T. Che, S. Li, and X. Wang, "Graph neural networks for image-guided disease diagnosis: A review," *iRADIOLOGY*, vol. 1, no. 2, p. 151–166, 2023.

[4] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *Healthcare*, vol. 11, no. 12, p. 1808, 2023.

[5] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Curran Associates Inc., 2020.

[6] X. Wang, Y. Fang, Q. Wang, P.-T. Yap, H. Zhu, and M. Liu, "Self-supervised graph contrastive learning with diffusion augmentation for functional MRI analysis and brain disorder detection," *Medical Image Analysis*, vol. 101, p. 103403, 2025.

[7] D. Bhati, F. Neha, and M. Amiruzzaman, "A Survey on Explainable Artificial Intelligence (XAI) Techniques for Visualizing Deep Learning Models in Medical Imaging," *Journal of Imaging*, vol. 10, no. 10, p. 239, 2024.

[8] I. D. Mienye, G. Obaido, N. Jere, E. Mienye, K. Aruleba, I. D. Emmanuel, and B. Ogbuokiri, "A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges," *Informatics in Medicine Unlocked*, vol. 51, p. 101587, 2024.

[9] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized Explainer for Graph Neural Network," 2020.

[10] E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot, "Automated anatomical labelling atlas 3," *NeuroImage*, vol. 206, p. 116189, 2020.

[11] M. Proietti, A. Ragno, B. L. Rosa, R. Ragno, and R. Capobianco, "Explainable AI in drug discovery: self-interpretable graph neural network for molecular property prediction using concept whitening," *Machine Learning*, vol. 113, no. 4, p. 2013–2044, 2023.

[12] N. A. Valous, F. Popp, I. Zörnig, D. Jäger, and P. Charoentong, "Graph machine learning for integrated multi-omics analysis," *British Journal of Cancer*, vol. 131, no. 2, p. 205–211, 2024.

[13] R. Scaringi, G. Fiameni, G. Vessio, and G. Castellano, "GraphCLIP: Image-graph contrastive learning for multimodal artwork classification," *Knowledge-Based Systems*, vol. 310, p. 112857, 2025.

[14] X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola, and J. S. Duncan, "BrainGNN: Interpretable Brain Graph Neural Network for fMRI Analysis," *Medical Image Analysis*, vol. 74, p. 102233, 2021.

[15] Z. Zhou, Q. Wang, X. An, S. Chen, Y. Sun, G. Wang, and G. Yan, "A novel graph neural network method for Alzheimer's disease classification," *Computers in Biology and Medicine*, vol. 180, p. 108869, 2024.

[16] Y. Zhang, X. He, Y. H. Chan, Q. Teng, and J. C. Rajapakse, "Multimodal graph neural network for early diagnosis of Alzheimer's disease from sMRI and PET scans," *Computers in Biology and Medicine*, vol. 164, p. 107328, 2023.

[17] W. Ju, Y. Wang, Y. Qin, Z. Mao, Z. Xiao, J. Luo, J. Yang, Y. Gu, D. Wang, Q. Long, S. Yi, X. Luo, and M. Zhang, "Towards Graph Contrastive Learning: A Survey and Beyond," 2024.

[18] B. Pfeifer, H. Chereda, R. Martin, A. Saranti, S. Clemens, A.-C. Hauschild, T. Beißbarth, A. Holzinger, and D. Heider, "Ensemble-GNN: federated ensemble learning with graph neural networks for disease module discovery and classification," *Bioinformatics*, vol. 39, no. 11, 2023.

[19] E. Lella and G. Vessio, "Ensembling complex network 'perspectives' for mild cognitive impairment detection with artificial neural networks," *Pattern Recognition Letters*, vol. 136, p. 168–174, 2020.

[20] A. G. Valerio, K. Trufanova, S. de Benedictis, G. Vessio, and G. Castellano, "From segmentation to explanation: Generating textual reports from MRI with LLMs," *Computer Methods and Programs in Biomedicine*, vol. 270, p. 108922, 2025.

[21] X. Hu, Z. Sun, Y. Nian, Y. Wang, Y. Dang, F. Li, J. Feng, E. Yu, and C. Tao, "Self-Explainable Graph Neural Network for Alzheimer Disease and Related Dementias Risk Prediction: Algorithm Development and Validation Study," *JMIR Aging*, vol. 7, p. e54748, 2024.

[22] the ADNI team, "Alzheimer's Disease Neuroimaging Initiative (ADNI) Database," adni.loni.usc.edu, 2003–2025, data used in this work were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. ADNI was launched in 2003 as a public–private partnership to study Alzheimer's disease biomarkers and progression.

[23] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, no. 21, p. e104–e107, 2017.

[24] S. Chithra and R. Vijayabhanu, *A Comprehensive Review and Current Methods for Classifying Alzheimer's Disease Using Feature Extraction and Machine Learning Techniques*. Springer Nature Singapore, 2023, p. 735–747.

[25] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: a comprehensive review," *Computational Social Networks*, vol. 6, no. 1, 2019.

[26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," 2018.

[27] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," 2018.

[28] W. Ågren, "The NT-Xent loss upper bound," 2022.

[29] S. Abushakra and e. a. Doraiswamy, "Hippocampal Atrophy on Magnetic Resonance Imaging as a Surrogate Marker for Clinical Benefit and Neurodegeneration in Early Symptomatic Alzheimer's Disease: Synthesis of Evidence from Observational and Interventional Trials," *CNS Drugs*, vol. 40, no. 2, p. 199–214, 2025.

[30] G. B. Frisoni and e. a. Hansson, "New landscape of the diagnosis of Alzheimer's disease," *The Lancet*, vol. 406, no. 10510, p. 1389–1407, 2025.