Explainable Graph Neural Networks for Psychiatry Disorder Diagnosis Using Brain Networks

1st Nesrine Jellali
Computer Science and Mathematics Department
Faculty of Economics and Management of Sfax
Sfax, Tunisia
nesrine.jellali@fsegs.usf.tn

2nd Rebh Soltani
REsearch Groups in Intelligent Machines
National Engineering School of Sfax
Sfax, Tunisia
rebh.soltani@enis.tn

3rd Hela Ltifi
Computer Science and Mathematics Department
Faculty of sciences and technology of Sidi Bouzid, University of Kairouan,Tunisia.
hela.ltifi@fstsbz.rnu.tn

***Abstract*—Graph Neural Networks (GNNs) are a revolutionary game-changing approach toward psychiatric diagnosis because of their incomparable capability for modeling complex relations in neuroimaging data. Herein, we introduce an explainable high-powered GNN-based model designed to address the challenge of distinguishing patients with Major Depressive Disorder (MDD) from healthy t method's foundation is on the following new suggestions: feature extraction, advanced hyperparameter adjustment, and powerful explainable GNN (X-GNN). Our model, tested on the Rest-Meta-MDD dataset, demonstrated exceptional performance while achieving state-of-the-art performance.**

***Index Terms*—Psychiatry Disorder, Explainability, Graph Neural Networks, Causal Generation, Brain Network**

I. Introduction

In recent years, emerging advanced neuroimaging modalities, including functional magnetic resonance imaging of the resting state, have helped shape our understanding of brain connectivity in psychiatric disorders [1]. These imaging modalities can characterize functional networks within the brain, modeled as graphs with nodes representing regions and edges representing functional connections. These developments are especially valuable in comprehending psychological disorders like MDD, which show intricate patterns of brain connectivity.

MDD is one of the most common and severe mental health illnesses, adding greatly to the world load of sickness and suffering [1]. It's also characterized by persistent emotions of melancholy, guilt, and worthlessness, along with an elevated risk of suicide [2]. The underlying pathophysiological mechanisms are complex and remain unclear.

Recently, MDD has increasingly been conceptualized as a disorder of abnormal functional integration across distributed brain regions involved in emotional and cognitive regulation [3]. However, previous studies have reported inconsistent patterns of brain abnormalities, primarily due to limited sample sizes and variability in analysis workflows. To address these challenges, the REST-meta-MDD consortium compiled the largest coordinated resting-state fMRI dataset of MDD patients and healthy controls, including over 1,000 participants: http://rfmri.org/REST-meta-MDD.

We present X-GNN, an explainable graph neural network that finds the most significant brain connection patterns connected to MDD diagnosis. In the first step, we utilize strong preprocessing algorithms to enhance the input data, assuring high-quality graph representations. Unlike techniques needing an auxiliary interpretative network, X-GNN is naturally interpretable. It learns two unique subgraph-level representations: α, capturing causal relationships, and β, expressing non-causal characteristics. This is done utilizing a graph variational autoencoder architecture, regularized by a conditional mutual information (CMI) requirement to enable meaningful disentanglement. To further boost performance and stability, we apply hyperparameter tuning, assuring robust and dependable outcomes. This paper is constructed as follows: Section 2 includes a review of relevant work. Section 3 describes our suggested approach. Section 4 presents the experimental setup. Section 5 provides and analyzes the findings, comparing X-GNN's performance versus current approaches. Finally, Section 6 conclude the study.

II. Related Work

A. Graph Neural Networks

Real-world applications naturally represent data as graphs; common examples are found in social networks, information systems, chemistry, and biology [8], [12]. Graphs have a rich way of modeling complex relational and node-level information. However, their very own nature of complexity hinders their analysis and representation effectively [13]. GNNs have emerged in recent years as the de facto paradigm to perform machine learning on graph-structured data. By recursively aggregating information from neighbouring nodes, GNNs effectively capture the graph topology and node-level features, thus enabling strong and flexible modeling capabilities [9].

It aimed at unraveling the complex structure of relationships in graph-structured data by inferring a latent representation of both individual entities and their contextual roles within the greater network. In a way, by creating unique embeddings, GNNs can encode not only intrinsic node attributes but also

the complex interactions defining a graph's topology. This therefore, summarizes the ability to:

- Predict node behavior: Forecast the property of individual nodes, such as predicting which members of a social network are most likely to churn or which financial transaction is fraudulent [11].
- Comprehend global graph properties: Classify entire graphs by extracting meaningful patterns, such as deciding on the toxicity of a molecular compound in drug discovery [14].
- Infer missing links: Predict unobserved connections that empower applications such as personalized product recommendations in e-commerce or protein-protein interaction predictions in bioinformatics [15].

B. Model for Psychiatric Diagnosis

Recent research have progressively used FC as a neurological biomarker for constructing automated diagnosis models for mental diseases. Traditionally, these models employ a two-stage training strategy: first, feature selection approaches highlight key FC links, which are subsequently input into a classification model. Feature selection procedures include group-level statistical tests (e.g., t-tests, rank-sum tests) and unsupervised dimensionality reduction techniques like as principal component analysis (PCA) and tensor decomposition [16]. For categorization, shallow machine learning models like random forests (RF) and support vector machines (SVM) remain commonly utilized [17]. However, these standard approaches fail to represent the nonlinear and topological complexity of brain networks, limiting their efficacy, particularly on large-scale psychiatric datasets. Additionally, the two-stage training technique generally leads to incorrect forecasts. To solve these constraints, XGNN presents an end-to-end, self-explaining model that directly identifies causative biomarkers, boosting both interpretability and diagnostic accuracy.

III. Proposed Approach

We propose an Explainable Graph Convolutional Network (XGCN) framework for brain network classification using resting-state fMRI (rs-fMRI) data. The model captures brain connectivity via a graph representation learning paradigm, integrating k-NN-based graph construction for robust structure discovery. Multi-layer message passing with optimized activation mitigates over-smoothing, while GNNExplainer enhances interpretability, supporting biomarker discovery in neuroscience.

A. Graph Construction

We represent each subject as a node in a graph $G = (V, E)$, where:

- $V = \{v_1, v_2, ..., v_N\}$ are nodes corresponding to subjects.
- E defines subject-to-subject similarities using k-NN with cosine similarity.
- $X \in \mathbb{R}^{N \times d}$ is the feature matrix, where each node has a d-dimensional feature vector.

The adjacency matrix A is constructed using cosine similarity:

$$\text{sim}(i, j) = \frac{X_i \cdot X_j}{\|X_i\| \|X_j\|} \tag{1}$$

An edge exists between nodes i and j if j is among the k-nearest neighbors of i.

We apply StandardScaler for feature normalization:

$$X' = \frac{X - \mu}{\sigma} \tag{2}$$

B. Graph Convolutional Network (GCN) Architecture

To learn discriminative node representations, we employ a multi-layer Graph Convolutional Network (GCN). The core update rule is:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \tag{3}$$

where $H^{(l)}$ is the node feature matrix at layer l, $\tilde{A}$ is the adjacency matrix with self-loops, and $\tilde{D}$ is the degree matrix.

For the final layer, we apply Softmax for classification:

$$\hat{Y} = \text{Softmax}\left(H^{(L)}W^{(L)}\right) \tag{4}$$

C. Training Objective and Optimization

We use the cross-entropy loss function:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}) \tag{5}$$

Optimized using the Adam optimizer with a learning rate of 0.01 and weight decay $\lambda = 5 \times 10^{-4}$.

D. Model Interpretability: GNNExplainer

To enhance interpretability, we integrate GNNExplainer, which identifies the most influential nodes and edges. GNNExplainer optimizes an edge importance function, aiding in model transparency and biomarker discovery.

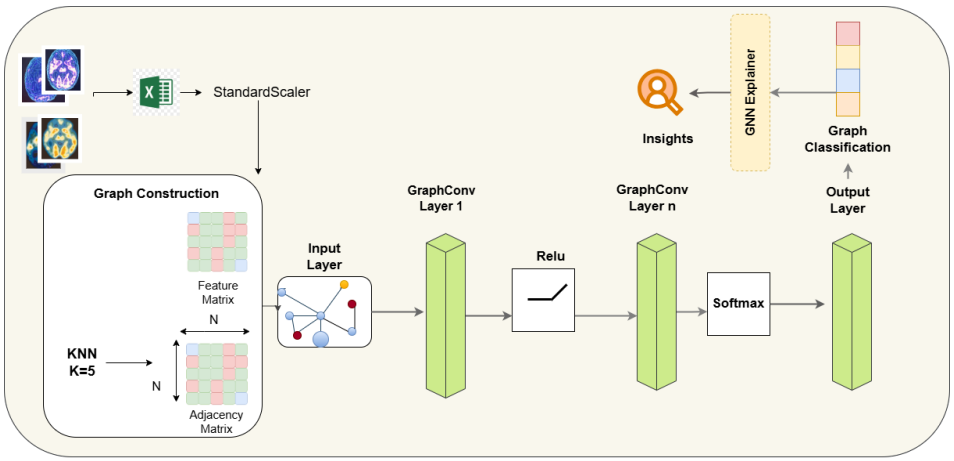


Fig. 1: Explainable Graph Convolutional Network Architecture.

E. Training Procedures

Algorithm 1 Overview of GNN Training and Explanation

Require: Dataset D, Number of neighbors k, Training epochs E, Node to explain n_e.
Ensure: Trained GNN model, Explanation for node n_e.

Load and preprocess the dataset
Scale numerical features using `StandardScaler`
Map labels: MDD $\rightarrow$ 0, Controls $\rightarrow$ 1
Construct the graph G
Build adjacency matrix with k-nearest neighbors
Convert adjacency matrix to graph structure using DGL
Add self-loops to G
Define and train the GNN model
Define GNN model with two `GraphConv` layers
Initialize training and testing masks (80%-20%)
for $t = 1$ to E **do**
    Compute logits for G
    Compute loss using `CrossEntropyLoss`
    Backpropagate and update parameters using Adam optimizer
end for
Evaluate the GNN model
Evaluate GNN model on test data =0

IV. Experiments and Results

A. Dataset Description

REST-meta-MDD Yan et al. (2019): It is one of the largest MDD dataset including more than 2000 participants from twenty-five independent research groups4. In this study, 1604 participants (848 MDDs and 794 HCs). More informations can be found at REST-meta-MDD dataset.

TABLE I: Dataset Statistics.

Dataset	Edges	Nodes	Graphs	Classes
Rest-meta-MDD	11900	2380	1	2

The comprehensive statistics of this dataset are shown in Table I.

B. Preprocessing

The preprocessing of the REST-meta-MDD dataset involved standard steps using **SPM12** and **DPABI** in MATLAB. This included slice timing correction, realignment for head motion, spatial normalization to the MNI152 template, and smoothing with an 8mm Gaussian kernel. The smoothed image $I_s(x, y, z)$ was obtained by convolving the original image I with a Gaussian kernel G:

$$I_s(x, y, z) = I(x, y, z) * G(x, y, z; \sigma),$$

where σ is derived from FWHM:

$$\sigma = \frac{\text{FWHM}}{2\sqrt{2 \ln 2}}.$$

We used the **Brainnetome Atlas** for brain parcellation into **246 ROIs**, extracting the mean time series $s_i(t)$ for each region. Functional connectivity matrices C_{ij} were computed using Pearson correlation:

$$C_{ij} = \frac{\text{cov}(s_i, s_j)}{\sigma_{s_i} \sigma_{s_j}},$$

resulting in a symmetric 246×246 matrix. The upper triangular part was vectorized and stored in CSV format, with missing values imputed using **K-nearest neighbors (KNN)** imputation ($k = 5$):

$$\hat{x}_{ij} = \frac{1}{k} \sum_{l=1}^{k} x_{lj}.$$

Feature standardization was applied via Z-score normalization:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}.$$

We reduced dimensionality using **Principal Component Analysis (PCA)** while retaining 100 components, preserving over 95% of variance:

$$X_{\text{PCA}} = XW.$$

Finally, the dataset was split into **80% training** and **20% test sets**, ensuring balanced representation of MDD patients and healthy controls. A graph representation was constructed, modeling each subject as a node, with edges defined through a thresholded functional connectivity approach, preserving significant connections. This preprocessing pipeline provided a robust foundation for training the **Explainable Graph Convolutional Network (X-GCN)** for brain network classification.

C. Assessment metrics for classification effectiveness

We use the following metrics to evaluate each binary classification's results:

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \tag{8}$$

$$\text{Sensitivity (SEN)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{9}$$

$$\text{Specificity (SPE)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \tag{10}$$

$$\text{Precision (PRE)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{11}$$

$$\text{F1-Score} = \frac{2 \times \text{PRE} \times \text{SEN}}{\text{PRE} + \text{SEN}} \times 100\% = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{12}$$

where TP, TN, FP, and FN stand for the prediction data's true positives, true negatives, false positives, and false negatives, respectively.

The confusion matrix (see Fig. 2) represents the classification performance of the model in differentiating between two classes: MDD and Controls. From a total of 476 samples, the model correctly classified 238 samples as MDD (True Positives) and 237 samples as Controls (True Negatives). There was only 1 instance of a False Negative, where an MDD sample was misclassified as Control, and no False Positives were observed. This leads to a very high accuracy of 99.79%, which is the ratio of correctly classified instances to the total number of samples $\frac{238+237}{476}$. The near-perfect classification performance, as shown in the matrix, shows that the model is highly accurate in identifying and distinguishing between the two classes with minimal misclassification errors.

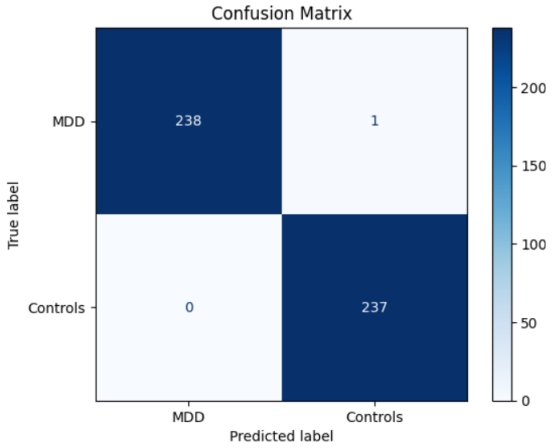


Fig. 2: Confusion Matrix for Rest-Meta-MDD Dataset Classification.

The ROC curve shown in Fig 3 presents a complete description of the classifier performance by plotting the True Positive Rate, Sensitivity, against the False Positive Rate at various thresholds. This is evidence of a very effective model with extremely small trade-offs between sensitivity and specificity: the near-vertical rise at the beginning of the curve, then a flat plateau at maximum True Positive Rate.

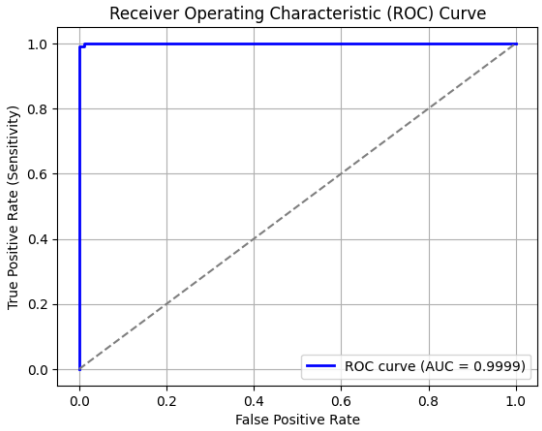
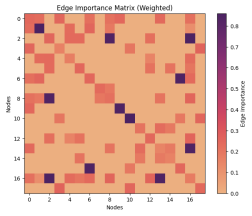


Fig. 3: ROC Curve for Rest-Meta-MDD Dataset Classification.

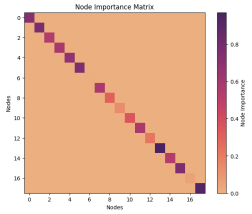
An AUC value of 0.9999 reflects excellent performance in distinguishing between MDD and Control classes. This high AUC value shows that it is an extremely good classifier with a very high probability of ranking a random positive example higher than a random negative example. The result leads to the robustness and preciseness of the model, allowing for a low rate of Type I (false positive) and Type II errors (false negative). In addition, this points toward very good generalization capability for the proposed approach, which can classify unseen data properly with high discriminative power for all thresholds.

These visualizations (cf. Figures 4a 4c, 4d and 4b) provide concrete intuition about how the model selects and uses the most informative parts of the graph. The edge-based analysis

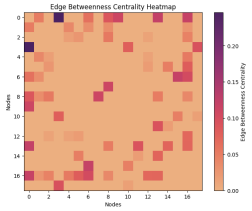
shows that indeed the model assigns different importance to the connections, with a number of edges being much more relevant; these are likely pointing to the main pathways along which information flows critically. Such differences suggest that the model successfully differentiates between strong and weak relationships, thus contributing to more accurate predictions.



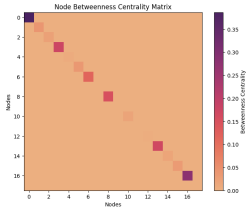
(a) Edge Importance Matrix



(b) Node Importance Matrix



(c) Edge Betweenness Centrality Heatmap



(d) Node Betweenness Centrality Heatmap

Fig. 4: Comparison of Different Heatmaps

The node-level results show a similar pattern, in that some nodes are much more important than others, which may reflect their centralities or their influences within the graph structure. High-importance nodes could be critical hubs or pivotal points in the aggregation and propagation of information within the graph.

Furthermore, heatmaps of centrality highlight how the model depicts structural dependencies with regard to nodal and functional roles of node and edge strengths. These facts hereby validate our model, rightfully emphasizing those very graph components which contribute most toward its predictions to make the GCN more interpretable and robust.

V. Discussion

As shown in Table II, our proposed model has shown significant improvements compared to the state-of-the-art methods based on its application to the REST-meta-MDD dataset. The model outperformed the existing approaches in almost all the metrics. These results reflect that the model effectively learns and utilizes the graph-based representations within the REST-meta-MDD dataset, thus providing clear improvements over prior methodologies on this challenging brain disease dataset.

TABLE II: Evaluation of Proposed Model Against State-of-the-Art Methods.

Model	Accuracy	F1-score	Precision	Sensitivity	Specificity
SVM [5]	59.7	61.2	61.2	61.1	58.2
DGCNN [4]	72.1	69.9	62.5	79.2	67.1
RF [6]	62.3	63.5	63.3	63.7	60.7
GCN [7]	67.4	68.4	68.2	68.7	66.0
CI-GNN [10]	65 (Avg)	61 (Avg)	-	-	-
Our model	**98.94**	**98.32**	**98.38**	**98.32**	**98.21**

TABLE III: Parameter Values for Each Method.

Method	**Parameter**	**Value**
[7]	Dropout rate	0.5
	Batch size	50
	Epoch	100
	Learning rate	0.001
[5]	C	100.0
	Kernel	rbf
	Degree	3
	Gamma	0.01
[6]	Max features	2
	Max depth	100
	Min samples split	200
	Min samples leaf	50
[10]	Min samples split	200
	Min samples leaf	50
Our model (X-GCN)	Number of Neighbors (k)	5
	Hidden Layer Size	16
	Learning Rate	0.01
	Number of Epochs	50

Another critical contribution of this study is the integration of GNNExplainer into the model. In fact, as the demand for explainable AI grows-in particular, for healthcare applications-interpretabily becomes as important as performance. GNNExplainer bridges the gap between model performance and interpretability by identifying the most influential nodes and edges for model decisions. It enables practitioners to understand the underlying reasoning for predictions, making such models powerful but more trustworthy, which is an essential requirement in medical domains.

Furthermore, our analysis showed that although GNNs had achieved great results in predicting mental disorders, there are still very important trade-offs between model complexity and interpretability. This will imply a tradeoff between the desired model performance in complex, deeper-layer architecture for

less transparency; clinicians would face difficulty in interpretation and hence trusted results. That means the reverse is that, with simpler models, although offering interpretability, it may be compromised in performance over the temporal complexities. This balance is still a point of future research, and a proper balance between model sophistication and interpretability is key for moving forward with practical applications of GNNs in psychiatry.

Other takeaways from the current study would relate to hyperparameter optimization in arriving at ideal results. Also, as identified from the widely differing parameter sets adopted by state-of-the-art models, a set of carefully chosen hyperparameters makes a wide difference in terms of model efficiency and predictive performance. Our observations have underlined the importance of careful tuning, especially in sensitive domains like that of mental health problems, where improvement in prediction precision might have seriously beneficial real-life consequences. As present in the previous table,the explainability results of our X-GCN model underlined the most important brain connectivity patterns that distinguish MDD patients from healthy controls. The top feature and edge importances point to the most influential neural connections, while sparsity ensures that the model focuses only on relevant features. The drop in Fidelity- compared to Fidelity+ confirms that the selected features are crucial for classification. These differences between subjects confirm that our model captures personalized changes in brain connectivity and does not depend on a single biomarker; therefore, reinforcing biological validity and interpretability of the model.

VI. Conclusion

This work highlights how GNNs can exhibit promising performance for the diagnosis of mental disorders, underlining the contribution of temporal graphs in modeling the dynamic relationships between patients. Herein, by integrating GNNExplainer, we show that GNNs are able to provide strong predictive performance along with model interpretability, an important factor in healthcare applications. The research further underlines the importance of hyperparameter optimization, where this careful tuning may grossly enhance model accuracy. Going forward, further investigation in hybrid models and developing more efficient methods for hyperparameter searching will be the key to realizing the full potential of GNNs in clinical settings. This study fortifies the now-flourishing domain of explainable AI in healthcare and sets the bedrock for further work toward accuracy with transparency in the diagnosis of mental health conditions.

References

[1] Ferrari, Alize J., Fiona J. Charlson, Rosana E Norman, Scott B. Patten, Greg Freedman, Christopher J. L. Murray, Theo Vos and Harvey A. Whiteford. “Burden of Depressive Disorders by Country, Sex, Age, and Year: Findings from the Global Burden of Disease Study 2010.” PLoS Medicine 10 (2013).

[2] Sharp, Lisa Kay and Martin S Lipsky. “Screening for depression across the lifespan: a review of measures for use in primary care settings.” American family physician 66 6 (2002): 1001-8 .

[3] Zhao, Y.-J., Mingying Du, Xiaoqi Huang, Su Lui, Zhuangfei Chen, J. Liu, Y Luo, Xiuli Wang, Graham J. Kemp and Qiyong Gong. “Brain grey matter abnormalities in medication-free patients with major depressive disorder: a meta-analysis.” Psychological Medicine 44 (2014): 2927 - 2937.

[4] Zhu, Manyun, Yu Quan and Xuan He. “The classification of brain network for major depressive disorder patients based on deep graph convolutional neural network.” Frontiers in Human Neuroscience 17 (2023).

[5] Hearst, Dumais, Osuna, Platt and Scholkopf, ”Support vector machines,” in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998.

[6] Breiman, L.. “Random Forests.” Machine Learning 45 (2001): 5-32.

[7] Kipf, Thomas and Max Welling. “Semi-Supervised Classification with Graph Convolutional Networks.” ArXiv abs/1609.02907 (2016).

[8] Cho, Eunjoon, Seth A. Myers and Jure Leskovec. “Friendship and mobility: user movement in location-based social networks.” Knowledge Discovery and Data Mining (2011).

[9] Hamilton, William L., Zhitao Ying and Jure Leskovec. “Inductive Representation Learning on Large Graphs.” Neural Information Processing Systems (2017).

[10] Zheng, Kaizhong, Shujian Yu and Badong Chen. “CI-GNN: A Granger Causality-Inspired Graph Neural Network for Interpretable Brain Network-Based Psychiatric Diagnosis.” Neural networks : the official journal of the International Neural Network Society 172 (2023): 106147 .

[11] Gupta, Atika, Priya Matta, and Bhasker Pant. ”Graph neural network: Current state of Art, challenges and applications.” Materials Today: Proceedings 46 (2021): 10927-10932.

[12] You, Jiaxuan, Bowen Liu, Rex Ying, Vijay S. Pande and Jure Leskovec. “Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation.” Neural Information Processing Systems (2018).

[13] Zhou, Jie, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu and Maosong Sun. “Graph Neural Networks: A Review of Methods and Applications.” ArXiv abs/1812.08434 (2018).

[14] Jellali, Nesrine, Rebh Soltani, and Hela Ltifi. ”An Improved Eulerian Echo State Network for Static Temporal Graphs.” In International Conference on Intelligent Systems Design and Applications, pp. 307-318. Cham: Springer Nature Switzerland, 2023.

[15] Zhang, Muhan, and Yixin Chen. ”Link prediction based on graph neural networks.” Advances in neural information processing systems 31 (2018).

[16] Zhang, Z., Allen, G. I., Zhu, H., Dunson, D. (2019). Tensor network factorizations: Relationships between brain structural connectomes and traits. Neuroimage, 197, 330-343.

[17] H. Rubin-Falcone, F. Zanderigo, B. Thapa-Chhetry, M. Lan, J. M. Miller, M. E. Sublette, M. A. Oquendo, D. J. Hellerstein, P. J. McGrath, J. W. Stewart, et al., Pattern recognition of magnetic resonance imaging-based gray matter volume measurements classifies bipolar disorder and major depressive disorder, Journal of affective disorders 227 (2018) 498–505

[18] Soltani, Rebh, Emna Benmohamed, and Hela Ltifi. ”Hybrid Quanvolutional Echo State Network for Time Series Prediction.” ICAART (2). 2024.

[19] Soltani, Rebh, Emna Benmohamed, and Hela Ltifi. ”Topology-adaptive Bayesian optimization for deep ring echo state networks in speech emotion recognition.” Neural Computing and Applications 37.1 (2025): 399-416.

[20] Bargougui, Ikhlas, Rebh Soltani, and Hela Ltifi. ”Refining High-Quality Labels Using Large Language Models to Enhance Node Classification in Graph Echo State Network.”