Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 6, No. 3, 2026

Multimodal Brain Imaging for Brain Disorder Classification: A Review of Graph Neural Networks and Transformer-Based Methods

Ting Zhao *

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 454000, PR China

Abstract: With the rapid advancement of neuroimaging technologies and deep learning methods, multimodal brain imaging analysis has emerged as an important research direction for the early diagnosis and progression assessment of neurological disorders. Different neuroimaging modalities, such as structural magnetic resonance imaging (sMRI), functional magnetic resonance imaging (fMRI), and diffusion tensor imaging (DTI), provide complementary information from anatomical, functional, and structural connectivity perspectives. Compared with single-modality analysis, multimodal learning can offer a more comprehensive characterization of pathological alterations in the brain. At the same time, the increasing use of connectome-based representations has introduced graph-structured data into neuroimaging analysis, making graph neural networks (GNNs) particularly suitable for this field. More recently, Transformer-based architectures have further enhanced the ability of deep models to capture long-range dependencies and complex interactions across temporal windows, graph nodes, and heterogeneous modalities. As a result, the integration of multimodal learning, graph neural networks, and Transformer models has become a promising paradigm for intelligent brain disorder classification. This review focuses on these three key aspects. First, the importance and major strategies of multimodal brain imaging fusion are discussed. Second, the development and applications of graph neural networks in connectome analysis and brain disease classification are summarized. Third, the role of Transformer-based models in multimodal fusion and spatiotemporal brain network modeling is analyzed. Finally, the current challenges and future research directions are discussed. This review aims to provide a structured and theoretically grounded overview of recent methodological progress in multimodal neuroimaging-based diagnosis.

Keywords: Multimodal brain imaging, Graph Neural Network, Transformer, Brain disease classification, Connectome, spatiotemporal modeling.

1. Introduction

Brain disorders, especially neurodegenerative and neurodevelopmental diseases such as Alzheimer's disease (AD), mild cognitive impairment (MCI), Parkinson's disease (PD), and autism spectrum disorder (ASD), have become major health challenges worldwide. Early diagnosis of these disorders is essential because timely intervention may help delay disease progression [1], improve quality of life, and reduce the burden on healthcare systems. However, early-stage brain disorders are often difficult to identify using conventional clinical observations alone, since pathological changes may emerge long before obvious behavioral symptoms become apparent. For this reason, neuroimaging-based computer-aided diagnosis has attracted increasing attention in recent years [2].

Neuroimaging provides noninvasive tools for examining structural, functional, and connectivity-related changes in the brain. Structural MRI is commonly used to characterize morphological abnormalities such as cortical thinning, gray matter atrophy, and regional volume reduction. Functional MRI captures blood-oxygen-level-dependent signals and reflects spontaneous neural activity as well as functional interactions among brain regions. Diffusion tensor imaging provides information about white matter fiber tracts and can be used to construct structural connectivity networks. Since different modalities reveal different aspects of brain pathology, it is natural to consider integrating them within a unified analytical framework [3].

Compared with traditional single-modality analysis, multimodal brain imaging offers at least three major advantages. First, it improves the completeness of disease representation by combining complementary information. Second, it helps reduce the risk of relying excessively on any one noisy or incomplete source of information [4]. Third, it provides a more biologically plausible description of neurological disorders, whose pathological mechanisms often involve complex interactions between structural degeneration, functional dysregulation, and connectivity disruption. See Figure 1, For these reasons, multimodal learning has gradually become a central topic in neuroimaging-based disease classification.

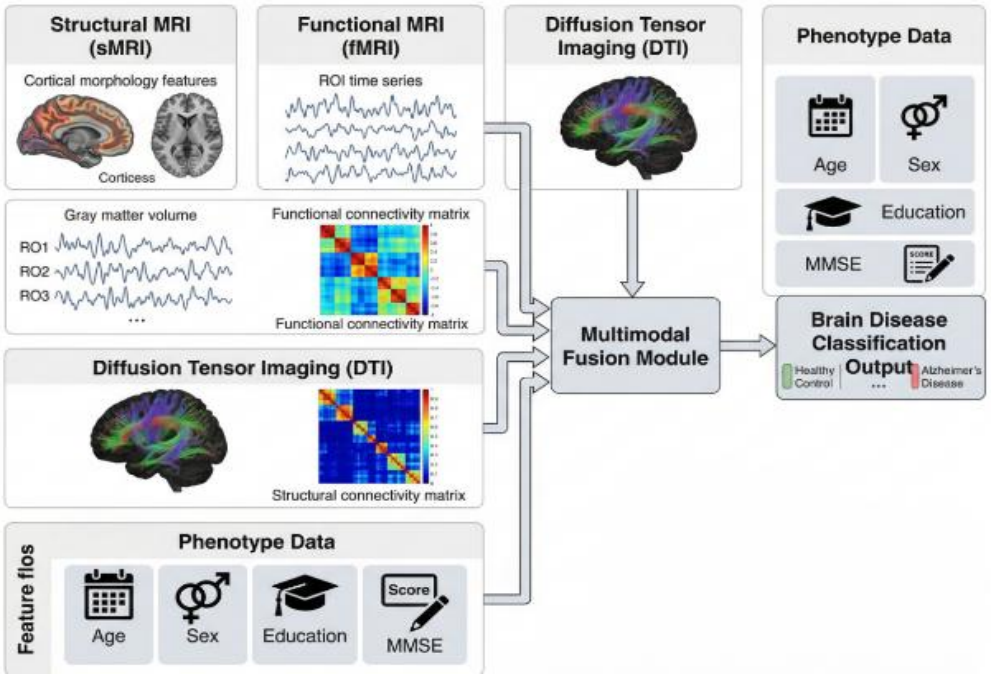

Figure 1. Overview of multimodal brain imaging data and their complementary roles in brain disease classification

At the methodological level, however, multimodal brain imaging also introduces substantial challenges. The data are typically high-dimensional but collected from limited numbers of subjects. Different modalities have heterogeneous statistical properties and may not be naturally aligned in a shared representation space. In addition, many neuroimaging studies increasingly rely on connectome-based analysis, where the brain is represented as a graph rather than as a conventional Euclidean signal. This means that standard machine learning models and even ordinary convolutional neural networks are often insufficient to capture the intrinsic topological structure of the data [5].

Graph neural networks have emerged as a powerful solution to these problems. By representing brain regions as nodes and inter-regional relationships as edges [6], GNNs can naturally model connectome data and extract disease-relevant topological features. In recent years, graph convolutional networks, graph attention networks, graph pooling architectures, and dynamic graph models have all been applied to brain disease classification tasks. These models have demonstrated strong ability to learn discriminative representations from both structural and functional brain networks.

At the same time, Transformer models have rapidly reshaped many areas of machine learning because of their ability to capture long-range dependencies through self-attention. In neuroimaging research, Transformers have been introduced for several purposes: multimodal feature interaction, temporal modeling of dynamic functional connectivity, graph sequence analysis, and even joint modeling of regional spatial relations and temporal dependencies. The combination of GNNs and Transformers is particularly promising because it allows models to simultaneously learn local graph structure and global long-range interactions [7].

In view of these developments, this review focuses on three highly related themes: multimodal brain imaging, graph neural networks, and Transformer-based modeling. Rather than treating them as separate topics, this paper emphasizes how they increasingly converge into a unified framework for intelligent diagnosis. The remainder of this review is organized as follows. Section 2 discusses the role of multimodal neuroimaging and major multimodal fusion strategies. Section 3 reviews graph neural network methods for connectome analysis and brain disease classification. Section 4 focuses on Transformer-based models and their integration with multimodal and graph-based methods. Section 5 summarizes major challenges and future directions. Section 6 concludes the paper.

2. Multimodal Brain Imaging for Brain Disease Classification

(1) Motivation for Multimodal Brain Imaging Analysis

Neurological disorders are highly complex and rarely manifest in only one dimension of brain organization. In many diseases, structural abnormalities, altered functional coordination, and white matter degradation coexist and influence each other. A single-modality approach can only observe part of this pathological landscape. For example, sMRI may reveal tissue atrophy but cannot directly characterize dynamic communication among brain regions. fMRI may detect functional disruption but is less capable of describing underlying anatomical pathways. DTI captures structural connectivity but does not directly measure neural activity. Therefore, multimodal neuroimaging is fundamentally important because it provides complementary observations of brain pathology.

Formally, assume that a subject is described by M imaging modalities. The multimodal input can be represented as

$$\mathcal{X} = X^{(1)}, X^{(2)}, \ldots, X^{(M)} \tag{1}$$

Where $X^{(m)}$ denotes the feature representation extracted from modality m. In a typical scenario, $X^{(1)}$, $X^{(2)}$, and $X^{(3)}$ may correspond to sMRI, fMRI, and DTI, respectively. The goal of multimodal learning is to map $\mathcal{X}$ into a unified latent representation Z that contains more discriminative information than any single modality alone.

The value of multimodal imaging can be understood from both clinical and computational perspectives. Clinically, different disease stages may affect different biological

systems unevenly. Some subjects may exhibit more prominent structural atrophy, while others may display earlier functional dysconnectivity. Computationally, combining modalities can improve robustness because one modality may compensate for limitations in another. This is especially relevant in disorders such as AD and MCI, where disease-related patterns are subtle and distributed across multiple brain systems.

In addition to the three commonly discussed imaging modalities, phenotype variables are often incorporated as auxiliary information. Age, sex, education level, genetic risk indicators, and cognitive assessment scores such as MMSE may all contribute valuable diagnostic priors. In practice, the fusion of imaging and phenotype information often leads to better performance than using imaging data alone. This indicates that multimodal learning in neuroimaging should not be interpreted narrowly as only combining multiple scans; rather, it should be seen as integrating heterogeneous but complementary information sources related to disease state [8].

(2) Major Modalities and Their Characteristics

Among available neuroimaging modalities, sMRI, fMRI, and DTI are among the most widely used in multimodal studies.

sMRI provides high-resolution anatomical information and is commonly used to quantify morphological features such as cortical thickness, hippocampal volume, gray matter density, and regional atrophy. In neurodegenerative disease research, these structural biomarkers are highly important because tissue degeneration is often one of the clearest signs of disease progression. However, morphological features alone may not capture how disease alters interactions among brain regions.

fMRI, especially resting-state fMRI, is widely used to construct functional connectivity networks. After preprocessing and ROI parcellation, a time series is extracted from each brain region, and pairwise relationships are measured using statistical metrics such as Pearson correlation. The resulting connectivity matrix reflects synchronous activity patterns across the brain. Functional changes are often highly informative in early disease stages because altered coordination among regions may emerge before obvious structural damage becomes visible.

DTI provides a way to study white matter integrity and structural connectivity. Through diffusion-based fiber tracking, DTI estimates the existence and strength of anatomical pathways between regions. This modality complements both sMRI and fMRI because it describes the physical substrate that constrains communication across the brain.

These modalities differ not only in content but also in representation form. sMRI is often processed into region-level scalar features or volumetric maps. fMRI may be represented as time series, static connectivity matrices, or dynamic graph sequences. DTI may be expressed as tractography-derived adjacency matrices or edge-weighted structural graphs. Because of such differences, multimodal fusion is not a trivial concatenation problem. Instead, it requires careful consideration of how heterogeneous representations can be aligned, encoded, and integrated [8].

(3) Main Strategies for Multimodal Fusion

Multimodal fusion methods are generally categorized into early fusion, intermediate fusion, and late fusion.

In early fusion, features from all modalities are concatenated at the input level:

$$Z = \left[X^{(1)} \parallel X^{(2)} \parallel \cdots \parallel X^{(M)}\right] \tag{2}$$

Where $\parallel$denotes feature concatenation. The advantage of this approach lies in its simplicity. Once all features are arranged in a unified vector or matrix, a downstream model can be trained directly. However, early fusion also has obvious limitations. It assumes that features from different modalities are directly comparable, but in practice they may differ substantially in dimension, scale, and semantic meaning. Direct concatenation can therefore amplify noise and redundancy.

Late fusion takes a different approach. Each modality is processed independently to produce its own prediction, and the final output is obtained by weighted combination:

$$\hat{y} = \sum_{m=1}^{M} \alpha_m \hat{y}^{(m)} \tag{3}$$

Where $\hat{y}^{(m)}$is the prediction from modality m, and $\alpha_m$is the fusion weight. Late fusion is flexible and allows modality-specific encoders to be optimized independently. However, because fusion occurs only at the decision level, this method often fails to exploit deep interactions between modalities.

Intermediate fusion has become the most common strategy in modern multimodal learning. Each modality is first encoded into a latent representation:

$$H^{(m)} = f_m\left(X^{(m)}\right) \tag{4}$$

And then these high-level representations are combined through a fusion module:

$$Z = \phi\left(H^{(1)}, H^{(2)}, \ldots, H^{(M)}\right) \tag{5}$$

Here, $f_m(\cdot)$denotes the encoder for modality m, and $\phi(\cdot)$denotes the fusion function, which may be concatenation, attention, gating, bilinear interaction, or another learnable mechanism. Intermediate fusion is generally more effective because it preserves modality-specific structure while still allowing the model to learn cross-modal relationships.

In multimodal neuroimaging, the choice of fusion strategy depends heavily on the representation type of each modality. If all modalities are already encoded as region-level vectors, then intermediate fusion with attention or gating is often effective. If some modalities are graphs and others are tabular phenotype features, then a hybrid fusion architecture is needed. Therefore, multimodal fusion is closely tied to encoder design, which motivates the use of specialized models such as graph neural networks and Transformers. As Shown in the Figure 2 below, Comparison of early fusion, intermediate fusion, and late fusion strategies in multimodal brain imaging analysis.

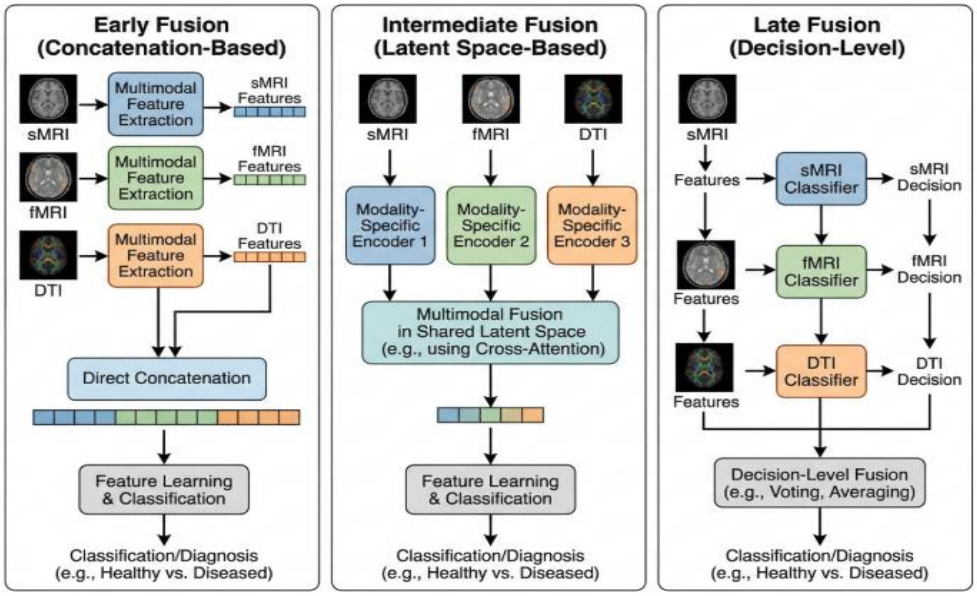

Figure 2. Comparison of early fusion, intermediate fusion, and late fusion strategies in multimodal brain imaging analysis

(4) Cross-Modal Attention and Adaptive Fusion

One of the most important recent developments in multimodal learning is the use of attention mechanisms to enable adaptive fusion. Rather than treating all modalities as equally informative, attention-based models can learn which modality, region, or feature dimension deserves more emphasis for a given task or even for a given subject.

A generic attention mechanism can be formulated as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{6}$$

Where Q, K, and V denote query, key, and value matrices, respectively. In cross-modal settings, one modality may provide the query while another provides keys and values. This allows the model to compute how strongly one modality should attend to another.

For example, if $H^{(f)}$ denotes functional features and $H^{(s)}$ denotes structural features, then a cross-modal interaction can be written as

$$\tilde{H}^{(f)} = \text{Attention}\left(Q^{(f)}, K^{(s)}, V^{(s)}\right) \tag{7}$$

Which means that functional representation is refined under structural guidance. Similar formulations can be used in the opposite direction or extended to multiple modalities.

The major advantage of adaptive attention-based fusion is that it moves beyond simple averaging or concatenation. Instead, it learns task-dependent and data-dependent interaction patterns. This is particularly useful in brain disease classification because different modalities may have different diagnostic value across subjects and disease stages. Attention-based fusion also naturally connects multimodal learning with Transformer architectures, which are fundamentally built on stacked self-attention and cross-attention operations [9].

(5) Challenges in Multimodal Brain Imaging

Despite its advantages, multimodal neuroimaging remains challenging. First, full multimodal datasets are rare. Many public databases contain incomplete modalities across subjects, and the number of subjects who possess all required scans is often limited. Second, modality heterogeneity introduces complex alignment problems. Third, fusion models with high representational power can easily overfit small datasets. Fourth, interpretability becomes more difficult as fusion mechanisms become more sophisticated.

Another important issue is that multimodal information may be complementary but not always equally reliable. Imaging quality can vary across scans, and some modalities may be noisier or less informative in certain cases. Therefore, future multimodal frameworks need to be not only powerful but also robust, uncertainty-aware, and clinically interpretable.

3. Graph Neural Networks for Brain Connectome Analysis

(1) Why Graph Neural Networks Are Suitable for Brain Imaging

The brain is naturally organized as a network. Whether structural, functional, or multimodal relationships are considered, the connectome provides a graph-based representation in which regions are nodes and pairwise interactions are edges. This makes graph neural networks especially suitable for neuroimaging analysis [10].

Shown in the Figure 3 below, Brain connectome representation and graph neural network modeling for neuroimaging analysis.

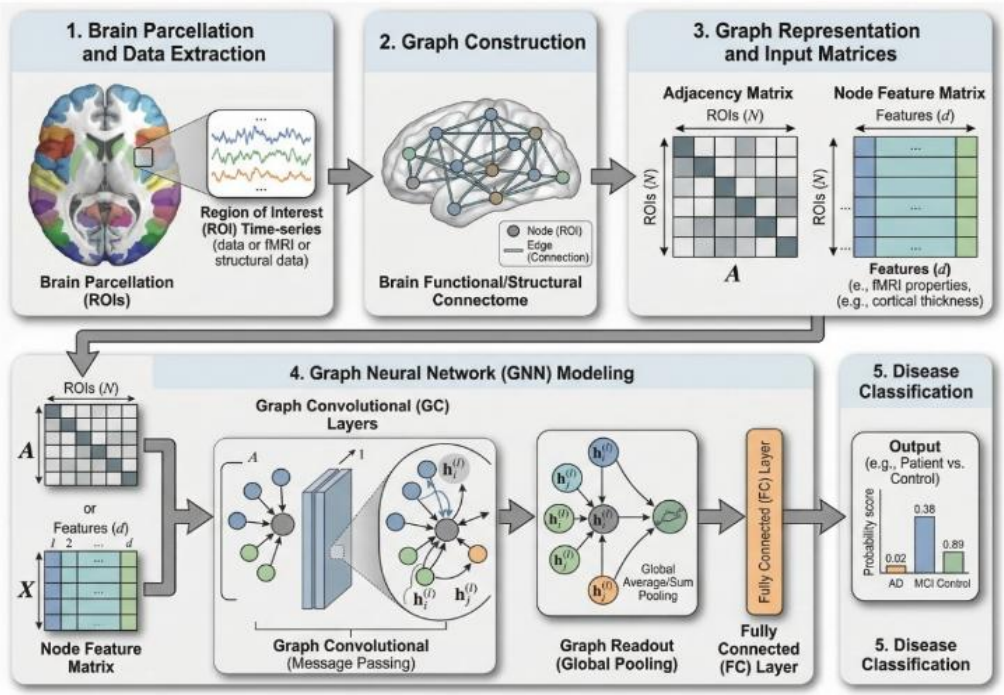

Figure 3. Brain connectome representation and graph neural network modeling for neuroimaging analysis

A brain graph can be defined as

$$G = (V, E, X) \tag{8}$$

Where V is the set of nodes, E is the set of edges, and X is the node feature matrix. In connectome analysis, V usually corresponds to ROIs, E corresponds to structural or functional connectivity, and X may contain morphological, signal-based, or multimodal regional attributes.

Traditional learning methods usually flatten the graph into hand-engineered features or summary statistics. While this may reduce complexity, it discards rich topological information. In contrast, GNNs directly propagate and aggregate information over the graph structure, preserving local connectivity patterns and global network organization.

A generic graph message passing layer can be written as

$$h_i^{(l+1)} = \sigma\left(W^{(l)} h_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} W_n^{(l)} h_j^{(l)} \right) \tag{9}$$

Where $h_i^{(l)}$ denotes the representation of node i at layer l, $\mathcal{N}(i)$ is the set of neighboring nodes, $W^{(l)}$ and $W_n^{(l)}$ are learnable parameters, c_{ij} is a normalization constant, and σ is an activation function. This update captures the key idea that each node representation should depend on its own state as well as the states of connected nodes [11].

(2) Because disease-related abnormalities are often distributed across interconnected brain systems rather than isolated regions, such graph-based neighborhood aggregation is highly meaningful for brain disorder classification.

Graph Convolutional Networks in Neuroimaging

Among GNN models, graph convolutional networks have been especially influential in brain disease classification. A commonly used GCN layer is formulated as

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)} \right) \tag{10}$$

Where $\tilde{A} = A + I$ is the adjacency matrix with self-loops, $\tilde{D}$ is the corresponding degree matrix, $H^{(l)}$ is the feature matrix at layer l, and $W^{(l)}$ is a trainable weight matrix.

This formulation effectively smooths and propagates information across connected nodes. In neuroimaging tasks, GCNs have been used to classify AD, MCI, ASD, and other conditions by learning from functional or structural brain graphs. Their success comes from the fact that they can exploit the full connectivity structure rather than relying on handcrafted graph statistics.

However, standard GCNs also have limitations. If too many graph convolution layers are stacked, node representations may become overly similar, a phenomenon known as over-smoothing. In addition, fixed adjacency-based aggregation may not adequately distinguish the relative importance of different neighbors. These limitations motivated the development of more flexible graph architectures, including graph attention networks and dynamic graph models [12].

(3) Graph Attention Networks and Adaptive Neighborhood Modeling

Graph attention networks address the limitation of uniform aggregation by assigning different weights to different neighbors. Instead of assuming that all connected regions contribute equally, GAT learns adaptive coefficients:

$$\alpha_{ij} = \frac{exp\left(\text{LeakyReLU}\left(a^{\top}[Wh_i \| Wh_j]\right)\right)}{\sum_{k \in \mathcal{N}(i)} exp\left(\text{LeakyReLU}(a^{\top}[Wh_i \| Wh_k])\right)} \tag{11}$$

Where a and W are trainable parameters. The updated node representation is then computed as a weighted sum of neighboring features.

This mechanism is particularly appealing in brain imaging because not all connections are equally relevant to disease classification. Some brain regions or pathways may carry stronger pathological signals than others. By learning attention coefficients, GAT can highlight critical interactions and improve interpretability [12].

Attention on graphs also forms a conceptual bridge between GNNs and Transformers. While GAT focuses attention on graph neighborhoods, Transformer-style attention can be extended to broader relational contexts, including whole-graph interactions and temporal dependencies.

(4) Graph-Level Representation and Brain Disease Classification

The final goal in many neuroimaging tasks is graph-level classification rather than node-level prediction. That is, the entire brain graph must be mapped to a disease label. To achieve this, GNN models typically combine node-level representation learning with graph readout or pooling operations [13].

Let $H \in \mathbb{R}^{N\times d}$ denote the final node representation matrix. A graph-level embedding can be obtained through a readout function:

$$z_G = \text{Readout}(H) \tag{12}$$

Where the readout may be mean pooling, max pooling, attention pooling, or hierarchical graph pooling. The final classification is then given by

$$\hat{y} = \text{MLP}(z_G) \tag{13}$$

The choice of readout strategy matters because disease-related patterns may not be uniformly distributed across nodes. Simple average pooling may dilute important localized signals, while attention-based readout can focus on informative regions.

In multimodal brain classification, graph-level representation may be derived from one modality or from several modality-specific graphs. These graph embeddings can then be fused with phenotype vectors or additional modality features, further emphasizing the close connection between GNN design and multimodal fusion.

(5) Dynamic Graph Neural Networks for Functional Connectivity

One of the most important limitations of static graph analysis is that it ignores temporal variation in functional connectivity. Brain activity is inherently dynamic, and the interactions among regions fluctuate over time. To address this issue, dynamic graph neural networks have been developed.

Given an fMRI sequence, a sliding-window approach can be used to construct a sequence of graphs:

$$\mathcal{A} = A^{(1)}, A^{(2)}, \ldots, A^{(T_w)} \tag{14}$$

Where each $A^{(t)}$ corresponds to the functional connectivity matrix in window t. The GNN encoder is then applied to each graph:

$$H^{(t)} = \text{GNN}\left(A^{(t)}, X^{(t)}\right) \tag{15}$$

This produces a sequence of graph embeddings or node representations, which can then be processed by a temporal module. Compared with static analysis, dynamic graph learning can capture transient states, switching behavior, and evolving connectivity patterns that may be closely related to disease progression.

Shown in the Figure 4 below, Dynamic functional connectivity construction and graph sequence learning based on sliding-window fMRI analysis.

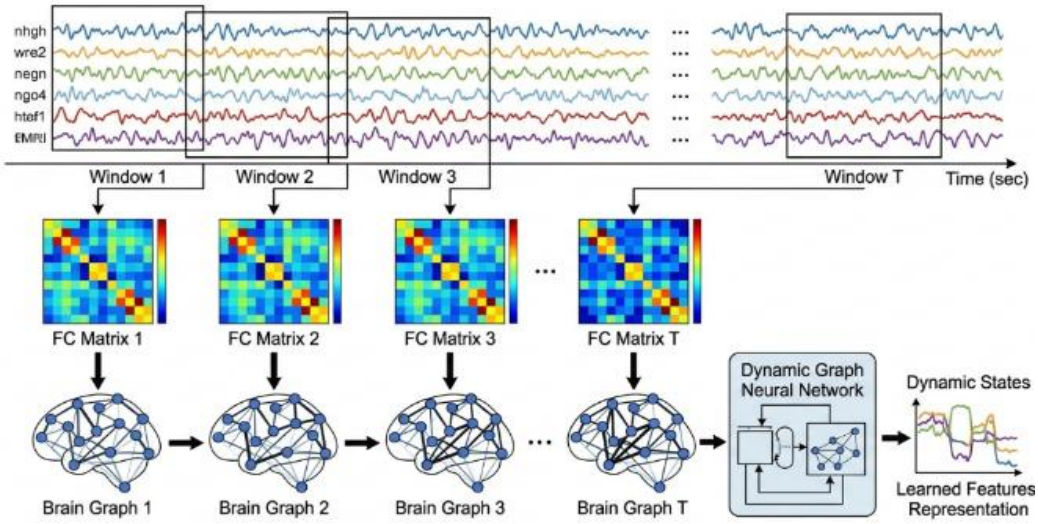

Figure 4. Dynamic functional connectivity construction and graph sequence learning based on sliding-window fMRI analysis

Dynamic graph models are especially relevant to your work because they naturally align with the idea of combining DGCN and Transformer. The graph component models spatial topology in each window, while the temporal component captures dependencies across windows [14].

(6) Limitations of Current GNN Methods in Neuroimaging

Although GNNs have achieved strong results, they are not without limitations. First, the quality of graph learning depends heavily on how the graph is constructed. Choices such as atlas definition, thresholding strategy, and edge weighting can substantially affect results. Second, many GNN methods assume a fixed graph, which may not be appropriate for dynamic or subject-adaptive settings. Third, graph models can become difficult to interpret when deep or highly nonlinear. Fourth, many studies still use relatively small datasets, making it difficult to evaluate generalization rigorously.

These limitations motivate the integration of GNNs with more flexible modeling frameworks, especially Transformer-based architectures that can complement graph learning with stronger global interaction modeling [15].

4. Transformer Models for Multimodal and Spatiotemporal Brain Imaging

(1) Why Transformer Matters in Neuroimaging

The Transformer has become one of the most influential model architectures in modern machine learning. Its central innovation is self-attention, which allows every element in a sequence to interact directly with every other element. Unlike recurrent models, Transformers can model long-range dependencies more effectively and in a more parallelizable manner.

In neuroimaging, this capability is highly valuable for

several reasons. First, dynamic brain graphs involve temporal dependencies across windows that may extend beyond local neighborhoods. Second, multimodal fusion requires learning interactions across heterogeneous feature sets. Third, even within a single graph, long-range relations between distant but functionally linked regions may be important. These considerations make Transformer-based modeling highly attractive [16].

The standard attention operation is given by

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{16}$$

For self-attention, Q, K, and V are all projections of the same input. For cross-attention, they may come from different modalities or stages of the model. This flexibility is one reason why Transformers are so well suited for multimodal learning.

(2) Transformer for Temporal Modeling of Dynamic Brain Graphs

In dynamic brain imaging, each time window yields a graph representation $H^{(t)}$. These representations can be treated as a sequence and fed into a Transformer:

$$H_t = \text{Transformer}\left(H^{(1)}, H^{(2)}, \dots, H^{(T_w)}\right) \tag{17}$$

The Transformer encoder captures long-range temporal dependencies among windows. Unlike ordinary recurrent models, it does not assume a strictly sequential bottleneck and can therefore model relationships between distant time windows more directly.

This is particularly important in brain disorder classification because disease-related temporal patterns may not be defined by only local transitions. For example, certain recurring connectivity states or long-range fluctuations may distinguish patients from healthy controls. Transformer-based temporal modeling provides a natural way to capture such global sequence structure.

To preserve order information, positional encoding is usually added:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{18}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{19}$$

This helps the Transformer distinguish the temporal order of graph embeddings. In spatiotemporal brain modeling, positional information is essential because the meaning of a connectivity pattern often depends on when it occurs.

(3) Transformer for Multimodal Interaction Learning

Beyond temporal modeling, Transformer can also serve as a powerful multimodal fusion module. Suppose modality-specific encoders produce representations $H^{(1)}, H^{(2)}, \dots, H^{(M)}$. These representations can be treated as modality tokens and processed jointly by a Transformer encoder. In this way, the model learns interactions among modalities in a shared attention space.

A multimodal Transformer can be expressed as

$$Z = \text{Transformer}\left(H^{(1)}, H^{(2)}, \dots, H^{(M)}\right) \tag{20}$$

Alternatively, cross-attention can be used to let one modality attend selectively to another. This is beneficial because imaging modalities do not contribute equally to all aspects of diagnosis. For example, structural features may help guide the interpretation of functional abnormalities, while functional signals may highlight disease-relevant effects not visible in anatomy alone.

Transformer-based multimodal learning is especially attractive because it unifies modality interaction, feature weighting, and higher-order dependency modeling within a single framework. Compared with ordinary concatenation or shallow attention mechanisms, it offers much richer fusion capacity [17].

(4) GNN-Transformer Hybrid Architectures

Perhaps the most promising direction in current research is the combination of GNNs and Transformers. These two architectures are complementary. GNNs are strong at encoding local graph topology and structured neighborhood interactions. Transformers are strong at modeling global dependencies and flexible cross-element relations. By combining them, a model can capture both local and global aspects of brain organization.

A typical hybrid architecture can be written as

$$H_s^{(t)} = \text{GNN}\left(A^{(t)}, X^{(t)}\right) \tag{21}$$

$$H_t = \text{Transformer}\left(H_s^{(1)}, H_s^{(2)}, \dots, H_s^{(T_w)}\right) \tag{22}$$

$$\hat{y} = \text{MLP}\left(\text{Readout}(H_t)\right) \tag{23}$$

In this design, the GNN acts as a spatial encoder for each graph snapshot, while the Transformer acts as a temporal encoder over graph embeddings. If multiple modalities are involved, the architecture can be further extended so that each modality has its own graph encoder, after which cross-modal Transformer blocks are used for fusion.

This paradigm is directly aligned with modern multimodal neuroimaging research. It allows the model to integrate local connectome topology, temporal dynamics, and cross-modal interactions within one unified framework. Shown in the Figure 5 below, A multimodal GNN-Transformer framework for spatiotemporal brain disease classification. For disorders characterized by distributed and time-varying abnormalities, such architectures are particularly promising.

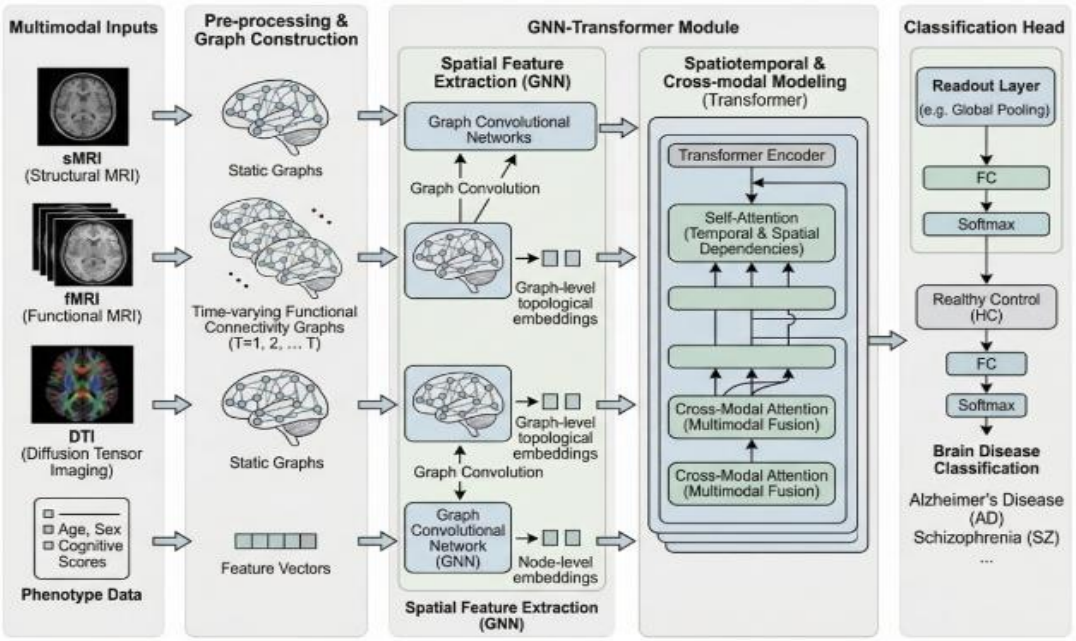

Figure 5. A multimodal GNN-Transformer framework for spatiotemporal brain disease classification

(5) Strengths and Limitations of Transformer-Based Methods

Transformer-based methods offer several clear strengths. They can model long-range dependencies, support flexible interaction structures, and naturally implement attention-based fusion. They are also highly modular and can be combined with GNNs, CNNs, or tabular encoders.

However, they also have limitations. First, self-attention can be computationally expensive, especially when applied to long sequences or large graphs. Second, Transformer models often require more data than simpler architectures to train effectively. Third, in neuroimaging, where datasets are often relatively small, careful regularization and architectural design are necessary to avoid overfitting. Fourth, although attention maps are sometimes interpreted as explanations, the interpretability of Transformers remains an open issue.

Nevertheless, given their flexibility and modeling power, Transformers are likely to remain central to future multimodal neuroimaging research [18].

5. Integrating Multimodal Learning, GNNs, and Transformer

The most important methodological trend in recent years is the convergence of multimodal learning, graph neural networks, and Transformers into a unified pipeline. This integration is not accidental. Each of these components solves a different but closely related problem.

Multimodal learning addresses the heterogeneity of information sources. GNNs address the graph-structured nature of brain connectomes. Transformers address long-range and cross-modal dependency modeling. Together, they form a natural architecture for modern brain disorder classification.

A general integrated framework can be summarized as follows. First, each modality is encoded separately, with graph-based modalities using GNN encoders and non-graph modalities using suitable MLP or CNN encoders:

$$H^{(m)} = f_m\left(X^{(m)}\right), m = 1,2,\dots,M \tag{24}$$

Second, if dynamic graphs are used, each temporal slice is encoded through a graph module:

$$H^{(m\cdot t)} = \mathrm{GNN}\left(A^{(m\cdot t)}, X^{(m\cdot t)}\right) \tag{25}$$

Third, a Transformer module performs temporal modeling and/or cross-modal interaction learning:

$$Z = \mathrm{Transformer}\left(\{H^{(m\cdot t)}\}\right) \tag{26}$$

Finally, a readout and classifier produce the prediction:

$$\hat{y} = \mathrm{Classifier}\left(\mathrm{Readout}(Z)\right) \tag{27}$$

This unified view is helpful because it clarifies how different methodological innovations relate to one another. Rather than treating multimodal fusion, graph learning, and Transformer modeling as isolated design choices, it shows that they are components of a common spatiotemporal and cross-modal representation learning framework.

6. Current Challenges and Future Directions

Despite significant progress, several challenges remain.

The first is data limitation. Complete multimodal datasets are still small, especially when strict inclusion criteria are applied. This restricts the reliable training of deep hybrid models. Self-supervised learning, transfer learning, and data-efficient pretraining may therefore become increasingly important.

The second is robustness to missing modalities. In clinical practice, some subjects may lack one or more scans. Future multimodal frameworks should be able to degrade gracefully when modalities are missing rather than failing entirely.

The third is model interpretability. Clinical application requires understanding not only whether a subject is classified as diseased but also which brain regions, networks, or temporal patterns contributed to that decision. Attention visualization, region importance scoring, and biologically informed graph priors may help improve interpretability.

The fourth is graph construction uncertainty. Since connectome definition depends on preprocessing choices, future models may benefit from learning graph structure adaptively rather than assuming a fixed adjacency matrix.

The fifth is generalization across centers and scanners.

Multi-site domain shift remains a major obstacle to real-world deployment. Domain adaptation, federated learning, and harmonization strategies are therefore promising future directions.

Finally, future research should move toward clinically meaningful evaluation. Rather than focusing only on accuracy, it is important to consider robustness, calibration, interpretability, and the ability to support individualized disease assessment.

7. Conclusion

Multimodal brain imaging has become a key research direction in neuroimaging-based diagnosis because it integrates complementary structural, functional, and connectivity information about the brain. At the same time, graph neural networks provide a natural and effective way to model connectome data, while Transformer architectures offer powerful tools for learning long-range temporal and cross-modal dependencies. The combination of these three methodological directions has created a promising paradigm for brain disease classification.

Current research shows that multimodal fusion can improve the completeness and robustness of disease representation, GNNs can capture topology-aware brain network patterns, and Transformers can model complex interactions across time and modality. However, important challenges remain, including limited sample size, modality heterogeneity, graph uncertainty, limited interpretability, and weak cross-center generalization. Addressing these issues will be crucial for translating advanced computational models into practical clinical tools. Overall, the joint development of multimodal learning, graph neural networks, and Transformer-based modeling is expected to continue shaping the future of intelligent brain disorder diagnosis.

Acknowledgements

We thank all the anonymous reviewers for their hard reviewing work

References

[1] Chen Jian, Zheng Li, Hu Yuzhu, et al. Traffic flow matrix-based graph neural network with attention mechanism for traffic flow prediction [J]. Information Fusion, 2024, 104: 102146.

[2] Du Xingze, Liu Huizhou, Shen Bowen, et al. Detection-driven adaptive semantic feature weight for multi-modality image fusion [J]. Engineering Applications of Artificial Intelligence, 2026, 163: 112874.

[3] França Lucas GS, Ciarrusta Judit, Gale-Grant Oliver, et al. Neonatal brain dynamic functional connectivity in term and preterm infants and its association with early childhood neurodevelopment [J]. Nature Communications, 2024, 15(1): 16.

[4] Guo Meng-Hao, Xu Tian-Xing, Liu Jiang-Jiang, et al. Attention mechanisms in computer vision: A survey [J]. Computational visual media, 2022, 8(3): 331-68.

[5] Liao Wenlong, Bak-Jensen Birgitte, Pillai Jayakrishnan Radhakrishna, et al. Short-term power prediction for renewable energy using hybrid graph convolutional network and long short-term memory approach [J]. Electric Power Systems Research, 2022, 211: 108614.

[6] Liu Jinyuan, Lin Runjia, Wu Guanyao, et al. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion [J]. International Journal of Computer Vision, 2024, 132(5): 1748-75.

[7] Liu Qinghao, Zhu Yuehao, Liu Min, et al. MBUNeXt: Multibranch Encoder Aggregation Network Based on Layer-Fusion Strategy for Multimodal Brain Tumor Segmentation [J]. IEEE Transactions on Neural Networks and Learning Systems, 2025.

[8] Yang Guangrui, Li Ming, Feng Han, et al. Deeper insights into deep graph convolutional networks: Stability and generalization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[9] Zhou Jie, Jie Biao, Wang Zhengdong, et al. LCGNet: Local sequential feature coupling global representation learning for functional connectivity network analysis with fMRI [J]. IEEE Transactions on Medical Imaging, 2024, 43(12): 4319-30.

[10] Bessadok Alaa, Mahjoub Mohamed Ali, Rekik Islem. Graph neural networks in network neuroscience [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(5): 5833-48.

[11] Cao Tangwei, Lin Runwei, Zheng Yinuo, et al. A novel approach analysing the dynamic brain functional connectivity for improved MCI detection [J]. IEEE Transactions on Biomedical Engineering, 2023, 71(1): 207-16.

[12] Chen Zhiqian, Chen Fanglan, Zhang Lei, et al. Bridging the gap between spatial and spectral domains: A unified framework for graph neural networks [J]. ACM Computing Surveys, 2023, 56(5): 1-42.

[13] Cheung Liege, Wang Yun, Lau Adela SM, et al. Using a novel clustered 3D-CNN model for improving crop future price prediction [J]. Knowledge-Based Systems, 2023, 260: 110133.

[14] Dai Quanyu, Wu Xiao-Ming, Xiao Jiaren, et al. Graph transfer learning via adversarial domain adaptation with graph convolution [J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(5): 4908-22.

[15] Ding Chaoyue, Sun Shiliang, Zhao Jing. MST-GAT: A multimodal spatial–temporal graph attention network for time series anomaly detection [J]. Information Fusion, 2023, 89: 527-36.

[16] Liu Tao, Jiang Aimin, Zhou Jia, et al. GraphSAGE-based dynamic spatial–temporal graph convolutional network for traffic prediction [J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(10): 11210-24.

[17] Yadav Satya Prakash, Zaidi Subiya, Mishra Annu, et al. Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN) [J]. Archives of Computational Methods in Engineering, 2022, 29(3): 1753-70.

[18] Zuo Qiankun, Shen Yanyan, Zhong Ning, et al. Alzheimer's disease prediction via brain structural-functional deep fusing network [J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 4601-12.