MAPI-GNN: Multi-Activation Plane Interaction Graph Neural Network for Multimodal Medical Diagnosis

Ziwei Qin*, Xuhui Song*, Deqing Huang, Na Qin, Jun Li†

Institute of Systems Science and Technology, School of Electrical Engineering, Southwest Jiaotong University, China
qinziwei, sxh2023@my.swjtu.edu.cn, qinna, elehd@swjtu.edu.cn, dirk.li@outlook.com

Abstract

Graph neural networks are increasingly applied to multimodal medical diagnosis for their inherent relational modeling capabilities. However, their efficacy is often compromised by the prevailing reliance on a single, static graph built from indiscriminate features, hindering the ability to model patient-specific pathological relationships. To this end, the proposed Multi-Activation Plane Interaction Graph Neural Network (MAPI-GNN) reconstructs this single-graph paradigm by learning a multifaceted graph profile from semantically disentangled feature subspaces. The framework first uncovers latent graph-aware patterns via a multi-dimensional discriminator; these patterns then guide the dynamic construction of a stack of activation graphs; and this multifaceted profile is finally aggregated and contextualized by a relational fusion engine for a robust diagnosis. Extensive experiments on two diverse tasks, comprising over 1300 patient samples, demonstrate that MAPI-GNN significantly outperforms state-of-the-art methods.

Code — https://github.com/HecateBlair/MAPI-GNN

Introduction

Multimodal medical imaging is pivotal for comprehensive disease diagnosis, as it leverages diverse physical principles (Minaee et al. 2020; Fan et al. 2020; Azizi et al. 2021)—such as capturing anatomical structure with MRI and metabolic activity with PET—to provide synergistic insights that are unattainable from any single modality (Behrad and Abadeh 2022; Xie et al. 2021; Li et al. 2024). However, effectively fusing this intrinsically heterogeneous data remains a formidable challenge (Duan et al. 2024; Mi, Li, and Zhou 2020; Alfeo, Cimino, and Vaglini 2022) . The core difficulty lies in reconciling disparate data structures and semantic information to form a cohesive pathological representation (Gao et al. 2021; Boehm et al. 2022). While deep learning models, particularly Convolutional Neural Networks (CNNs) (Zhang et al. 2023; Kuttala, Subramanian, and Oruganti 2023; Wang et al. 2021; He et al. 2020; Zhou et al. 2024; Lecun and Bottou 1998), have advanced feature extraction, their inherent grid-based processing struggles to explicitly model the complex, high-dimensional, and often subtle non-Euclidean relationships between different modalities (Xu et al. 2024; Zhao et al. 2024; Lipkova et al. 2022). This limitation has clinical ramifications: failing to capture complex cross-modal patterns can obscure pathological markers, compromising accuracy and clinical outcomes.

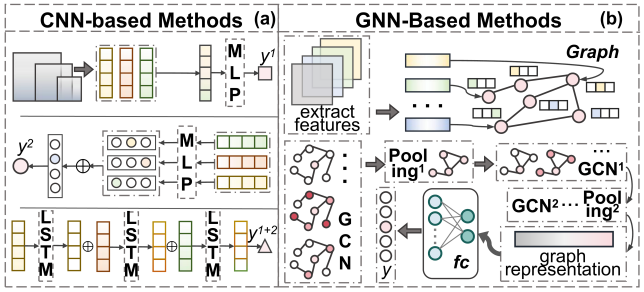

Figure 1: Conceptual fusion strategies. (a) CNN-based methods use fixed fusion points (e.g., early/late), which limits the modeling of complex inter-modal dependencies. (b) GNNs offer a flexible paradigm, using a graph topology for explicit relationship modeling and hierarchical aggregation.

Graph Neural Networks (GNNs) (Wu et al. 2020; Khemani et al. 2024; Wu et al. 2022) offer a powerful framework for relational reasoning. The common GNN-based paradigm (conceptually illustrated in Fig. 1) constructs a graph from pre-extracted features for fusion and analysis (Lin et al. 2023; Tan et al. 2022). While this approach has achieved notable success, we argue its efficacy for handling the unique complexities of multimodal medical data is curtailed by three critical limitations:

1) Indiscriminate feature representation: Existing methods often conflate diagnostically relevant features with redundant or noisy information, which can in turn impair the model's downstream reasoning process. **2) Static graph topology:** They typically rely on a single, predefined graph structure, which is inherently insufficient to capture the diverse and patient-specific relationships within complex multimodal data. **3) Localized message passing:** Most GNNs confine message passing to local neighborhoods, which restricts their ability to capture crucial long-range dependencies and form a holistic, global understanding of the data.

*These authors contributed equally.
†Corresponding author.

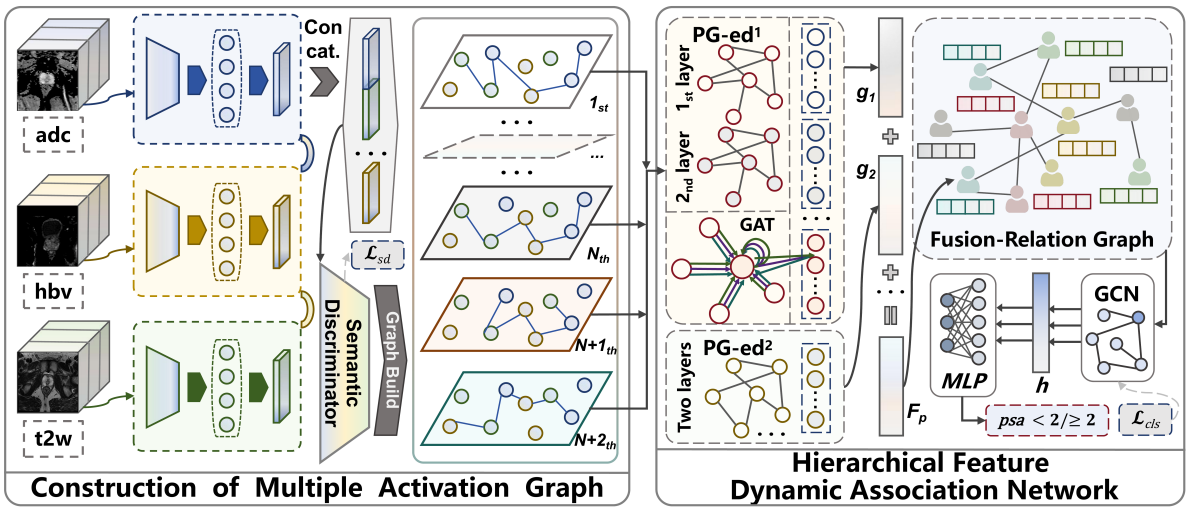

Figure 2: Overview of the MAPI-GNN architecture. Stage I (detailed in Fig. 3) generates multiple, semantically-aware activation graphs from patient-specific multimodal data. Stage II (detailed in Fig. 4) then performs a hierarchical, two-level fusion on these graphs, first modeling intra-sample relationships and then inter-sample dependencies for the final diagnosis.

Thus, we propose the Multi-Activation Plane Interaction Graph Neural Network (MAPI-GNN), which shifts the paradigm from a single, static graph to learning a dynamic, multifaceted graph profile per patient, systematically responding to the three aforementioned limitations. Specifically: **(1)** To mitigate **indiscriminate feature representation**, a Multi-Dimensional Feature Discriminator (MDFD) adaptively identifies salient features via semantic importance. **(2)** To counter **static graph topology**, these salient features guide a Multi-Activation Graph Construction Strategy (MAGCS) to build multiple, semantically-aware topologies. **(3)** Finally, to move beyond **localized message passing**, a Hierarchical Feature Dynamic Association Network (HFDAN) fuses the resulting intra-sample graphs and models inter-sample dependencies in a global graph for robust classification. Our main contributions are three-fold:

- A feature-driven paradigm for dynamic graph construction, where topology is learned from semantic salience rather than being predefined, enabling the model to capture more adaptive, patient-specific relationships.
- A hierarchical architecture that distills intra-sample insights from a manifold of activation graphs and contextualizes them within a global graph, enabling comprehensive and high-fidelity patient-level analysis.
- Comprehensive experimental validation on two diverse medical datasets (multi-parametric MRI, and CT with clinical data), demonstrating that our work achieves state-of-the-art performance and robust generalizability.

Related Work

CNN-based Multimodal Fusion

Effective feature fusion is central to multimodal medical diagnosis (Panayides et al. 2020; Cui et al. 2023). Many approaches are CNN-based (Fig. 1(a)), traditionally employing early fusion (Stahlschmidt, Ulfenborg, and Synnergren 2022; Bayoudh et al. 2022), which is sensitive to noise and misalignment, or late fusion, which combines predictions from modality-specific models but consequently misses crucial low-level correlations (Wang et al. 2019; Sharma, Sharma, and Kumar 2023). While more advanced hybrid strategies, sometimes employing attention mechanisms, exist, their reliance on fixed, grid-based operations fundamentally limits their ability to effectively discern salient information from heterogeneous sources (Karim et al. 2023; Xie et al. 2022). To address this limitation, our work introduces a feature discriminator to guide a more discerning and effective fusion process.

GNN-based Multimodal Fusion

GNN-based classification (Fig. 1(b)) offers a more powerful paradigm for relational modeling (Ding et al. 2024; Fei et al. 2023). Current methods are typically applied at two scales. In the context of medical diagnostics, these scales correspond to distinct analytical granularities, from localized feature analysis to holistic patient-profile assessment. While **node-level** approaches excel at localized tasks (Huang et al. 2024; Chen et al. 2022; Sarkar et al. 2023; Huang, Tang, and Chen 2022), they struggle to form a holistic, patient-level diagnosis due to their inherently limited receptive fields. In contrast, **graph-level** methods can analyze a patient's entire profile (Khoshraftar and Aijun 2024; Song, Li, and Qian 2022), sometimes using multiplex structures (D'Souza et al. 2024; Shao et al. 2020). However, their core limitation is the reliance on predefined, static graph topologies, which implicitly assume that a single relational structure is optimal for all patients. This rigidity prevents the model from adapting to patient-specific pathological markers. To mitigate this bottleneck, our work proposes a dynamic graph construction strategy and a hierarchical fusion network to capture both adaptive intra-sample and global inter-sample relationships.

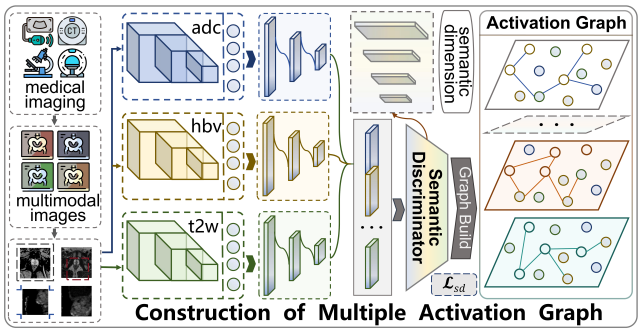

Figure 3: Workflow of Stage I: Multi-Activation Graph Construction. From compressed modality features, a Multi-Dimensional Feature Discriminator identifies salient activated features for multiple semantic dimensions, each guiding the construction of a corresponding activation graph.

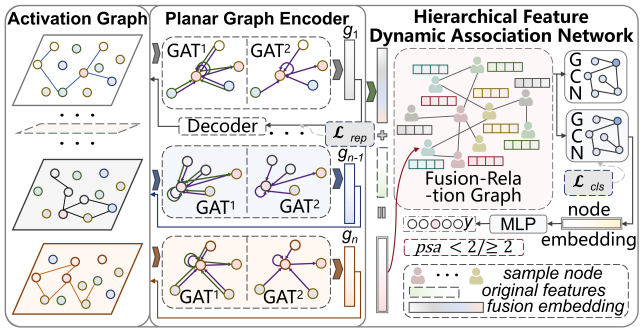

Figure 4: Workflow of the Hierarchical Feature Dynamic Association Network (Stage II): 1) Intra-sample encoding of multiple activation graphs into representations ($\mathbf{g}_m$); and 2) Inter-sample classification on a global graph of fused patient vectors ($\mathbf{F}_p$) processed by a GCN.

Method

Overall Architecture

Our proposed two-stage framework (Fig. 2) counters the static single-graph paradigm by learning a patient-specific, multifaceted graph profile. First, a stack of semantically-aware activation graphs is dynamically constructed from disentangled feature subspaces. This profile is then hierarchically aggregated, modeling both intra- and inter-sample relationships to yield a robust final diagnosis.

Stage I: Multi-Activation Graph Construction

This stage (Fig. 3) generates structured graphs from raw features by identifying discriminative feature subsets and using them to define graph topologies.

Multi-Dimensional Feature Discriminator Our Multi-Dimensional Feature Discriminator assesses feature importance across multiple learned semantic dimensions. Initial modality features, extracted and compressed via autoencoders, are concatenated into a vector $\mathbf{x} \in \mathbb{R}^C$. This vector is projected by the Multi-Dimensional Feature Discriminator, $F_{sd} : \mathbb{R}^C \to \mathbb{R}^M$, into an M-dimensional semantic space. The influence of feature i on semantic dimension m, $\mathbf{C}_m(i)$, is quantified via perturbation:

$$C_m(i) = \left| [F_{sd}(\mathbf{x})]_m - [F_{sd}(\hat{\mathbf{x}}^{(i)})]_m \right| \tag{1}$$

where $\hat{\mathbf{x}}^{(i)}$ is $\mathbf{x}$ with i-th feature ablated. For each dimension m, features with the highest influence scores are designated as activated features. To ensure the discriminator learns robust, disentangled semantic dimensions, training is guided by a composite loss, $\mathcal{L}_{sd}$, combining a primary reconstruction loss ($\mathcal{L}_{AE}$) with a regularization term $\mathcal{L}_{reg}(\Theta_{sd})$:

$$\mathcal{L}_{sd} = \mathcal{L}_{AE}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{reg}(\Theta_{sd}) \tag{2}$$

where, $\mathcal{L}_{reg}$ enforces desirable properties on the discriminator's weights Θ_{sd} by combining L1 and L2 regularization, and an orthogonality constraint on its linear layers $\lambda_{orth}\|W_{sd}W_{sd}^T - I\|_F^2$ to promote semantic independence. Here, $||\cdot||_F$ denotes the Frobenius norm and λ_{orth} is its corresponding weight coefficient.

Multi-Activation Graph Construction Strategy The Multi-Activation Graph Construction Strategy builds a unique activation graph $\mathcal{G}_m$ for each semantic dimension m. All M graphs share a common set of C nodes $\mathcal{V}$, corresponding to the feature dimensions of $\mathbf{x}$. For each graph $\mathcal{G}_m$, the edge set $\mathcal{E}_m$ connects activated nodes to their k-nearest activated neighbors (via Euclidean distance in the initial feature space). We propose a edge weighting scheme where the weight $w_{ij}^{(m)}$ is the average influence of the connected nodes:

$$w_{ij}^{(m)} = \frac{1}{2}(C_m(i) + C_m(j)) \tag{3}$$

This yields a set of M graphs, $\{\mathcal{G}_1, \ldots, \mathcal{G}_M\}$, each offering a unique and complementary semantic view of the feature relationships for a given patient.

Stage II: Hierarchical Feature Dynamic Association Network

The Hierarchical Feature Dynamic Association Network (Fig. 4) is tasked with processing the activation graphs from Stage I, hierarchically fusing their diverse representations to yield a final diagnosis.

Intra-Sample Encoding and Fusion Each activation graph $\mathcal{G}_m$ is processed by a Planar Graph Encoder (implemented with GAT (Velickovic et al. 2017)), yielding a 32-dimensional graph-level representation $\mathbf{g}_m$. Crucially, our pre-defined sparse topology $\mathcal{E}_m$ complements the GAT by constraining its attention mechanism to a small, semantically meaningful subset of feature relationships, guiding the model toward discriminative information. Furthermore, the learned attention coefficients are modulated by our pre-defined edge weights $w_{ij}^{(m)}$, ensuring that both learned patterns and a priori feature importance guide the final aggregation. The GAT layer aggregates neighbor information as:

$$\mathbf{h}_i' = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W} \mathbf{h}_j\right) \tag{4}$$

Method	ACC ↑	AUC ↑	PRE ↑	REC ↑	F1 ↑	SPE
CNN-based Methods						
LFF	0.8030	0.9605	0.9217	0.6364	0.7636	**0.9697**
MSC	0.8522	0.9380	0.8708	0.8181	0.8470	0.8863
DenseNet	0.7386	0.8078	0.7273	0.8333	0.7767	0.6250
CNN	0.7045	0.7867	0.7115	0.7708	0.7400	0.6250
GNN-based Methods						
SAGE	0.8068	0.8790	0.8650	0.7270	0.7900	0.8860
GTAD	0.8300	0.8200	0.7640	**0.9550**	0.8480	0.7850
MGNN-CMSC	0.8939	0.9592	0.8810	0.9487	0.9136	0.8148
LG-GNN	0.8977	0.9724	0.9149	0.8958	0.9053	0.9000
HGM2R	0.9242	0.9798	0.9246	0.9242	0.9242	0.9394
Transformer-based Methods						
ViT	0.9053	0.9728	0.8587	0.9491	0.9145	0.8069
Conventional Fusion Methods						
Early-fusion	0.7500	0.8255	0.7500	0.8125	0.7800	0.6750
Late-fusion	0.6591	0.6576	0.6667	0.7500	0.7059	0.5500
MAPI-GNN (Ours)	**0.9432**	**0.9838**	**0.9361**	0.9545	**0.9438**	0.9318

Table 1: Performance of different methods on the PI-CAI dataset. The best results are highlighted in **bold**.

where α_{ij} is the learned attention coefficient and $\mathbf{h}_j$ is the node's representation. A readout function then produces the graph representation $\mathbf{g}_m = \frac{1}{C}\sum_{i=1}^{C}\mathbf{h}_i^{(L)}$, by averaging the final node representations from the L-th (final) layer.

These M graph representations are concatenated with the initial feature vector $\mathbf{x}_p$ to form patient-level feature $\mathbf{F}_p$, preserving the complete feature profile from all M semantic views and the original features in a lossless manner:

$$\mathbf{F}_p = \text{Concat}(\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_M, \mathbf{x}_p) \tag{5}$$

To ensure the learned embeddings are informative, the Planar Graph Encoders are regularized by a representation loss, $\mathcal{L}_{rep}$, which penalizes the reconstruction error when decoding initial node features from the final embeddings:

$$\mathcal{L}_{rep} = \frac{1}{N \cdot M}\sum_{p=1}^{N}\sum_{m=1}^{M}\mathcal{L}_{MSE}\left(\mathbf{H}_{m,p}^{(0)}, \text{Decoder}(\mathbf{H}_{m,p}^{(L)})\right) \tag{6}$$

where $\mathbf{H}_{m,p}^{(0)}$ and $\mathbf{H}_{m,p}^{(L)}$ are the initial and final (L-th layer) node features for the m-th graph of the p-th patient.

Inter-Sample Classification To model inter-sample relationships, we construct a global Fusion-Relation Graph, whose nodes are patients featured by $\mathbf{F}_p$. This global graph is processed by a Graph Convolutional Network (GCN (Kipf and Welling 2016)), whose layer-wise propagation is:

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \tag{7}$$

where $\tilde{\mathbf{A}}$ is the adjacency matrix with self-loops, $\tilde{\mathbf{D}}$ is diagonal degree matrix, and $\mathbf{H}^{(l)}$ is the l-th layer's representation.

The final representations are fed to an MLP for classification. This final stage is supervised by the primary task objective, the standard Cross-Entropy (CE) loss:

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}(Y, \hat{Y}) \tag{8}$$

where Y and $\hat{Y}$ are the ground-truth labels and the model predictions, respectively.

Overall Training Objective

The proposed architecture is jointly trained end-to-end by minimizing a final composite objective function. This objective is carefully formulated as a weighted sum of the aforementioned loss components to effectively balance their respective contributions:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{rep}\mathcal{L}_{rep} + \lambda_{sd}\mathcal{L}_{sd} \tag{9}$$

where λ_{cls}, λ_{rep}, and λ_{sd} are hyperparameters to balance the losses. We set these to $\lambda_{cls} = 1.0$, $\lambda_{rep} = 0.3$, and $\lambda_{sd} = 1.0$ to prioritize the main classification task and the semantic disentanglement, while applying moderate regularization to the representation encoding.

Experiments

Datasets and Experimental Setup

Our primary benchmark is **the public PI-CAI 2022 Challenge dataset**, used for clinically significant Prostate Cancer (csPCa) classification. The modalities for this task include T2w, ADC, and HBV MRI. We construct our study cohort from the 220 csPCa cases with trusted expert annotations, pairing them with 220 randomly under-sampled benign cases to form a balanced 440-case dataset. Preprocessing follows the prior study (Kan et al. 2022). To validate the robustness and generalizability of our proposed architectural components, we also utilize **a multi-modal Coronary Heart Disease (CHD) dataset** for diagnosis (974 cases, under IRB No. KY2025331). The modalities for this task are CCTA (Coronary CT Angiography) scans and structured clinical data. For a rigorous and reproducible evaluation, all models are assessed using a five-fold cross-validation protocol with a fixed random seed. We report the mean of key performance metrics, including Accuracy (ACC), Area Under the Curve (AUC), F1-Score (F1), Precision (PRE), Recall (REC), and Specificity (SPE).

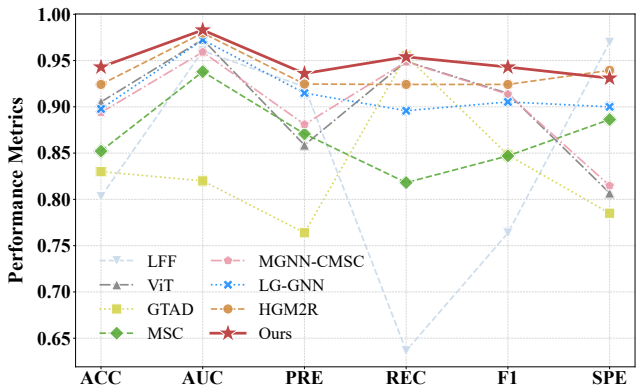

Figure 5: Visual comparison of MAPI-GNN against key baselines on the PI-CAI dataset. This figure visualizes the primary metrics presented in Table 1, highlighting our model's competitive performance.

Comparison with State-of-the-art Methods

Compared methods. To provide a comprehensive evaluation, we benchmark our proposed architecture against a diverse set of methods, spanning from foundational benchmarks to recent state-of-the-art (SOTA) approaches. These are grouped into four categories: 1) **CNN-based methods**, representing advanced architectures for multimodal learning, such as LFF (Holste et al. 2019) and MSC (Haarburger et al. 2019); 2) a wide range of **GNN-based methods** that constitute the current state-of-the-art in this domain, including GraphSAGE (Hamilton, Ying, and Leskovec 2017), GTAD (Sim et al. 2024), MGNN-CMSC (Li and Nabavi 2024), LG-GNN (Zhang et al. 2022), and HGM2R (Feng et al. 2023); 3) **Transformer-based Methods**, such as a standard Vision Transformer (ViT); and 4) **conventional fusion strategies** (Early and Late fusion) which serve as fundamental benchmarks. To ensure a fair and reproducible comparison, all models were re-implemented or run using their official code under identical experimental settings.

Results Analysis. The quantitative results in Table 1 demonstrate our model's state-of-the-art performance. A key observation is the general superiority of GNN-based methods over their CNN-based counterparts. For instance, the strongest GNN baseline (HGM2R, 0.9242 ACC) outperforms the best-performing CNN (MSC, 0.8522 ACC) and ViT (0.9053 ACC) baselines, affirming the advantages of explicit relational modeling for this task. Building upon this paradigm, our framework further advances the SOTA by outperforming the highly competitive HGM2R with a notable margin of 1.9 percentage points in Accuracy (0.9432 vs. 0.9242). Crucially, our model's ability to achieve top scores across multiple key metrics simultaneously demonstrates a well-balanced and robust performance, rather than being narrowly optimized for a single metric. This dominant performance validates the synergistic effectiveness of our core design principles: dynamic, feature-driven graph construction and our hierarchical fusion architecture.

Method	ACC	AUC	PRE	REC	F1
w/o MDFD	0.8500	0.9137	0.8515	0.8636	0.8533
w/o MAGCS	0.8205	0.9115	0.8024	0.8545	0.8266
w/o HFDAN	0.8364	0.9153	0.8234	0.8591	0.8402
MAPI-GNN	**0.9432**	**0.9838**	**0.9361**	**0.9545**	**0.9438**

Table 2: Ablation study on the PI-CAI dataset. Removing each of our three core components (MDFD, MAGCS, HFDAN) validates their respective contributions.

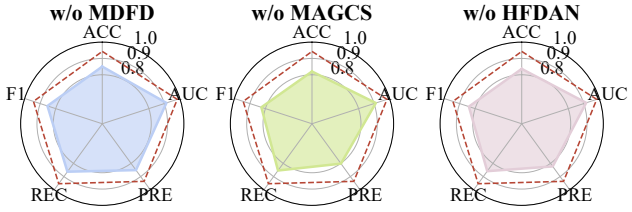

Figure 6: Radar chart visualization of the PI-CAI ablation study. The reduced area upon removing any single component visually confirms that all modules are essential.

Ablation Studies

We conducted ablation studies on the PI-CAI and CHD datasets by systematically removing each core component: the Multi-Dimensional Feature Discriminator (MDFD), the Multi-Activation Graph Construction Strategy (MAGCS), and the Hierarchical Feature Dynamic Association Network (HFDAN). The results (Table 2, Table 3, Fig. 6) confirm that all modules are integral, as removing any component causes a notable performance degradation across both datasets. Interestingly, component importance varies by data modality. On the PI-CAI (mpMRI) dataset, removing MAGCS incurred the most significant accuracy drop (12.3%), suggesting that capturing diverse inter-feature relationships is paramount. In contrast, on the CHD (heterogeneous CT and clinical) dataset, removing MDFD was most detrimental (6.9% drop), highlighting the need for salient feature selection. Furthermore, while some ablated models on the CHD dataset excel in specific metrics (e.g., AUC or PRE), our full model secures the best overall ACC and F1-Score (Table 3). This demonstrates that the synergistic interplay of our components, rather than any single module, is key to achieving robust, adaptable performance.

Qualitative Analysis

To illustrate how our Multi-Dimensional Feature Discriminator (MDFD) transforms the feature space, we visualize its impact via t-SNE (Fig. 8). The original features (Fig. 8(a)) show highly overlapping csPCa and benign classes, indicating poor separability. In stark contrast, the representations processed by the MDFD (Fig. 8(b)) exhibit clear class separation. This demonstrates that the MDFD successfully projects entangled raw features into a highly discriminative embedding space. Furthermore, the heatmap of this learned space (Fig. 7) reveals diverse activation patterns across samples, suggesting the capture of rich, multifaceted abstract

Method	ACC	AUC	PRE	REC	F1
w/o MDFD	0.8333	0.8672	0.7722	0.8839	0.8652
w/o MAGCS	0.8673	**0.9538**	0.8852	0.8710	0.8780
w/o HFDAN	0.8584	0.9306	**0.9245**	0.8033	0.8596
MAPI-GNN	**0.9027**	0.9206	0.8806	**0.9516**	**0.9147**

Table 3: Ablation study on the CHD dataset.

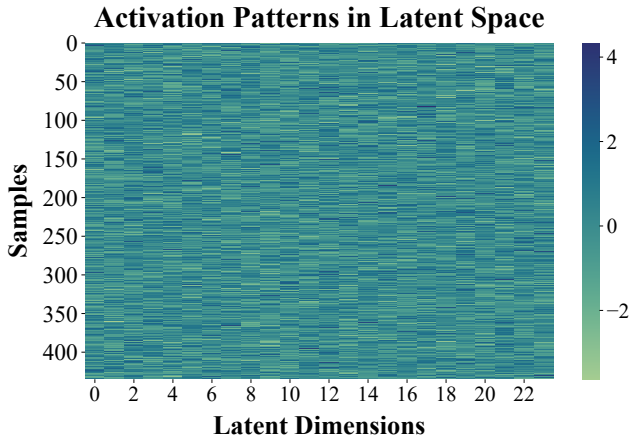

Figure 7: Heatmap of the MDFD latent space. It confirms the discriminator learns diverse, disentangled representations, as different semantic dimensions (columns) show distinct activation patterns across the patient samples (rows).

features. Together, these visualizations confirm the module's fundamental role in enhancing feature discriminability.

Comparison with PI-CAI Challenge Leaders

To situate our model's performance in a competitive context, we benchmark MAPI-GNN against the results of top-performing teams from the PI-CAI 2022 Challenge, including the SWANGEESE Team (Kan et al. 2022), PIMed Team (Li et al. 2022), and Guerbet Research (Debs et al. 2022). As the official submission channel is closed and test labels are private, a direct leaderboard comparison is impossible. We strictly followed the official protocol to provide the most rigorous benchmark possible, training on the public data and reporting performance on our cross-validation test fold using the official metrics (AUC, AP, and SCORE = (AUC + AP) / 2). As shown in Table 4 and Fig. 9, our architecture significantly surpasses these leading solutions, achieving a state-of-the-art SCORE of 0.9599 (AUC 0.9838, AP 0.9361). While this comparison is across different test sets, the substantial performance margin strongly suggests our method's architectural superiority and clinical potential.

Parameter Analysis

We analyze the sensitivity of our work to several key hyperparameters on the PI-CAI dataset to validate our final model configuration. Results are visualized in Fig. 10.

Number of Semantic Dimensions (M). This parameter defines the number of activation graphs. We analyzed its impact on the PI-CAI dataset: accuracy rose from 0.9136 (M=8) to 0.9302 (M=12) and plateaued around 0.9400 (M=16, 20), peaking at 0.9432 for M=24. We selected M=24 as the optimal trade-off between marginal performance gain and computational cost.

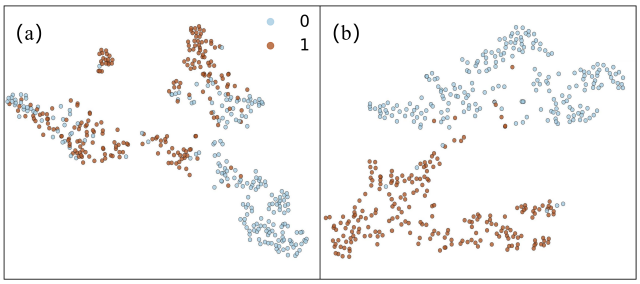

Figure 8: t-SNE visualization. (a) The original feature space, showing highly entangled classes. (b) After MDFD projection, the learned representations show clear separability, validating the creation of a discriminative embedding space.

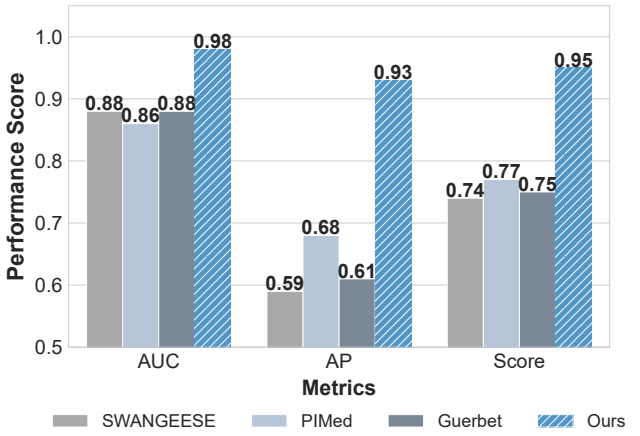

Figure 9: Visual benchmark vs. top PI-CAI 2022 Challenge teams. Our model's scores, obtained on our cross-validation test fold, are compared against the leaderboard scores.

Graph Construction Parameters (k and w). For the number of neighbors (k), testing $k \in \{3, 5, 10\}$ yielded accuracies of 0.9091, 0.9432, and 0.9205, respectively. We confirmed $k = 5$ as the optimal trade-off between performance and cost. For edge weighting (Eq. 3), we chose the average form. This links weights to discriminator scores (C_m) to guide GAT's attention without distorting individual scores.

Proportion of Activated Features (PAF). This value controls the sparsity of the constructed graphs, creating a critical trade-off between retaining sufficient signal information (higher PAF) and filtering out potential noise (lower PAF). The results in Table 5 indicate that selecting 5% (PAF=0.05) of features provides the optimal balance.

Feature Perturbation Method (FPM). We also validate our choice of the zeroing-out perturbation strategy for calcu-

Team	AUC	AP	SCORE
SWANGEESE Team	0.8860	0.5930	0.7400
PIMed Team	0.8650	0.6810	0.7730
Guerbet Research	0.8890	0.6150	0.7520
Ours	**0.9838**	**0.9361**	**0.9599**

Table 4: Benchmark against top teams from the PI-CAI 2022 Challenge leaderboard (SCORE = (AUC + AP) / 2).

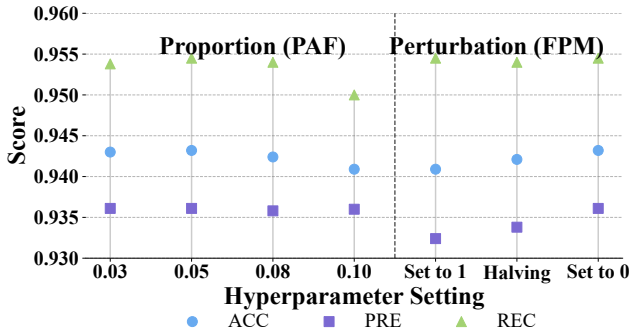

Figure 10: Hyperparameter sensitivity analysis on the PI-CAI dataset. PAF: Proportion of Activated Features, FPM: Feature Perturbation Method.

lating the crucial feature influence scores. As detailed in Table 6, this method proves most effective for accurately identifying critical diagnostic features compared to other strategies (e.g., halving or setting to one).

Computational Complexity

To assess clinical feasibility, we analyzed MAPI-GNN's computational footprint on a single NVIDIA A30 GPU (165W TDP). Our model is lightweight, with 12.27M parameters and a total of 1.925 GFLOPs for the 440-case PI-CAI dataset, averaging 4.38 MFLOPs per case. With an average inference time of 45 ms per case, the model enables near real-time diagnostics. This high accuracy, fast inference, and MFLOPs-scale footprint confirm our framework as a practical and efficient solution for clinical deployment.

Discussion

On the Synergy of CNNs and GNNs. Our experiments affirm a key synergy. While GNNs alone outperform CNNs in relational modeling, our model demonstrates that using CNNs for potent feature extraction and GNNs to model their inter-dependencies yields superior performance to either architecture in isolation.

Clinical Applicability. The architecture shows strong clinical relevance. Its high specificity (0.931) and recall (0.954) promise to reduce unnecessary biopsies while maintaining high detection rates. Furthermore, its efficient, end-to-end design enables streamlined inference, enhancing deployment feasibility over complex multi-stage methods.

Limitations and Future Work. Despite its strong performance, we identify key avenues for future work. First, the current framework assumes complete modalities; enhancing its robustness to missing data is a critical next step for real-world utility. Second, while validated on two tasks, extending the framework to more diseases and data types (e.g., PET, histopathology, genomics) is needed to establish broader generalizability. While our analysis shows the learned semantic space is discriminative, developing methods, such as those linking to traditional radiomics, to map these abstract dimensions to concrete pathological concepts would be valuable towards greater clinical interpretability.

Proportion (PAF)	ACC	PRE	REC
0.03	0.9430	0.9361	0.9538
0.10	0.9409	0.9360	0.9500
0.08	0.9424	0.9358	0.9540
0.05	**0.9432**	**0.9361**	**0.9545**

Table 5: Sensitivity to the Proportion of Activated Features (PAF). This value controls graph sparsity, and an optimal trade-off is achieved at 5%.

Perturbation Method	ACC	PRE	REC
Set to 1	0.9409	0.9324	0.9545
Halving	0.9421	0.9338	0.9540
Zeroing-out	**0.9432**	**0.9361**	**0.9545**

Table 6: Comparison of different feature perturbation methods (FPM). The zeroing-out strategy used in our model yields the best performance.

Conclusion

In this work, we introduce Multi-Activation Plane Interaction Graph Neural Network, a framework for multimodal medical diagnosis that, instead of relying on the prevailing static single-graph paradigm, learns patient-specific graph topologies directly from the data. At the core of our method is a two-stage process that first dynamically constructs a multifaceted graph profile for each patient and then performs a hierarchical fusion across these graphs to model both intra- and inter-sample relationships. Extensive experiments on diverse medical datasets demonstrate that our proposed framework achieves state-of-the-art performance. Ablation studies further confirm that its synergistic components are all integral to its success. By providing a more adaptive, powerful, and data-driven approach to multimodal fusion, our work represents a significant step toward more accurate, reliable, and ultimately more trustworthy computer-aided diagnosis in complex clinical scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62401481, Natural Science Foundation of Sichuan Province under Grant 2025ZNSFSC1450, Fundamental Research Funds for the Central Universities under Grant 2682024CX067, China Postdoctoral Science Foundation under Grant 2024M752683.

References

Alfeo, A. L.; Cimino, M. G.; and Vaglini, G. 2022. Degradation stage classification via interpretable feature learning. *Journal of Manufacturing Systems*, 62: 972–983.

Azizi, S.; Mustafa, B.; Ryan, F.; Beaver, Z.; Freyberg, J.; Deaton, J.; Loh, A.; Karthikesalingam, A.; Kornblith, S.; and Chen, T. 2021. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3478–3488.

Bayoudh, K.; Knani, R.; Hamdaoui, F.; and Mtibaa, A. 2022. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8): 2939–2970.

Behrad, F.; and Abadeh, M. S. 2022. An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications*, 200: 117006.

Boehm, K. M.; Khosravi, P.; Vanguri, R.; Gao, J.; and Shah, S. P. 2022. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, 22(2): 114–126.

Chen, H.; Huang, Z.; Xu, Y.; Deng, Z.; Huang, F.; He, P.; and Li, Z. 2022. Neighbor Enhanced Graph Convolutional Networks for Node Classification and Recommendation. arXiv preprint / online article. Accessed via online repository.

Cui, C.-X.; Yang, H.; Wang, Y.; Zhao, S.; Asad, Z.; Coburn, L. A.; Huo, Y.; et al. 2023. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 5(2): 022001.

Debs, N.; Routier, A.; Abi-Nader, C.; et al. 2022. Deep learning for detection and diagnosis of prostate cancer from bpMRI and PSA: Guerbet's contribution to the PI-CAI 2022 Grand Challenge. Guerbet Research, Villepinte, France. noelie.debs@guerbet.com.

Ding, L.; Li, C.; Jin, D.; and Ding, S. 2024. Survey of spectral clustering based on graph theory. *Pattern Recognition*, 110366.

Duan, J.; Xiong, J.; Li, Y.; and Ding, W. 2024. Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion*, 102536.

D'Souza, N. S.; Wang, H.; Giovannini, A.; Foncubierta-Rodriguez, A.; Beck, K. L.; Boyko, O.; and Syeda-Mahmood, T. F. 2024. Fusing modalities by multiplexed graph neural networks for outcome prediction from medical data and beyond. *Medical Image Analysis*, 93: 103064.

Fan, D.-P.; Zhou, T.; Ji, G.-P.; Zhou, Y.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE transactions on medical imaging*, 39(8): 2626–2637.

Fei, Z.; Guo, J.; Gong, H.; Ye, L.; Attahi, E.; and Huang, B. 2023. A gnn architecture with local and global-attention feature for image classification. *IEEE Access*, 11: 110221–110233.

Feng, Y.; Ji, S.; Liu, Y.-S.; Du, S.; Dai, Q.; and Gao, Y. 2023. Hypergraph-based multi-modal representation for open-set 3D object retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 2206–2223.

Gao, X.; Shi, F.; Shen, D.; and Liu, M. 2021. Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in Alzheimer's disease. *IEEE journal of biomedical and health informatics*, 26(1): 36–43.

Haarburger, C.; Baumgartner, M.; Truhn, D.; Broeckmann, M.; Schneider, H.; Schrading, S.; Kuhl, C.; and Merhof, D. 2019. Multi-scale Curriculum CNN for Context-aware Breast MRI Malignancy Classification. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*, volume 11767 of *Lecture Notes in Computer Science*, 495–503. Springer. ISBN 978-3-030-32250-2.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30, 1024–1034. Curran Associates, Inc.

He, A.; Li, T.; Li, N.; Wang, K.; and Fu, H. 2020. CABNet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 40(1): 143–153.

Holste, G.; Partridge, S. C.; Rahbar, H.; Biswas, D.; Lee, C. I.; and Alessio, A. M. 2019. End-to-End Learning of Fused Image and Non-Image Features for Improved Breast Cancer Classification from MRI. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3294–3303.

Huang, Z.; Li, K.; Jiang, Y.; Jia, Z.; Lv, L.; and Ma, Y. 2024. Graph Relearn Network: Reducing performance variance and improving prediction accuracy of graph neural networks. *Knowledge-Based Systems*, 301(000): 15.

Huang, Z.; Tang, Y.; and Chen, Y. 2022. A graph neural network-based node classification model on class-imbalanced graph data. *Knowledge-Based Systems*, 244: 108538–.

Kan, H.; Anhui, H.; Qiao, L.; Shi, J.; and An, H. 2022. Implementation Method of the PI-CAI Challenge (SWANGEESE Team). PI-CAI 2022 Grand Challenge. Challenge Report.

Karim, S.; Tong, G.; Li, J.; Qadir, A.; Farooq, U.; and Yu, Y. 2023. Current advances and future perspectives of image fusion: A comprehensive review. *Information Fusion*, 90: 185–217.

Khemani, B.; Patil, S.; Kotecha, K.; and Tanwar, S. 2024. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*, 11(1): 18.

Khoshraftar, S.; and Aijun, A. N. 2024. A Survey on Graph Representation Learning Methods. ACM Digital Library / online article. Accessed via online repository.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kuttala, R.; Subramanian, R.; and Oruganti, V. R. M. 2023. Multimodal hierarchical CNN feature fusion for stress detection. *IEEE Access*, 11: 6867–6878.

Lecun, Y.; and Bottou, L. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, B.; and Nabavi, S. 2024. A multimodal graph neural network framework for cancer molecular subtype classification. *BMC bioinformatics*, 25(1): 27.

Li, J.; Zhao, Y.; Zhang, H.; LiMember, W. J.; Fu, C.; Lian, C.; and Shan, P. 2024. Image Encoding and Fusion of Multimodal Data Enhance Depression Diagnosis in Parkinson's Disease Patients. *IEEE Transactions on Affective Computing*.

Li, X.; Vesal, S.; Saunders, S.; John, S.; Soerensen, C.; Jahanandish, H.; Moroianu, S.; Bhattacharya, I.; Fan, R. E.; Sonn, G. A.; et al. 2022. The Prostate Imaging: Cancer AI (PI-CAI) 2022 Grand Challenge (PIMed Team). Departments of Radiology and Urology, Stanford University, Stanford, CA 94305, USA; Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA. PI-CAI Grand Challenge Report.

Lin, Y.; Lu, K.; Yu, S.; Cai, T.; and Zitnik, M. 2023. Multimodal learning on graphs for disease relation extraction. *Journal of Biomedical Informatics*, 143: 104415.

Lipkova, J.; Chen, R. J.; Chen, B.; Lu, M. Y.; Barbieri, M.; Shao, D.; Vaidya, A. J.; Chen, C.; Zhuang, L.; and Williamson, D. F. 2022. Artificial intelligence for multimodal data integration in oncology. *Cancer cell*, 40(10): 1095–1110.

Mi, J.-X.; Li, A.-D.; and Zhou, L.-F. 2020. Review study of interpretation methods for future interpretable machine learning. *IEEE Access*, 8: 191969–191985.

Minaee, S.; Kafieh, R.; Sonka, M.; Yazdani, S.; and Soufi, G. J. 2020. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. *Medical image analysis*, 65: 101794.

Panayides, A. S.; Amini, A.; Filipovic, N. D.; Sharma, A.; Tsaftaris, S. A.; Young, A.; Foran, D.; Do, N.; Golemati, S.; and Kurc, T. 2020. AI in medical imaging informatics: current challenges and future directions. *IEEE journal of biomedical and health informatics*, 24(7): 1837–1857.

Sarkar, D.; Roy, S.; Malakar, S.; and Sarkar, R. 2023. A modified GNN architecture with enhanced aggregator and message passing functions. *Engineering Applications of Artificial Intelligence*, 122: 106077.

Shao, W.; Peng, Y.; Zu, C.; Wang, M.; Zhang, D.; and Initiative, A. D. N. 2020. Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, 80: 101663.

Sharma, A.; Sharma, K.; and Kumar, A. 2023. Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion. *Neural computing and applications*, 35(31): 22935–22948.

Sim, J.; Lee, M.; Wu, G.; and Kim, W. H. 2024. Multi-modal Graph Neural Network with Transformer-Guided Adaptive Diffusion for Preclinical Alzheimer Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 511–521. Springer.

Song, X.; Li, J.; and Qian, X. 2022. Diagnosis of glioblastoma multiforme progression via interpretable structure-constrained graph neural networks. *IEEE Transactions on Medical Imaging*, 42(2): 380–390.

Stahlschmidt, S. R.; Ulfenborg, B.; and Synnergren, J. 2022. Multimodal deep learning for biomedical data fusion: a review. *Briefings in bioinformatics*, 23(2): bbab569.

Tan, K.; Huang, W.; Liu, X.; Hu, J.; and Dong, S. 2022. A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. *Artificial Intelligence in Medicine*, 126: 102260.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wang, H.; Wang, S.; Qin, Z.; Zhang, Y.; Li, R.; and Xia, Y. 2021. Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Medical Image Analysis*, 67: 101846.

Wang, Z.; Li, M.; Wang, H.; Jiang, H.; Yao, Y.; Zhang, H.; and Xin, J. 2019. Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features. *IEEE Access*, 7: 105146–105158.

Wu, S.; Sun, F.; Zhang, W.; Xie, X.; and Cui, B. 2022. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5): 1–37.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24.

Xie, F.; Yuan, H.; Ning, Y.; Ong, M. E. H.; Feng, M.; Hsu, W.; Chakraborty, B.; and Liu, N. 2022. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of biomedical informatics*, 126: 103980.

Xie, X.; Wang, X.; Liang, Y.; Yang, J.; Wu, Y.; Li, L.; Sun, X.; Bing, P.; He, B.; and Tian, G. 2021. Evaluating cancer-related biomarkers based on pathological images: a systematic review. *Frontiers in oncology*, 11: 763527.

Xu, X.; Li, J.; Zhu, Z.; Zhao, L.; Wang, H.; Song, C.; Chen, Y.; Zhao, Q.; Yang, J.; and Pei, Y. 2024. A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis. *Bioengineering*, 11(3): 219.

Zhang, H.; Song, R.; Wang, L.; Zhang, L.; Wang, D.; Wang, C.; and Zhang, W. 2022. Classification of brain disorders in rs-fMRI via local-to-global graph neural networks. *IEEE transactions on medical imaging*, 42(2): 444–455.

Zhang, Y.; He, X.; Liu, Y.; Ong, C. Z. L.; Liu, Y.; and Teng, Q. 2023. An end-to-end multimodal 3D CNN framework with multi-level features for the prediction of mild cognitive impairment. *Knowledge-Based Systems*, 281: 111064.

Zhao, Y.; Li, X.; Zhou, C.; Pen, H.; Zheng, Z.; Chen, J.; and Ding, W. 2024. A review of cancer data fusion methods based on deep learning. *Information Fusion*, 102361.

Zhou, F.; Hu, S.; Du, X.; and Lu, Z. 2024. Motico: an attentional mechanism network model for smart aging disease risk prediction based on image data classification. *Computers in biology and medicine*, 178: 108763.