Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 4, No. 2, 2025

A Review of Multimodal Medical Image Classification Based on Deep Learning

Jiahao Song*

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

* **Corresponding author**: Jiahao Song

Abstract: Medical imaging plays an important role in the field of modern medicine. It provides key information about the internal structure and biological activities of the human body for clinical diagnosis and treatment. However, single-modality medical imaging is limited by the imaging principle and is difficult to fully present the characteristics of specific organs or lesions, which restricts the accuracy and comprehensiveness of clinical diagnosis. Multimodal medical image fusion technology can more comprehensively and accurately reflect the characteristics of lesions by integrating the complementary information of different imaging modalities. In recent years, it has become a research hotspot in the field of medical image analysis. In this paper, model-based and model-independent multimodal fusion methods are first introduced, and then the most popular neural network model and its application in multimodal medical images are elaborated in detail. Finally, the future development trend of multimodal medical image classification is prospected.

Keywords: Multimodal fusion; Neural networks; Medical image classification.

1. Introduction

Medical imaging plays an important role in the field of modern medicine. It provides key information about the internal structure and biological activities of the human body for clinical diagnosis and treatment. However, due to the technical characteristics of different imaging devices, single-modality images are often difficult to fully present the characteristics of specific organs or lesions, which to a certain extent restricts the accuracy and comprehensiveness of clinical diagnosis. Therefore, in order to make full use of the rich and heterogeneous features presented by different modal medical images and improve the accuracy of image interpretation, the use of multimodal medical images for disease classification has become a research field that has attracted much attention [1, 2].

In the field of medical diagnosis and treatment, medical imaging technology plays an indispensable role and provides key characteristic information for clinical decision-making. Different medical imaging modalities have different sensitivities and resolution characteristics for human tissue structure and pathological changes based on their unique imaging principles [3,4]. Taking computed tomography (CT) as an example, this technology obtains cross-sectional images through X-ray rotation scanning. Its advantage is that it can provide high-resolution anatomical structure information, which is of great value in the detection and diagnosis of craniocerebral injuries, tumor lesions and cardiovascular diseases. In contrast, magnetic resonance imaging (MRI) uses strong magnetic fields and radiofrequency pulses for imaging. Its outstanding soft tissue contrast resolution gives it a unique advantage in the detection of neurological diseases (such as brain and spinal lesions), motor system injuries (such as joint lesions) and tumors. In addition, nuclear medicine imaging technology achieves disease diagnosis and evaluation by introducing radioactive tracers. Among them, positron emission tomography (PET), as a representative technology in this field, plays an important role in tumor metabolic imaging and neurological function evaluation. Traditional medical diagnosis mainly relies on single-modality imaging examinations. However, this mode has obvious limitations. Single-modality images are often limited by the imaging principle and may have problems such as noise interference, insufficient contrast, and limited resolution. These factors will affect the accuracy of diagnosis. Multimodal image fusion technology can help to make up for these shortcomings and improve image quality and the accuracy and reliability of diagnosis [5].

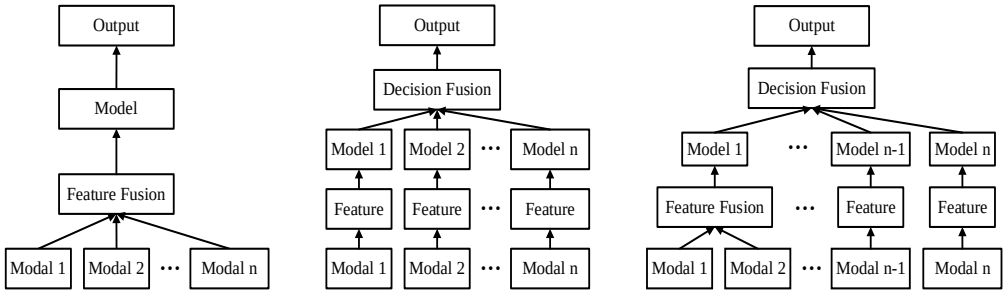
In this article, we first introduce the multimodal fusion method. Then we discuss the most popular neural network method in detail. Finally, we analyze the challenges of using deep learning to assist multimodal image diagnosis, and look forward to future research methods.

2. Multimodal Fusion Method

To date, a large number of multimodal fusion methods have been proposed, which can be primarily categorized into model-agnostic and model-based approaches.

2.1. Model-agnostic approach

The model-agnostic method does not rely on a specific deep learning method in the multi-modal fusion process. According to feature fusion, this method can be divided into: early fusion, late fusion and hybrid fusion. The structures of the three fusion methods are shown in Figure 1.



(a) Early Fusion (b)Late Fusion (c) Hybird Fusion

Fig.1 Three model-independent multimodal fusion methods

(1) Early Fusion: In order to solve the heterogeneity problem of multimodal medical imaging data in the original space, this strategy first extracts independent features from each modality data, and then fuses them in the feature space, so it is also called feature-level fusion. This fusion method makes full use of the complementarity and correlation of the underlying features between different modalities. Through the interactive integration of feature space, it can achieve effective fusion of multi-source information in the early processing stage, providing richer feature representation for subsequent analysis. Specifically, early fusion achieves preliminary integration of multimodal information by extracting basic features such as edges and textures of each modality and establishing a mapping relationship between features. This processing method helps to retain more original data information and provide a reliable feature basis for subsequent diagnostic decisions. A commonly used fusion rule is addition, which is a simple addition operation of all features corresponding to all modalities. The formula is as follows:

$$z = f(w_1^T v_1 + w_1^T v_1 + \cdots + w_n^T v_n) \quad (1)$$

where z is the fused multimodal feature, w is the weight matrix, and v_n is the feature input of modality n. This method is simple to operate, but its disadvantage is that it is easy to cause semantic loss in the later stage.

Another fusion rule is the multiplication rule, which uses tensor calculation to fuse the eigenvectors of all modes into a unified tensor. The formula is as follows:

$$z = \begin{bmatrix} v^1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} v^2 \\ 1 \end{bmatrix} \otimes \cdots \begin{bmatrix} v^n \\ 1 \end{bmatrix} \quad (2)$$

where v^n represents the feature input of different modes, and $\otimes$ represents the outer product operator.

(2) Late Fusion: This method integrates information at the decision-making level, so it is also called decision-level fusion. The core idea is to build an independent deep learning model for each modality for training, and then integrate the output results of each sub-model through specific fusion rules to obtain the final diagnostic decision. The advantage of this method lies in its high flexibility: on the one hand, different modalities can adopt the network architecture and training strategy that best suits their data characteristics; on the other hand, each model maintains relative independence to avoid mutual interference between modalities. In terms of fusion strategy, late fusion mainly adopts rule-based integration methods, including but not limited to maximum fusion, average fusion, probability-based Bayesian fusion, more complex ensemble learning and other methods. This hierarchical processing strategy not only reduces the complexity of model training, but also can effectively integrate the discriminative information of each modality and improve the reliability of the final decision.

(3) Hybrid Fusion: As an integrated fusion strategy, this method organically combines the technical advantages of early fusion and late fusion to form a multi-level feature fusion system. Specifically, hybrid fusion first realizes the preliminary integration of multimodal data at the feature level (early fusion), and then performs deep information fusion at the decision level (late fusion). This dual fusion mechanism can fully explore the complementary information of multimodal data at different levels. It is worth noting that in this multi-level fusion framework, the strategy combination and weight distribution at different fusion stages are the key factors that determine the performance of the model. They need to be carefully designed and optimized according to specific task requirements and data characteristics to achieve the best synergy of each fusion level.

Through in-depth analysis of the three fusion methods, the following conclusions can be drawn: Each fusion strategy has its own unique advantages and limitations. Although early fusion performs well in capturing feature correlation, it is easy to cause the model to overfit the training data and reduce the generalization ability due to premature feature integration. Late fusion can effectively integrate the high-level semantic information of each modality, but due to its independent training characteristics, it cannot achieve the collaborative optimization of multimodal data at the classifier level. Although hybrid fusion combines the advantages of the first two methods and provides greater design flexibility, its complex network structure and multi-level fusion mechanism significantly increase the computational burden and training difficulty of the model. Therefore, in practical applications, there is no universal optimal fusion solution, but it is necessary to weigh the advantages and disadvantages of each method according to specific clinical needs, data characteristics, and computing resources, and select the most suitable fusion strategy. This task-oriented adaptive selection is the key to realizing intelligent diagnosis of medical images.

2.2. Model-based approach

The model-based approach solves the multimodal fusion problem from the perspective of implementation technology and models. Common methods include Multiple Kernel Learning (MKL), Graphical Model (GM) and Neural Networks (NNs) methods.

(1) MKL is an extension of the kernel support vector machine (SVM) method. This method regards the kernel as a similarity function between data points and uses different kernels to correspond to different perspectives of the data, so it can better and more flexibly fuse heterogeneous data. The advantage of this method lies in the flexibility of kernel selection. Through the feature mapping capabilities of different basic kernels, different feature components of heterogeneous data can be solved through corresponding kernel functions.

(2) GM mainly uses technical means such as image segmentation, feature splicing and relationship prediction to achieve the fusion processing of shallow or deep image features, and finally generates a comprehensive modality fusion result. The significant advantage of this type of method is that it can make full use of the spatial structure information and time series characteristics of medical imaging data, and at the same time supports embedding domain expert knowledge into the model in the form of a graph structure, which not only enhances the interpretability of the model, but also improves the rationality of feature representation. However, this modeling method that strongly relies on prior knowledge limits the generalization ability of the model to a certain extent, and may cause the model to perform poorly when facing out-of-distribution data. However, this modeling method that strongly relies on prior knowledge limits the generalization ability of the model to a certain extent, and may cause the model to perform poorly when facing out-of-distribution data.

(3) As one of the most mainstream fusion technologies, NNs have the core advantage of powerful feature learning capabilities and end-to-end training mechanisms. This type of method can automatically learn the complex nonlinear relationships between multimodal data, and achieve effective fusion of different modal features through a deep network architecture, with good scalability and adaptability. Compared with traditional non-neural network methods, deep neural networks can learn more complex decision boundaries, thereby better capturing the potential correlations between multimodal data. However, this type of method also has obvious limitations: first, its "black box" characteristics lead to poor interpretability of the model, which is an important challenge in fields that require high credibility, such as medical diagnosis; second, deep neural networks usually require large-scale annotated data sets for training to achieve ideal performance, which may limit its scope of application when medical imaging data is relatively difficult to obtain; finally, this type of method has high computing resource requirements and may face challenges in scenarios with high real-time requirements.

3. Neural network models and applications

3.1. Convolutional Neural Networks

3.1.1. Brief Introduction of Convolutional Neural Network

Convolutional neural networks (CNNs) are a type of deep learning architecture specifically designed to process grid-like data structures (such as images) and play an important role in the fields of computer vision and pattern recognition [6,7]. By performing multi-level feature extraction and nonlinear transformation, it can automatically learn meaningful feature representations from raw data [8]. The main structure of CNNs is convolutional layer [9], activation layer, pooling layer and fully connected layer. The function of the convolutional layer is feature extraction. Based on local correlation, it focuses on the influence of pixels around each pixel. Global feature extraction is achieved by continuously sliding the convolution kernel. Global feature extraction is achieved by continuously sliding the convolution kernel. The activation layer performs nonlinear mapping on the data processed by the convolutional layer. Common activation functions include Sigmoid, ReLU [10], GELU [11], etc. The pooling layer is also called the downsampling layer [12]. It retains important features and prevents overfitting of the data. The neurons in each layer of the fully connected layer are connected to all neurons in the previous layer and usually play the role of classification. As shown in Figure 2, the CNN model with brain tumor images as input includes convolutional layer, pooling layer and fully connected layer.

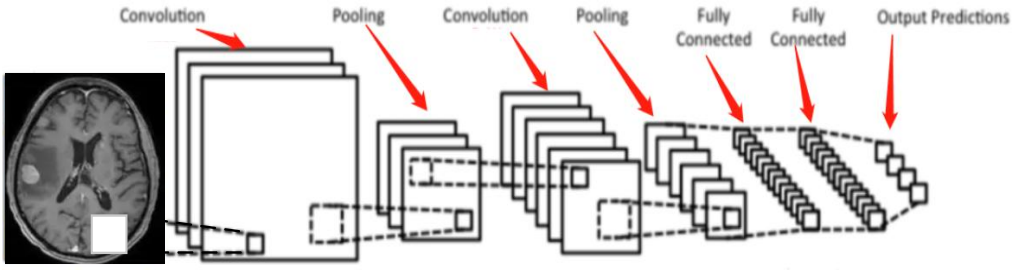


Fig.2 Classic CNN framework diagram

3.1.2. Convolutional Neural Networks for Multimodal Medical Image Diagnosis

Multimodal medical images contain patient case information from different observation angles. The fusion of multimodal images helps to make full use of the complementary information provided by different modalities, which plays an important role in disease diagnosis [13,14]. With the continuous research of CNNs, the network has shown excellent performance in multimodal medical image diagnosis tasks.

In brain disease detection, Kong et al. [15] combined MRI and PET images into a composite fusion mode, which not only provides information on the brain's anatomical structure and metabolic function, but also effectively highlights key features through image registration and noise reduction. To further improve feature extraction capabilities, Ismail et al. [16] proposed an integrated learning architecture based on multimodal image fusion, UltiAz-Net. This architecture cleverly fuses the output features of three classic networks, AlexNet [17], Inception-V3 [18], and ResNet [19], and introduces the MOGOA nature-inspired algorithm to achieve automatic optimization of the network structure, significantly improving the accuracy of brain image classification. However, multimodal fusion also brings about the problem of increased computational complexity. Although the multi-stream deep CNN architecture proposed by Ge et al. [20] effectively alleviates the overfitting phenomenon through multi-sensor feature fusion and 2D image enhancement technology, its complex network structure may lead to reduced computational efficiency. Considering that different modalities have different dimensional characteristics [21], Tu

et al. [22] developed an innovative multimodal feature conversion and fusion diagnosis model. The model first uses a feature filtering algorithm based on the degree of influence to remove redundant features, and then designs a hybrid framework combining artificial neural networks (ANN) and CNN to successfully achieve accurate diagnosis of Alzheimer's disease. This hierarchical feature processing strategy not only reduces the computational complexity, but also improves the generalization ability of the diagnostic model.

In the field of ocular disease diagnosis, researchers have significantly improved the diagnostic effect through innovative multimodal fusion methods. Li et al. [23] proposed a hierarchical fusion strategy that systematically integrates the multidimensional features of the network by deeply mining the complementary information between modalities, and achieved the best performance in the classification task of diabetic retinopathy dataset. To further improve the generalization ability of the model and reduce the dependence on training data, Yoo et al. [24] innovatively designed a two-stage multimodal transfer learning framework: the first stage uses a pre-trained CNN to extract deep features from fundus images, and the second stage effectively fuses these features with standardized clinical data, and finally achieves accurate detection of retinopathy by training the classifier. In the field of glaucoma detection, Huang et al. [25] developed a probabilistic deep learning framework that not only extracts features through CNN, but also innovatively introduces a probability layer to quantify the uncertainty of the prediction, so that the model can simultaneously output the diagnosis results and their corresponding confidence scores, providing more reliable support for clinical decision-making.

CNNs also shows excellent performance in other disease detection. Qian et al. [26] combined ultrasound B-mode images with color Doppler data for accurate classification of breast masses. Specifically, they used CNN to extract spatial patterns and texture information related to mass features from ultrasound B-mode images, and captured hemodynamic and vascular distribution features from color Doppler data, and finally achieved classification through feature fusion. To further improve the multimodal feature fusion effect, Le et al. [27] designed a new similarity loss function, combined it with the traditional classification loss function, and integrated it into the back-propagation training process of CNN, which significantly improved the diagnostic accuracy of prostate diseases. Ge et al. [28] proposed an innovative deep convolutional neural network (DCNN) architecture and developed a significant feature descriptor. The model can simultaneously process clinical images and dermoscopic images of a single lesion, effectively learn single-modal features and cross-modal representations, and thus achieve accurate classification of skin cancer. In response to the heterogeneity challenge of stem cell cancer, Li et al. [29] introduced a discriminative feature learning mechanism: first, by designing a specific loss function to narrow the feature distance of tumors in the same category, while expanding the feature differences between tumors in different categories; second, an adaptive weighting strategy was proposed to dynamically adjust the contribution of each modality, increase the weight of low loss value modalities, and reduce the influence of high loss value modalities, thereby optimizing the overall performance of the model.

Multimodal neuroimaging data, due to its complementary information features, provides an important foundation for the development of advanced network architectures. Studies have shown that the effective fusion of neuroimaging information of different modalities can not only build a more complex network system, but also significantly improve the diagnostic performance of the model. In particular, in the application of transfer learning, pre-trained models, as network initialization or feature extractors, have been proven to significantly accelerate the model convergence process while improving the learning efficiency and performance of network terminal tasks [30,31]. However, as the network depth increases, the improvement of model accuracy is often accompanied by an exponential increase in computational complexity. This trade-off requires researchers to carefully consider when designing the model: on the one hand, it is necessary to ensure that the network has sufficient expressive power to capture complex pathological features, and on the other hand, it is necessary to control the model complexity to avoid overfitting and waste of computational resources. Therefore, how to find the optimal balance between reducing model complexity and maintaining diagnostic accuracy has become an important research direction in the current field of medical image analysis.

Table 1 summarizes some representative works of CNNs on multimodal medical images.

Tab.1 CNNs on Multimodal Medical Images

Autor	Model	Disease
Kong et al.	3D CNN	Alzheimer’s disease
Ismail et al.	UltiAz-Net	Alzheimer’s disease
Ge et al.	CNN	Glioma
Tu et al.	ANN + CNN	Alzheimer’s disease
Li et al.	CNN	Glaucoma
Yoo et al.	DeepPDT-Net	Glaucoma
Huang et al.	CNN	Glaucoma
Qian et al.	CNN	Breast masses
Le et al.	CNN	Prostate
Ge et al.	DCNN	Skin disease
Li et al.	AGDAF	Hepatocellular carcinoma

3.2. Vision Transformer

3.2.1. Brief Introduction of Vision Transformer

Transformer [32] is a revolutionary neural network architecture. Its core innovation lies in the use of the self-attention mechanism. It was originally designed to solve sequence-to-sequence (seq2seq) learning tasks in the field of natural language processing (NLP) [33]. Compared with the traditional recurrent neural network (RNN) and its improved version, the long short-term memory network (LSTM), the Transformer model shows significant advantages in capturing long-distance dependencies, mainly due to its global attention mechanism. At the same time, since it abandons the recursive structure and is completely based on matrix operations, it has also greatly improved the efficiency of parallel computing. However, the original Transformer architecture is mainly designed for sequence data processing and cannot be directly applied to computer vision tasks. This limitation was overcome by the Google research team in 2020. They proposed the Vision Transformer (ViT) [34], which successfully extended the Transformer architecture to computer vision tasks such as image classification by segmenting the image into a fixed-size patch sequence and

introducing position encoding.

3.2.2. Vision Transformer for Multimodal Medical Image Diagnosis

ViT abandons the hierarchical convolutional structure of traditional CNN and adopts an architecture based on the self-attention mechanism, which successfully achieves effective modeling of the global context information of the image. This innovative design has shown great potential in the field of medical image diagnosis, and many studies have confirmed its application value.

To solve the problem of limited sample size of brain imaging data, Lyu et al. [35] innovatively adopted a transfer learning strategy to transfer the pre-trained ViT model to the brain imaging dataset. They used 2D MRI images as input and ViT as the backbone network, and achieved an accuracy of 95.3% in the Alzheimer's disease (AD) diagnosis task, which was significantly better than the traditional method. Zhu et al. [36] proposed the BraInf, which innovatively integrated representation learning, feature distillation and classification tasks into a unified framework. By introducing a multi-head self-attention mechanism, BraInf can effectively process high-dimensional MRI data, and at the same time use the structured distillation layer to perform feature downsampling, which significantly reduces the computational complexity while retaining key diagnostic features. In response to the special needs of 3D medical image analysis, Jang et al. [37] developed the M3T classification model. The model adopts an innovative two-stage feature extraction strategy: first, 3D CNN is used to extract disease-related local features from 3D MRI images, and then these features are input into ViT for multi-plane and multi-slice feature fusion, thereby achieving an overall representation of 3D MRI images. This hybrid architecture fully leverages the advantages of CNN in local feature extraction and ViT in global feature modeling, providing a new solution for 3D medical image analysis.

Although ViT has achieved remarkable success on large natural image datasets, its application in medical image analysis faces an important challenge: medical datasets usually have limited sample size, which makes it difficult to meet ViT's requirement for large-scale pre-training data. To address this problem, researchers have proposed a hybrid architecture strategy that combines ViT with convolutional neural networks (CNNs) [38]. This scheme can fully utilize the advantages of CNN in local feature extraction and ViT in global context modeling. The TransMed framework proposed by Dai et al. [39] is a typical representative in this direction. This framework innovatively integrates CNN and ViT for parotid tumor diagnosis. TransMed adopts a two-stage feature learning strategy: first, CNN is used to process multimodal medical images and convert them into feature sequences; then, ViT is used to learn the complex relationships between these feature sequences and finally achieve tumor classification. This architectural design not only retains the ability of CNN to extract low-level visual features, but also takes advantage of ViT's advantage in modeling high-level semantic relationships. Zhang et al. [40] proposed a more advanced MMIF (Multi-Modal Interaction Framework) framework, which contains two core components: category-constrained parallel ViT (CCPViT) and multimodal representation alignment network (MRAN). CCPViT learns the key features of different modalities through a parallel processing mechanism, effectively solving the problem of misaligned multimodal data. MRAN uses a cross-attention mechanism to deeply explore the interactive representations between cross-modal data by cascading encoded images and decoded texts, which not only achieves modality alignment but also significantly improves the accuracy of abnormality recognition. This dual architecture design provides a new research paradigm for multimodal medical data analysis.

Table 2 summarizes some representative works of ViT on multimodal medical images.

Tab.2 ViT on Multimodal Medical Images

Autor	Model	Disease
Lyu et al.	ViT	Alzheimer’s disease
Zhu et al.	BraInf	Alzheimer’s disease
Jang et al.	M3T	Alzheimer’s disease
Dai et al.	Trrans Med	Parotid gland tumors
Zhang et al.	MMIF	Fetal distress

3.3. Graph Neural Networks

3.3.1. Brief Introduction of Graph Neural Networks

Graph neural network (GNNs) is a deep learning model specifically designed to process data with complex relationships. Its core advantage is that it can directly use graph structures for information processing. Compared with CNNs and ViT models, GNNs can effectively process complex graph structure data by modeling irregular non-Euclidean data [41].

The architecture of the graph neural network is shown in Figure 3. The model iteratively updates the feature matrix and adjacency matrix of the input graph by stacking multiple graph convolution layers, thereby gradually extracting and optimizing the high-level feature representation of the node and its neighborhood. The core lies in the design of the graph convolution operation. Different graph convolutions define different information aggregation methods. Specifically, the graph convolution generates a new feature representation by combining the features of the node itself with the features of its neighboring nodes. This flexible information aggregation mechanism enables the graph convolutional neural network to effectively capture local and global patterns in graph structured data, thereby improving the performance of the model.

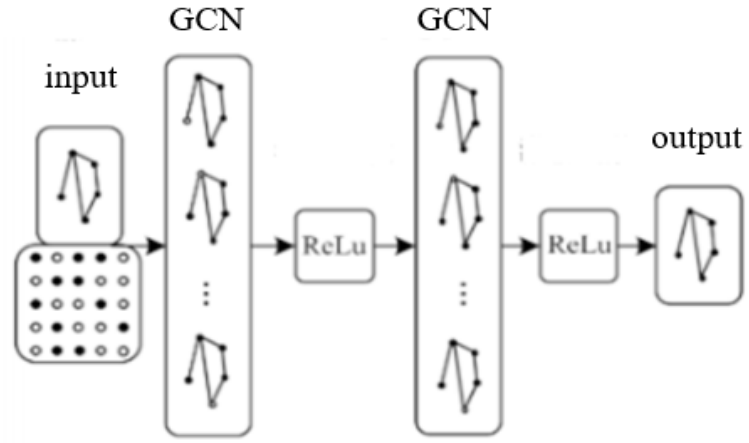


Fig.3 Fundamental Architectures of Graph Neural Networks

3.3.2. Graph Neural Networks for Multimodal Medical Image Diagnosis

As an emerging neural network, GNN has been widely used in the field of multimodal medicine. For example, Xing et al. [42] proposed a multimodal depression detection framework called EMO-GCN. The model introduced an institutional learning mechanism to reconstruct the graph structure through sparse attention after pooling, and used multiple GNNs to extract structural and acoustic features from EEG signals and speech respectively, achieving effective multimodal feature fusion. Considering the high heterogeneity and limited sample size of medical image data, Liu et al. [43] constructed a VMM-DGCN network, which used convolutional filters with different kernel sizes to capture the multi-scale feature representation of each subject, and introduced non-imaging information into the feature representation of each scale to construct multiple group graphs. Secondly, DeepGCN was used to extract high-level features of the group graph to complete the diagnosis of autism. However, most models ignore the spatiotemporal topological characteristics of brain networks. Xu et al. [44] developed an adaptive multi-channel GCN fusion framework with graph contrast learning. They first divided the ROI-based sequence signals into multiple overlapping time windows to construct a dynamic brain network representation. Secondly, they used adaptive multi-channel GCN to extract multimodal spatial features with contrast constraints, and input two stacked long short-term memory units to capture the temporal information transmitted by the time windows. Finally, they used MLP to realize the prediction of multimodal brain networks.

In the diagnosis of brain diseases, Chen et al. [45] selected structural magnetic resonance imaging (sMRI) and resting-state functional magnetic resonance imaging (rs-fMRI) data, analyzed the node features and edge characteristics, and finally achieved an accuracy of 95.8% through GNN. In order to avoid reconstructing the graph network and utilize the relationship between subjects, Tian et al. [46] proposed an scalable hierarchical graph convolutional network (EH-GCN). This network combines ResNet and GCN modules to integrate image features and connection features between different brain regions of interest (ROIs) into the feature representation of each subject, thereby extracting structural and functional connection features between brain ROIs. However, the limitation of this method is that it uses all brain ROIs for analysis without in-depth exploration of the impact of ROI selection on the final classification results, which may cause redundant information to interfere with model performance. Zhang et al. [47] proposed a method that combines imaging data with phenotypic data, captures individual features by constructing a brain network, and uses these features to represent the association between individuals and subjects in the potential population. This method further enriches the feature representation by introducing phenotypic data, but does not fully consider the multidimensional characteristics of image features. In terms of interpretability, Li et al. [48] constructed an interpretable iGLCN network by learning the optimal underlying latent graph to dynamically adjust the graph structure, which can better assist physicians in diagnosis.

Overall, these methods have made significant progress in processing multimodal medical images using graph convolutional networks, but they still face some challenges, such as optimizing ROI selection, improving computational efficiency, and further exploring multimodal data fusion. Future research can dig deeper in these directions to improve the performance and application scope of the model.

Table 3 summarizes some representative works of GCNs on multimodal medical images.

Tab.3 GCNs on Multimodal Medical Images

Autor	Model	Disease
Xing et al.	EMO-GCN	Depressive disorder
Liu et al.	VMM-DGCN	Autism spectrum disorder
Xu et al.	MSTGC	Epilepsy
Chen et al.	GCN	Schizophrenia
Tian et al.	EH-GCN	Alzheimer's disease
Zhang et al.	GCN	Alzheimer's disease
Li et al.	iGLCN	Parkinson's disease

4. Challenges and Prospects

4.1. Challenges

In recent years, the field of multimodal medical image classification has made significant progress, providing strong support for disease diagnosis, prognosis prediction and treatment plan formulation. By integrating medical image information of different modalities, such as CT, MRI, PET, etc., deep learning models can more comprehensively capture lesion characteristics and improve classification accuracy and robustness. However, this field still faces many challenges and also has great development potential.

(1) Multimodal medical image data comes from various sources, including CT, MRI, X-ray, ultrasound, pathological images, etc. Different modal data have significant differences in resolution, dimension, grayscale range, imaging principle, etc. For example, CT images provide high-resolution anatomical structure information, while MRI images are better at soft tissue imaging. This heterogeneity makes it difficult to directly fuse different modal data, and it is necessary to develop specialized algorithms to align and fuse multimodal information.

(2) Medical image annotation requires the participation of professional doctors, which is costly, time-consuming, and the annotation results may be subjective. In the medical field, different medical institutions use different equipment, set parameters, and scan times, which results in different presentation effects in images even for the same disease.

(3) The lack of model interpretability is a problem that all neural networks need to solve. Deep learning models are often viewed as "black boxes" and their decision-making process is difficult to explain. This is particularly important in the medical field, because doctors need to understand the basis of the model's judgment in order to be confident in the diagnosis and make correct clinical decisions. Some studies use model visualization to analyze the model's output results, such as using Grad-CAM [49] to visualize the decision area of the image, but there is still a certain gap between this and actual clinical needs.

4.2. Prospects

Deep learning eliminates the complex feature extraction work of traditional diagnostic methods, extracts deep features from image data, and assists doctors in quickly determining the grade and type of brain tumors. Looking ahead, multimodal medical image classification methods will continue to expand in the following areas:

(1) Researching more effective multimodal fusion algorithms is the key to improving model performance. For example, attention-based methods can automatically learn the importance weights of different modal data, while graph neural networks can model the complex relationships between different modal data. In addition, multimodal data generation methods based on generative adversarial networks (GANs) can also be explored to expand the training dataset and improve the robustness of the model.

(2) Classification models for 3D medical images have great potential, and using 3D imaging data for disease diagnosis will definitely be a future development trend. Unlike traditional 2D images, 3D images can provide more comprehensive tumor morphology and spatial information, allowing the model to more accurately identify tumor boundaries and sizes, thereby significantly improving the accuracy of classification. However, the computational complexity of processing 3D data is very high, and more efficient algorithms and model frameworks need to be developed.

(3) Federated learning is a distributed machine learning method that can achieve data sharing and model training across medical institutions while protecting data privacy. This is particularly important for medical image analysis, because medical data is usually scattered across different medical institutions and involves patient privacy. Federated learning can effectively utilize multi-center data, improve the generalization ability of the model, and promote the development of medical artificial intelligence.

Reference

[1] G. Muhammad, F. Alshehri, F. Karray, A. E. Saddik, M. Alsulaiman, and T. H. Falk, "A comprehensive survey on multimodal medical signals fusion for smart healthcare systems," Information Fusion, vol. 76, pp. 355–375, Dec. 2021, doi: 10.1016/j.inffus.2021.06.007.

[2] M. A. Azam et al., "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics," Computers in Biology and Medicine, vol. 144, p. 105253, May 2022, doi: 10.1016/j.compbiomed.2022.105253.

[3] Y. Li et al., "A review of deep learning-based information fusion techniques for multimodal medical image classification," Computers in Biology and Medicine, vol. 177, p. 108635, Jul. 2024, doi: 10.1016/j.compbiomed.2024.108635.

[4] J. Lipkova et al., "Artificial intelligence for multimodal data integration in oncology," Cancer Cell, vol. 40, no. 10, pp. 1095–1110, Oct. 2022, doi: 10.1016/j.ccell.2022.09.012.

[5] H. Hermessi, O. Mourali, and E. Zagrouba, "Multimodal medical image fusion review: Theoretical background and recent advances," Signal Processing, vol. 183, p. 108036, Jun. 2021, doi: 10.1016/j.sigpro.2021.108036.

[6] R. Girshick, "Fast R-CNN," in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169.

[7] S. Soni, S. S. Chouhan, and S. S. Rathore, "TextConvoNet: A convolutional neural network based architecture for text classification," Applied Intelligence, vol. 53, no. 11, pp. 14249–14268, 2023.

[8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[9] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biol. Cybernetics, vol. 36, no. 4, pp. 193–202, Apr. 1980, doi: 10.1007/BF00344251.

[10] M. Khalid, J. Baber, M. K. Kasi, M. Bakhtyar, V. Devi, and N. Sheikh, "Empirical Evaluation of Activation Functions in Deep Convolution Neural Network for Facial Expression Recognition," in 2020 43rd International Conference on Telecommunications and Signal Processing (TSP), Jul. 2020, pp. 204–207. doi: 10.1109/TSP49548.2020.9163446.

[11] T. Li, F. Zhang, G. Xie, X. Fan, Y. Gao, and M. Sun, "A high speed reconfigurable architecture for softmax and GELU in vision transformer," Electronics Letters, vol. 59, no. 5, p. e12751, Mar. 2023, doi: 10.1049/ell2.12751.

[12] J. Hyun, H. Seong, and E. Kim, "Universal pooling – A new pooling method for convolutional neural networks," Expert Systems with Applications, vol. 180, p. 115084, Oct. 2021, doi: 10.1016/j.eswa.2021.115084.

[13] J. Zhang, X. He, Y. Liu, Q. Cai, H. Chen, and L. Qing, "Multi-modal cross-attention network for Alzheimer's disease diagnosis with multi-modality data," Computers in Biology and Medicine, vol. 162, p. 107050, Aug. 2023, doi: 10.1016/j.compbiomed.2023.107050.

[14] T. Zhang and M. Shi, "Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease," Journal of Neuroscience Methods, vol. 341, p. 108795, Jul. 2020, doi: 10.1016/j.jneumeth.2020.108795.

[15] Z. Kong, M. Zhang, W. Zhu, Y. Yi, T. Wang, and B. Zhang, "Multi-modal data Alzheimer's disease detection based on 3D convolution," Biomedical Signal Processing and Control, vol. 75, p. 103565, May 2022, doi: 10.1016/j.bspc.2022.103565.

[16] W. N. Ismail, F. R. P. P., and M. A. S. Ali, "A Meta-Heuristic Multi-Objective Optimization Method for Alzheimer's Disease Detection Based on Multi-Modal Data," Mathematics, vol. 11, no. 4, Art. no. 4, Jan. 2023, doi: 10.3390/math11040957.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks".

[18] C. Szegedy et al., "Going Deeper with Convolutions," Sep. 17, 2014, arXiv: arXiv:1409.4842. doi: 10.48550/arXiv.1409.4842.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[20] C. Ge, I. Y.-H. Gu, A. S. Jakola, and J. Yang, "Deep Learning and Multi-Sensor Fusion for Glioma Classification Using Multistream 2D Convolutional Networks," in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jul. 2018, pp. 5894–5897. doi: 10.1109/EMBC.2018.8513556.

[21] A. Puente-Castro, E. Fernandez-Blanco, A. Pazos, and C. R. Munteanu, "Automatic assessment of Alzheimer's disease diagnosis based on deep learning techniques," Computers in Biology and Medicine, vol. 120, p. 103764, May 2020, doi: 10.1016/j.compbiomed.2020.103764.

[22] Y. Tu, S. Lin, J. Qiao, Y. Zhuang, and P. Zhang, "Alzheimer's disease diagnosis via multimodal feature fusion," Computers in Biology and Medicine, vol. 148, p. 105901, Sep. 2022, doi: 10.1016/j.compbiomed.2022.105901.

[23] Y. Li et al., "Multimodal Information Fusion for Glaucoma and Diabetic Retinopathy Classification," in Ophthalmic Medical Image Analysis, Springer, Cham, 2022, pp. 53–62. doi: 10.1007/978-3-031-16525-2_6.

[24] T. K. Yoo et al., "DeepPDT-Net: predicting the outcome of photodynamic therapy for chronic central serous chorioretinopathy using two-stage multimodal transfer learning," Sci Rep, vol. 12, no. 1, p. 18689, Nov. 2022, doi: 10.1038/s41598-022-22984-6.

[25] [25] X. Huang et al., "Detecting glaucoma from multi-modal data using probabilistic deep learning," Front Med (Lausanne), vol. 9, p. 923096, 2022, doi: 10.3389/fmed.2022.923096.

[26] X. Qian et al., "A combined ultrasonic B-mode and color Doppler system for the classification of breast masses using neural network," Eur Radiol, vol. 30, no. 5, pp. 3023–3033, May 2020, doi: 10.1007/s00330-019-06610-0.

[27] M. H. Le et al., "Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks," Phys Med Biol, vol. 62, no. 16, pp. 6497–6514, Jul. 2017, doi: 10.1088/1361-6560/aa7731.

[28] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi, "Skin Disease Recognition Using Deep Saliency Features and Multimodal Learning of Dermoscopy and Clinical Images," in Medical Image Computing and Computer Assisted Intervention − MICCAI 2017, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds., Cham: Springer International Publishing, 2017, pp. 250–258. doi: 10.1007/978-3-319-66179-7_29.

[29] S. Li, Y. Xie, G. Wang, L. Zhang, and W. Zhou, "Attention guided discriminative feature learning and adaptive fusion for grading hepatocellular carcinoma with Contrast-enhanced MR," Computerized Medical Imaging and Graphics, vol. 97, p. 102050, Apr. 2022, doi: 10.1016/j.compmedimag.2022.102050.

[30] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27, Springer, 2018, pp. 270–279.

[31] Y. Yang et al., "Glioma grading on conventional MR images: a deep learning study with transfer learning," Frontiers in neuroscience, vol. 12, p. 804, 2018.

[32] A. Vaswani, "Attention is all you need," Advances in Neural Information Processing Systems, 2017.

[33] I. Sutskever, "Sequence to Sequence Learning with Neural Networks," arXiv preprint arXiv:1409.3215, 2014.

[34] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[35] Y. Lyu, X.-W. Yu, D. Zhu, and L. Zhang, "Classification of Alzheimer's Disease via Vision Transformer: Classification of Alzheimer's Disease via Vision Transformer," Petra, 2022, Accessed: Feb. 24, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Classification-of-Alzheimer%27s-Disease-via-Vision-of-Lyu-Yu/9da3fadf092c864f61d6fd1e8eab5a6ca2397194

[36] J. Zhu et al., "Efficient self-attention mechanism and structural distilling model for Alzheimer's disease diagnosis," Computers in Biology and Medicine, vol. 147, p. 105737, Aug. 2022, doi: 10.1016/j.compbiomed.2022.105737.

[37] J. Jang and D. Hwang, "M3T: three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022, pp. 20686–20697. doi: 10.1109/CVPR52688.2022.02006.

[38] L. Qiu et al., "Hierarchical multimodal fusion framework based on noisy label learning and attention mechanism for cancer classification with pathology and genomic features," Computerized Medical Imaging and Graphics, vol. 104, p. 102176, Mar. 2023, doi: 10.1016/j.compmedimag.2022.102176.

[39] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers Advance Multi-Modal Medical Image Classification," Diagnostics, vol. 11, no. 8, Art. no. 8, Aug. 2021, doi: 10.3390/diagnostics11081384.

[40] Y. Zhang, Y. Deng, Z. Zhou, X. Zhang, P. Jiao, and Z. Zhao, "Multimodal learning for fetal distress diagnosis using a multimodal medical information fusion framework," Front Physiol, vol. 13, p. 1021400, 2022, doi: 10.3389/fphys.2022.1021400.

[41] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," Feb. 22, 2017, arXiv: arXiv:1609.02907. doi: 10.48550/arXiv.1609.02907.

[42] T. Xing, Y. Dou, X. Chen, J. Zhou, X. Xie, and S. Peng, "An adaptive multi-graph neural network with multimodal feature fusion learning for MDD detection," Sci Rep, vol. 14, no. 1, p. 28400, Nov. 2024, doi: 10.1038/s41598-024-79981-0.

[43] S. Liu, S. Wang, C. Sun, B. Li, S. Wang, and F. Li, "DeepGCN based on variable multi-graph and multimodal data for ASD diagnosis," CAAI Transactions on Intelligence Technology, vol. 9, no. 4, pp. 879–893, 2024, doi: 10.1049/cit2.12340.

[44] R. Xu, Q. Zhu, S. Li, Z. Hou, W. Shao, and D. Zhang, "MSTGC: Multi-Channel Spatio-Temporal Graph Convolution Network for Multi-Modal Brain Networks Fusion," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 2359–2369, 2023, doi: 10.1109/TNSRE.2023.3275608.

[45] X. Chen et al., "Discriminative analysis of schizophrenia patients using graph convolutional networks: A combined multimodal MRI and connectomics analysis," Front. Neurosci., vol. 17, Mar. 2023, doi: 10.3389/fnins.2023.1140801.

[46] X. Tian, Y. Liu, L. Wang, X. Zeng, Y. Huang, and Z. Wang, “An extensible hierarchical graph convolutional network for early Alzheimer’s disease identification,” Computer Methods and Programs in Biomedicine, vol. 238, p. 107597, Aug. 2023, doi: 10.1016/j.cmpb.2023.107597.

[47] Y. Zhang, X. He, Y. H. Chan, Q. Teng, and J. C. Rajapakse, “Multi-modal graph neural network for early diagnosis of Alzheimer’s disease from sMRI and PET scans,” Computers in Biology and Medicine, vol. 164, p. 107328, Sep. 2023, doi: 10.1016/j.compbiomed.2023.107328.

[48] F. Li et al., “Developing a Dynamic Graph Network for Interpretable Analysis of Multi-Modal MRI Data in Parkinson’s Disease Diagnosis,” in 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Jul. 2023, pp. 1–4. doi: 10.1109/EMBC40787.2023.10340672.

[49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” Int J Comput Vis, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.