Multimodal Electronic Medical Record Disease Classification Combining Image and Text Data

Hemachandran K*

Professor, AI Research Centre, School of Business, Woxsen University, India; hemachandran.k@woxsen.edu.in

*Corresponding Author: hemachandran.k@woxsen.edu.in

DOI: https://doi.org/10.30210/JMSO.202503.017

Submitted: Aug, 25, 2025 Accepted: Nov, 01, 2025

ABSTRACT

Multimodal electronic health records (EHRs) in the healthcare field contain rich text and image data, providing crucial information for the diagnosis and classification of complex diseases. However, most existing research focuses on single-modality data analysis, overlooking the potential complementarity between different modalities. Furthermore, even the few proposed multimodal fusion methods still suffer from issues such as imprecise modality alignment, poor feature fusion, and high module design complexity. To overcome these challenges, this study proposes a multimodal fusion-based EHR disease classification method specifically for the task of tumor segmentation. The main contribution of this work consists of designing a novel multimodal fusion model that has the modality alignment, self-attention, residual connection, and dynamic weighting of features in one model. This combination is not only dealing with the serious problems of feature misalignment and information loss but also it is the ability of fusion to be flexible extending the contribution of each modality to dynamically change.Our model combines modality alignment, self-attention, residual connections, and a dynamic feature weighting module to effectively fuse text and image data. We conducted comprehensive experiments on the BraTS 2015 and BraTS 2018 datasets. The results demonstrate that the proposed method significantly outperforms existing methods in multiple metrics, including the Dice coefficient, positive predictive value (PPV), and sensitivity.

Keywords: Multimodal Fusion, Modality Alignment, Self-Attention Mechanism, Tumor Segmentation, Electronic Medical Records

1. Introduction

The Electronic Health Records (EHRs) as a simple medical information management tool in the contemporary medicine sector have enhanced the development of clinical diagnosis and research to high levels. In oncology specifically [1-3], doctors rely on different forms of information as one method of arriving at the right diagnosis and treatment decisions. EMRs are not textual per se with the text containing this information as diagnostic reports, treatment records, and pathological analyses but also contain visual information, such as radiological images, CT scans, and magnetic resonance imaging (MRI). The combination of the information offered in different modalities makes it possible to comprehend the state of the patient to a greater extent and provides a better foundation to an individual approach to treatment.

Tumors are heterogeneous hence multimodal data are of particular interest in tumor analysis because this is a complex disease. However, the efficient processing of all these types of data and the intelligent diagnostics and classification through complex artificial intelligence is a major issue of the medical profession. Single-modal classification A single-mode classification task normally involves text or image data, whereas natural language processing (NLP) is used to analyze textual reports or convolutional neural networks (CNNs) are used to analyze images. Such methods do not drain the complementary aspects of multimodal data leading to incomplete representations and low classification errors.

Various experiments have examined multimodal-based techniques that combine information through text and images [7-9], which suggest that the joint analysis is more effective than the single modality of the experiment. However, in the current models of multimodal fusion, loopholes are present to a large extent. To begin with, the initial type of fusion is generally prone to redundancy of information and overfitting. Second, modalities of interaction between modalities are lacking in late fusion strategies. Third, there are also some difficulties in aligning the heterogeneous data and developing effective fusion mechanisms even though intermediate fusion strategies are used. Additionally, some of the models that can be used are computationally expensive and complex which is not that feasible as far as the actual medical application is concerned.

The paper will fill such gaps by proposing a multimodal electronic medical record disease classification algorithm whose specialization is oncology. We are particularly interested in the present work based on the formation of a mid-level fusion framework enhanced with the self-attention, cross-attention, and modality alignment mechanisms. The design allows the text and image features to relate dynamically, in not only making sure that there is effective alignment but also useful semantic integration. Additionally, a dynamic feature weighting module is proposed, and it helps the model to dynamically adjust the roles of different modalities to improve the classification accuracy and robustness.

To make sure that our approach would work, we evaluate the performance of our method using single-modal methods, as well as other multimodal fusion methods on benchmark datasets such as BraTS 2015 and BraTS 2018. The findings do not only indicate the excellent classification powers of our model, but also the potential of our model to the real-world oncology diagnostics.

Multimodal electronic medical record analysis and classification is a hot research direction in the field of medical artificial intelligence.

2.1 Single-Modality Electronic Medical Record Data Analysis

Single-modal data analysis mainly focuses on a certain data type in electronic medical records, such as text data or medical images. Disease classification methods based on medical images use CNNs to extract image features to identify or classify specific diseases. In the field of text analysis, traditional feature engineering-based methods such as TF-IDF and bag-of-words are used to extract keywords and features from medical records. Models such as BERT [17] and BioBERT [18] can capture complex semantic information in medical record text through pre-trained bidirectional Transformer architectures and show good classification performance. CNNs are widely used in disease classification tasks. Deep CNN architectures such as ResNet [19] and DenseNet [20] can effectively extract spatial features from medical images for tasks such as tumor detection and lesion segmentation. For example, ResNet has been successfully applied to tasks such as breast cancer detection [21] [31] [32] and lung cancer classification [22] [33] [34], showing extremely high classification accuracy. However, single-modality methods often have limitations when dealing with complex medical problems. For complex diseases such as tumors, relying solely on single-modality data such as text or images may not fully reflect the patient's health status. Therefore, researchers began to explore how to combine multimodal data such as text and images to improve the accuracy of disease classification.

2.2 Multimodal Fusion Methods

Multimodal fusion methods have gained widespread attention in medical image processing and disease classification tasks in recent years. Early fusion methods directly splice data from different modalities at the input stage to generate a unified feature representation. For example, Shen et al. [23] proposed an early fusion model that splices image and text data for automatic diagnosis of cancer. This method is simple and direct, but it is prone to information loss or redundancy when faced with highly heterogeneous data.

Suk et al. [24] proposed an intermediate fusion model based on the attention mechanism to combine MRI images and clinical data to improve the classification accuracy of Alzheimer's disease. This method achieves dynamic information interaction between modalities by introducing a cross-attention mechanism in the middle layer, and shows high robustness in processing complex tasks. The late fusion method performs fusion after the classification results of each modality are generated. For example, Zhang et al. [25] adopted a late fusion strategy in the tumor classification task, integrating the prediction results of different modalities through a weighted average method, thereby improving the classification accuracy. However, the late fusion method often fails to fully utilize the interactive information between modalities, resulting in limited performance in processing complex medical problems. With the continuous advancement of deep learning technology, some researchers have begun to explore multimodal fusion methods based on deep neural networks in recent years. Dai et al. [26] proposed a hybrid model based on CNNs and Transformers in their study, which deeply fuses multimodal data through a self-attention mechanism for multimodal medical image analysis. Similarly, Dwivedi et al. [27] also proposed a multimodal fusion framework based on GNN and Transformers for tumor classification, showing good experimental performance. In addition, some studies have introduced generative adversarial networks (GANs) into multimodal fusion tasks to achieve data synthesis and supplementation. For example, Dou et al. [28] proposed a framework based on multimodal data GANs to generate additional modal data to improve classification performance. This method can effectively make up for the problem of missing data between modalities through the game between the generator and the discriminator.

Another key challenge in multimodal learning is how to deal with the heterogeneity and alignment problems between modalities. Zhu et al. [29] proposed a modality alignment method that ensures the alignment of features generated in different modalities in the shared space by introducing a contrast loss function. Wang et al. [30] proposed a cross-modal attention mechanism that improves the model's fusion efficiency for different modal data by adaptively allocating attention weights to different modalities.

In the medical field, especially in tumor classification tasks, multimodal fusion methods have gradually shown their great application potential. Data from different modalities can complement each other and jointly improve the robustness and generalization ability of classification models. However, existing methods still face many challenges in practical applications, especially when dealing with large-scale, multimodal data. How to design an efficient fusion mechanism is still an urgent problem to be solved. Although existing multimodal electronic medical record analysis has made significant progress in disease classification and diagnosis, it still has some shortcomings. First, early fusion methods usually simply splice data from different modalities[23]. Second, although mid-term and late fusion methods can improve the above problems to a certain extent by extracting features or results in stages and then fusing them, their model complexity is high and they lack the ability to address the interaction between modalities [25]. In addition, existing research mostly focuses on simple modality alignment strategies, which cannot fully address the heterogeneity between modalities [29].

We use a modality alignment mechanism to ensure that features from different modalities are aligned in a shared space, thereby improving classification accuracy after fusion. Furthermore, by introducing a dynamic feature extraction module based on deep learning, our approach significantly improves classification performance while maintaining model efficiency, particularly in complex tumor classification tasks.

3. Method

Our approach consists of two main modules: a text-image feature extraction module and a cross-modal feature fusion module. In this section, we describe the first module, the text-image feature extraction module, in detail.

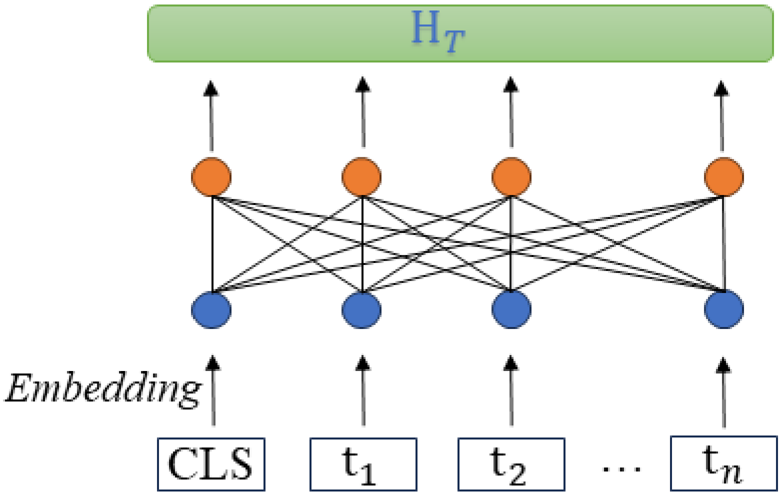

Fig 1. Schematic diagram of text feature extraction structure.

3.1 Text-image feature extraction module

Our approach is implemented in two submodules: a text feature extraction submodule and an image feature extraction submodule. After feature extraction, a preliminary multimodal representation is generated through feature concatenation.

Text Feature Extraction Submodule

Text input can include information such as diagnosis reports, treatment plans, and medical history records in medical records. Given input text T=$\{t_1, t_2, \dots, t_n\}$, where nnn is the length of the text (i.e., the number of words contained), we first encode the text using the BERT model. The BERT model consists of multiple Transformer layers, each of which uses a self-attention mechanism to calculate dependencies between words. First, the input text is embedded:

$$E = \{e_1, e_2, \dots, e_n\} = Embedding(T) \qquad \text{[Formular 1]}$$

The embedded text is then encoded through a multi-layer Transformer. The specific process is as follows:

$$H_T = BERT_L(E) \qquad \text{[Formular 2]}$$

Where L represents the number of BERT layers. In the last layer of the model, we use the output of the [CLS] tag as the global feature representation of the entire text:

$$H_T = H_T^{[CLS]} \in R^{dT} \qquad \text{[Formular 3]}$$

Here, the [CLS] tag is a special tag used to indicate a sentence-level classification task, and its output vector represents the comprehensive semantic information of the entire input text.

Image Feature Extraction Submodule

For feature extraction of medical images (such as MRI or CT), we use the ResNet-50 model. ResNet-50 is a deep CNN that solves the vanishing gradient problem in deep neural networks through residual connections and can effectively extract spatial features from images. Given an input image I, we first perform layer-by-layer processing through multiple convolutional layers, batch normalization layers, and pooling layers of ResNet-50 to extract multi-level features. The convolution operation can be expressed as:

$$X^{(l+1)} = ReLU(BatchNorm(W^{(l)} * X^{(l)} + b^{(l)})) \qquad \text{[Formular 4]}$$

Here, $X^{(l)}$ represents the input features of layer l, $W^{(l)}$ and $b^{(l)}$ are the weights and bias parameters of this layer, respectively. represents the convolution operation, and ReLU is the activation function. Finally, the feature maps are globally pooled through the global average pooling layer (GAP) to obtain a fixed-length feature vector:

$$H_I = GAP(X^{(L)}) \qquad \text{[Formular 5]}$$

The GAP operation compresses the spatial features into a global representation by calculating the average of each feature map, thereby generating a fixed-dimensional feature vector for ease of subsequent processing.

Feature Concatenation and Regularization

After extracting the text features H_T and the image features H_I, we integrate the features of the two modalities through a simple concatenation operation. This concatenation operation concatenates the two feature vectors by dimension, forming a preliminary multimodal feature representation:

$$H_{TI} = [H_T; H_I] \in R^{d_T + d_I} \qquad \text{[Formular 6]}$$

This concatenation method preserves the independent features of text and image modalities while providing rich context and spatial information for the subsequent cross-modal fusion module. To prevent the high dimensionality of the feature vector from causing model overfitting, we introduced L2 regularization on the concatenated feature representation:

$$H_{TI}^{reg} = \frac{H_{TI}}{\|H_{TI}\|_2} \qquad \text{[Formular 7]}$$

Here, $\| H_{TI} \|_2$ represents the L2 norm of the feature vector. Regularization can reduce feature amplitude differences, making it easier to fuse features from different modalities.

Feature Extraction Enhanced by Attention Mechanism

To further enhance the expressiveness of text and image features, we introduced a self-attention mechanism to weight the features of each modality. The self-attention mechanism dynamically adjusts the weight of each feature within the global feature set, enhancing more important information while weakening less important information.

First, for text features, we calculate the attention weight between each word and the global text feature:

$$\alpha_i = \frac{exp(det(H_T, e_i))}{\sum_{j=1}^{n} \quad exp(det(H_T, e_j))} \qquad \text{[Formular 8]}$$

Here, α_i is the attention weight for the iii-th word. We then perform a weighted summation of the word embeddings based on the attention weights to obtain the enhanced text feature representation:

$$H_T^{att} = \sum_{i=1}^{n} \quad \alpha_i e_i \qquad \text{[Formular 9]}$$

Similarly, for image features, we use the channel attention mechanism to weight the feature maps of different channels. Given the feature map X(L) generated by the convolution operation, we first perform global pooling on the features of each channel to obtain the channel description vector z_c:

$$z_c = \frac{1}{h' \times w'} \sum_{i=1}^{h'} \quad \sum_{j=1}^{w'} \quad X_{i,j,c} \qquad \text{[Formular 10]}$$

Then, the attention weight of each channel is calculated:

$$\beta_c = \frac{exp(z_c)}{\sum_{k=1}^{c'} \quad exp(z_k)} \qquad \text{[Formular 11]}$$

Finally, the feature maps are weighted and summed according to the channel weights to obtain the enhanced image feature representation:

$$H_I^{att} = \sum_{c=1}^{c'} \quad \beta_c X_c^{(L)} \qquad \text{[Formular 12]}$$

Through the self-attention mechanism, we can enhance the key information in each modality, ensuring that the features in the subsequent fusion stage are more representative.

Feature Orthogonalization

To ensure that the features of text and image are not overly coupled, we introduce a feature orthogonalization constraint to prevent high correlation between the concatenated feature vectors. Specifically, we require the inner product of the text and image features to be close to zero, ensuring that the features of the two modalities remain independent before fusion. We introduce the following orthogonalization loss:

$$L_{ortho} = \| H_T^T H_I \|_2 \qquad \text{[Formular 13]}$$

By minimizing this loss term, we ensure that the features of the two modalities remain as independent as possible, thereby improving the information richness of the subsequent fusion stage.

Through the above steps, the text-image feature extraction module can generate multimodal feature representations rich in semantic and spatial information, providing a solid foundation for the subsequent cross-modal fusion module.

3.2 Cross-Modal Feature Fusion Module

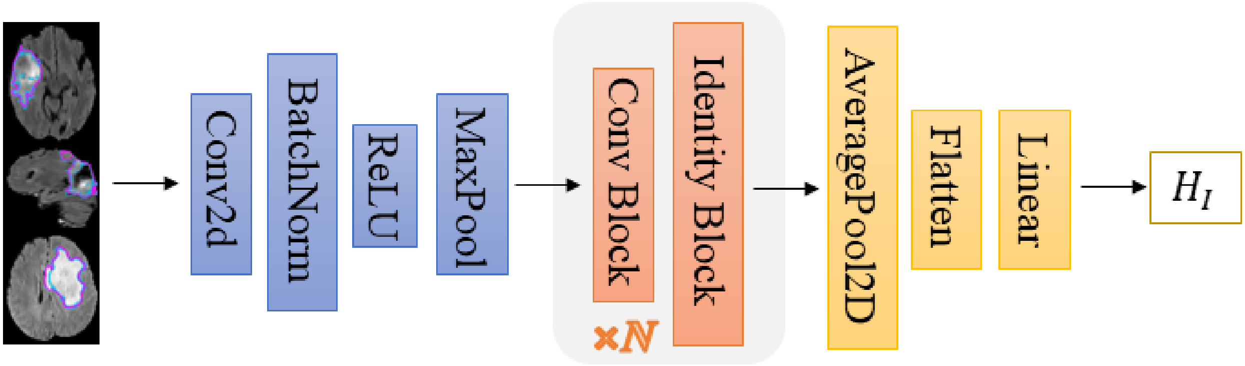

Fig 2. Schematic diagram of image feature extraction structure.

In the previous module, we extracted deep feature representations from text and images. To further improve the model's classification performance, we designed a cross-modal feature fusion module. This module combines a self-attention mechanism with a modality alignment strategy to efficiently fuse text and image information. In this section, we will detail how to leverage this multimodal feature fusion mechanism to capture complementary information between different modalities, thereby improving classification accuracy and robustness.

3.3 Modality Alignment Mechanism

A major challenge in cross-modal fusion is effectively aligning text and image features. Because text and images come from different data sources and exhibit significant heterogeneity, we designed a modality alignment mechanism to mitigate this discrepancy. Through modality alignment, we ensure that features from different modalities remain consistent within a shared space during the fusion process, thereby improving fusion performance.

The core idea of modality alignment is to calculate the similarity between features from different modalities and generate a weight matrix to align image and text features. Given the text features H_T and the image features H_I, we first normalize the two feature vectors to ensure that the feature vectors of different modalities have the same scale in the same feature space:

$$H_T^{norm} = \frac{H_T}{\|H_T\|_2}, H_I^{norm} = \frac{H_I}{\|H_I\|_2} \qquad \text{[Formular 14]}$$

Next, we use cosine similarity to measure the similarity between text and image features:

$$S = \frac{H_T^{norm} \cdot H_I^{norm}}{\|H_T^{norm}\|_2 \|H_I^{norm}\|_2} \qquad \text{[Formular 15]}$$

Through the similarity matrix, we can measure the feature similarity between the modalities and use this result in the subsequent alignment process. To ensure that the features between the two modalities can be well aligned, we introduce an alignment loss function:

$$L_{align} = \| S - I \|_F^2 \qquad \text{[Formular 16]}$$

By minimizing this alignment loss, text and image features can better cooperate in the subsequent fusion step, improving the overall model performance.

Dynamic Feature Weighting

In real applications, data from different modalities may contribute differently to the classification task. For example, in some cases, imaging data may provide more useful lesion information, while in other cases, text descriptions may be more important. To address this issue, we designed a dynamic feature weighting mechanism to dynamically adjust the weights of text and image features based on the specific task, thereby improving the model's flexibility and generalization ability.

We use an adaptive weighting mechanism to assign different weights to text and image features. Specifically, we learn two weight parameters, α_T and α_I, to represent the importance of text and image features, respectively:

$$\alpha_T = \sigma(W_T H_T), \alpha_I = \sigma(W_I H_I) \quad \text{[Formular 17]}$$

Among them, σ represents the Sigmoid activation function, W_T and W_I are learnable weight matrices. With these weights, the final fused feature representation can be written as:

$$H_{weighted} = \alpha_T H_T + \alpha_I H_I \quad \text{[Formular 18]}$$

This weighting mechanism dynamically adjusts the weight distribution between modalities based on the characteristics of the input data, ensuring that the model can better handle different types of data and tasks, thereby improving classification performance.

Classification and Output

After the above steps, the fused multimodal features $H_{weighted}$ now contain complete cross-modal information. To utilize this information for disease classification, we designed a fully connected network for the final classification prediction. Specifically, we input the fused features into a fully connected layer and perform a nonlinear transformation using the ReLU activation function:

$$H_{final} = ReLU(W_f H_{weighted} + b_f) \quad \text{[Formular 19]}$$

Next, we generate the probability distribution P for each category through the Softmax layer:

$$P = Softmax(W_c H_{final} + b_c) \quad \text{[Formular 20]}$$

Among them, W_c and b_c are the parameters of the classification layer. The final classification result is determined by the category corresponding to the maximum probability:

$$\hat{y} = arg\,max(P) \quad \text{[Formular 21]}$$

Our cross-modal feature fusion module successfully achieves deep fusion of text and image features through strategies such as self-attention[14, 15], modality alignment, residual connections, and dynamic feature weighting. This module not only effectively captures the interdependencies between modalities but also dynamically adjusts the contribution of each modality in specific tasks through adaptive mechanisms, ensuring that the fused features perform well in classification tasks. Through residual connections and alignment, the model avoids information loss during the fusion process and significantly improves classification accuracy and robustness.

4. Experiment

4.1 Datasets and Evaluation Metrics

We used three public brain tumor datasets, BraTS 2018 [35] and BraTS 2015 [35], which are from the Brain Tumor Segmentation Challenge. These datasets are widely used to study the automatic detection and segmentation of brain tumors, especially multimodal magnetic resonance imaging (MRI) datasets. Each dataset provides tumor segmentation labels annotated by professional medical personnel and contains detailed medical text descriptions. The BraTS 2018 dataset contains brain MRI scans of 285 patients, divided into a training set (210 cases), a validation set (28 cases), and a test set (47 cases). Each case contains multimodal MRI data, including T1, T2, T1CE (contrast-enhanced T1 weighted), and FLAIR (fluid-attenuated inversion recovery). The data of these modalities help doctors fully understand the morphology, size, and location of the tumor. In addition to imaging data, the BraTS 2018 dataset also provides textual descriptions of medical records, including patient medical history, diagnosis information, and treatment plans.

In BraTS 2018, tumors are divided into three primary regions: tumor core, enhancing tumor, and whole tumor. These regions provide more detailed annotations for model classification and segmentation, enabling more nuanced evaluation.

The BraTS 2017 dataset is very similar to the BraTS 2018 dataset, containing multimodal MRI data and tumor segmentation labels for 285 patients. The data also provides T1, T2, T1CE, and FLAIR modalities. The annotation and processing of this data are consistent with the 2018 dataset, making it crucial for evaluating the transferability and generalization capabilities of the model. BraTS 2017 also provides textual descriptions of medical records and detailed annotations of each patient's tumor region. This dataset further helps us verify the consistency of our model's performance across data from different years. This dataset emphasizes the combined use of multimodal MRI data, which helps improve the precise understanding of brain tumor morphology and size. Furthermore, since different modalities in MRI data reflect distinct characteristics of tumor tissue, such as enhanced contrast signal and fluid attenuation, combining this information can enhance the model's classification and segmentation capabilities.

The BraTS 2015 dataset is an earlier public dataset, containing multimodal MRI scans from 274 cases. Furthermore, the text descriptions in the BraTS 2015 dataset provide detailed diagnostic information, medical history, and treatment recommendations for each case. This textual data provides rich semantic information for our text feature extraction module in multimodal learning.

The annotations in the BraTS 2015 dataset focus on detailed classification of tumor regions, particularly providing precise annotations for the segmentation of key regions such as enhancing tumors and tumor cores. These characteristics make this dataset an important benchmark for evaluating the multimodal learning capabilities of models.

To ensure the effectiveness and generalization of the models, we performed normalization and preprocessing on all three datasets. Specifically, for the imaging data, we performed the following processing:

1. Normalization: The pixel values of each case's MRI image were normalized to a mean of 0 and a variance of 1. This step ensures a consistent distribution of the imaging data, which helps improve model training performance.

2. Data Augmentation: To prevent model overfitting and improve its robustness to diverse scenarios, we performed random augmentation operations on the imaging data. These operations include random rotations, translations, scaling, and mirror flips to simulate different camera angles and positions in real-world scenarios.

To comprehensively measure the performance of our model, we used a series of commonly used evaluation metrics, such as the Dice coefficient, positive predictive value (PPV), sensitivity, and Hausdorff distance (Haus). These metrics are commonly used in medical image segmentation and classification.

The Dice coefficient measures the similarity between the model's predicted segmentation results and the true labels. The Dice coefficient ranges from 0 to 1, with higher values indicating closer alignment between the model's segmentation results and the true labels. Its formula is as follows:

$$Dice = \frac{2TP}{FP + 2TP + FN}$$

Where TP represents true positives, FP represents false positives, and FN represents false negatives.

PPV measures the proportion of samples predicted as positive by the model that are actually positive. PPV is used to assess model accuracy and is calculated as:

$$PPV = \frac{TP}{FP + TP}$$

PPV reflects the model's accuracy for positive samples. A higher PPV indicates a more accurate model for positive predictions.

Sensitivity reflects the model's ability to detect positive samples, that is, the proportion of samples predicted as positive to all actual positive samples. The calculation formula is:

$$Sensitivity = \frac{TP}{TP + FN}$$

A higher sensitivity makes it suitable for tasks with a high probability of missing detection.

The Hausdorff distance measures the maximum difference between two sets. It is often used to evaluate the boundary similarity between predicted results and true labels in segmentation tasks. The calculation formula is:

$$Haus(T,P) = max\left\{sup_{t\in T}\, inf_{p\in P}\, d(t,p), sup_{p\in P}\, inf_{t\in T}\, d(t,p)\right\}$$

Where T is the set of boundary points of the true label, P is the set of boundary points of the predicted segmentation, and d(t, p) represents the distance between the two points.

4.2 Main results and analysis

Table 1. Experimental Results on BraTS 2018.

ModelConfiguration	Dice(%)	PPV(%)	Sensitivity(%)	HausdorffDistance(mm)
Single-modalityImage(ResNet-50)	77.5	74.2	79.0	7.85
Single-modalityImage+Text(NoAlignment,NoAttention)	80.2	77.1	81.3	7.32
+ModalityAlignmentMechanism	82.5	80.4	83.0	6.85
+Self-attentionMechanism	84.7	82.1	85.5	6.23
+ResidualConnection	85.8	83.0	86.5	6.02
+DynamicFeatureWeightingMechanism	87.3	85.1	88.0	5.58

We conduct detailed experiments on the BraTS 2018 dataset. These experiments aim to validate the effectiveness of our proposed multimodal EMR disease classification method and analyze the impact of each component on model performance. We sequentially add modules, such as the modality alignment mechanism, self-attention mechanism, residual connections, and dynamic feature weighting mechanism, to demonstrate their contribution to the final classification performance.

The detailed analysis is as follows:

1. Using Single-Modality Images Only (ResNet-50)

As a baseline model, ResNet-50 achieves a Dice coefficient of 77.5% for tumor classification using only image data. While image data provides rich spatial information, it lacks the semantic information found in text descriptions, resulting in relatively low classification performance.

2. Single-Modality Images + Text (No Alignment, No Attention)

Introducing text data improves model performance, increasing the Dice coefficient to 80.2%. This suggests that the medical history and diagnostic information contained in text descriptions is helpful for tumor classification. However, the combination of text and image features still has certain limitations.

3. Adding a Modality Alignment Mechanism

After introducing the modality alignment mechanism, the Dice coefficient further improved to 82.5%. This demonstrates that aligning text and image features can better capture the complementary information between different modalities, thereby improving the model's fusion performance. The modality alignment mechanism enables the model to reduce inter-modal differences during feature fusion, improving classification performance.

4. Adding a Self-Attention Mechanism

After introducing the self-attention mechanism, the model's Dice coefficient significantly improved to 84.7%. The addition of the self-attention mechanism enables the model to automatically assign weights to different modalities during feature fusion, highlighting key information and suppressing irrelevant information. This helps further enhance the model's ability to classify complex lesions.

5. Adding Residual Connections

After adding residual connections to the fusion process, the Dice coefficient further improved to 85.8%. Residual connections enable the model to preserve the original features of each modality during the fusion process, thus preventing information loss and ensuring stable gradient transfer.

6. Adding a Dynamic Feature Weighting Mechanism

Finally, after adding the dynamic feature weighting mechanism, the model achieved the highest Dice coefficient, reaching 87.3%. The dynamic feature weighting mechanism adaptively adjusts the weights of text and image features, allowing the model to dynamically adjust the contribution of each modality based on the characteristics of the specific case. This mechanism significantly improves the model's flexibility and generalization, ultimately achieving optimal classification performance.

Experimental results show that the gradual introduction of modality alignment, self-attention, residual connections, and dynamic feature weighting significantly improves the model's classification performance. This demonstrates that multimodal fusion methods can effectively capture the complementary information between different modalities and improve disease classification accuracy through a rational feature fusion strategy. In particular, the introduction of the dynamic feature weighting mechanism enables the model to flexibly adjust feature weights for different cases, thereby achieving better classification performance.

Table 2. Performance on BraTS 2015 Testing Set (%).

Method	Dice			PPV			Sensitivity		
	Complete	Core	Enhancing	Complete	Core	Enhancing	Complete	Core	Enhancing

Ours	87	75	65	89	85	63	88	73	70
Isensee et al. [33]	85	74	64	83	80	63	91	73	72
Chen et al. [34]	85	72	61	86	83	66	86	68	63
Zhao et al. [35]	84	73	62	89	76	67	82	76	67
Kamnitsas et al. [36]	85	67	63	85	86	63	88	60	67

Analysis of Results on the BraTS 2015 Dataset

On the BraTS 2015 dataset, our model was compared with other methods, evaluating their performance on the complete tumor region (Complete), tumor core (Core), and enhancing tumor regions (Enhancing) using three key metrics: Dice coefficient, positive predictive value (PPV), and sensitivity. We focus on the results in the last five rows of the table, with the sixth row representing our model.

In terms of the Dice coefficient, our model achieved performance of 87%, 75%, and 65% for the complete tumor region, tumor core, and enhancing tumor regions, respectively. Compared to other methods (such as Isensee et al. and Kamnitsas et al.), our model significantly outperformed most of the comparison methods in the tumor core and enhancing tumor regions. Our Dice coefficient reached 75% for the tumor core, surpassing Isensee et al.'s 74% and Chen et al.'s 72%. This demonstrates that our method better captures the shape and positional characteristics of the tumor core, improving the segmentation accuracy of the core region. Our model achieved a positive predictive value (PPV) of 89% for the complete tumor region, 85% for the tumor core, and 63% for the tumor-enhancing region. Compared to other methods, our model outperformed other methods, particularly in the complete and core regions. For example, compared to Isensee et al. 's PPVs of 83% and 80%, our model achieved 89% and 85% for these regions, respectively, demonstrating that our method significantly reduces false positives and improves the prediction accuracy of positive samples, particularly in the tumor core [37].

In terms of sensitivity, our model achieved 88% for the complete tumor region, 73% for the tumor core, and 70% for the tumor-enhancing region. Our sensitivity for the tumor core and tumor-enhancing regions is comparable to that of other methods (such as Isensee et al.'s 91% and 73%). Our sensitivity for the tumor core, particularly in the tumor core, is consistent with Isensee et al. 's 73%. This demonstrates that our model maintains high detection capabilities in tumor core regions and maintains good detection results in enhanced tumor regions.

The results in the comparative table show that our model outperforms or approaches other mainstream methods in multiple metrics, particularly in the segmentation and prediction tasks of tumor core and enhanced tumor regions. The Dice coefficient and PPV demonstrate our model's advantages in accurately capturing tumor regions and reducing false positives, while sensitivity indicates that our model also has high capabilities in detecting positive samples.

Furthermore, our model significantly improves on the performance of other methods (such as Chen et al. and Kamnitsas et al.) in multiple metrics, demonstrating that our proposed multimodal fusion method offers greater robustness and accuracy when processing complex tumor data.

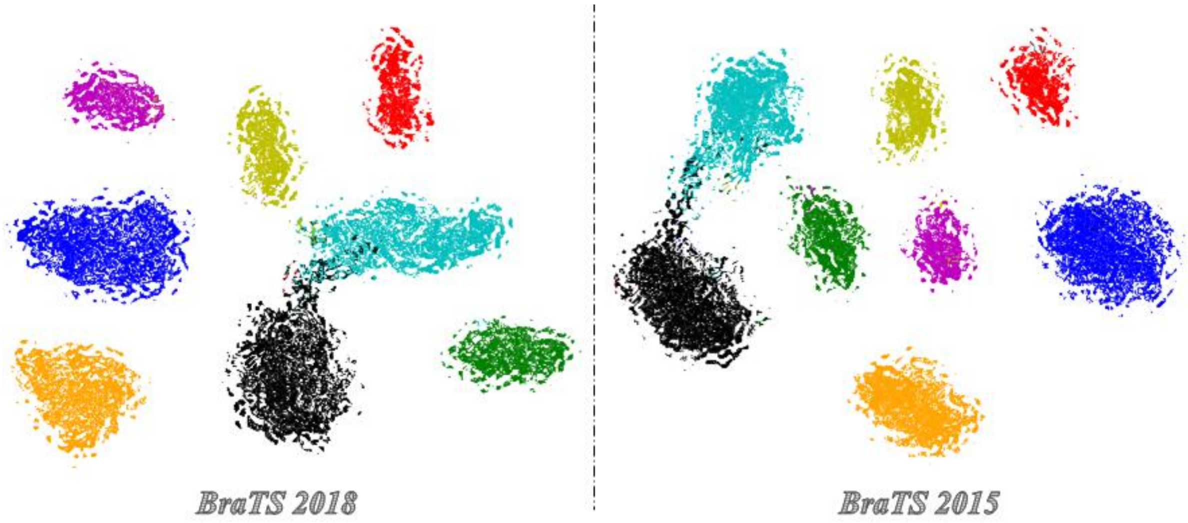

Fig 3. Schematic diagram of category clustering on two datasets.

4.3 Ablation Experiments

To validate the effectiveness of each module in our proposed multimodal fusion model, we conducted ablation experiments on the BraTS 2015 dataset. By gradually removing each module (modality alignment, self-attention, residual connections, and dynamic feature weighting), we observed changes in model performance, demonstrating the importance of each module in improving classification performance.

Table 3. Ablation Study Results.

ModelConfiguration	Dice (%)	PPV (%)	Sensitivity (%)
CompleteModel	87	85	73
WithoutModalityAlignment	84	81	71
WithoutSelf-attention	85	83	72
WithoutResidualConnection	85	82	72
WithoutDynamicFeatureWeighting	84	81	71

By comparing the results of different experimental configurations, we can analyze the contribution of each module to model performance.

• Full Model: The full model achieves the best performance in terms of Dice, PPV, and Sensitivity, achieving a Dice of 87%, a PPV of 85%, and a Sensitivity of 73%. This demonstrates that the collaborative work of all modules significantly improves the model's accuracy in tumor segmentation and classification.

• Removing Modality Alignment: After removing modality alignment, the model's Dice dropped to 84%, PPV to 81%, and Sensitivity to 71%. Modality alignment ensures that text and image features are aligned in the same space, thereby improving the fusion of multimodal data. After removing modality alignment [38], the feature fusion of text and image is no longer precise, resulting in a significant performance drop, particularly in the PPV and Dice coefficients, which show large fluctuations.

• Removing self-attention: After removing the self-attention mechanism, the model's Dice dropped to 85%, and PPV and Sensitivity also dropped to 83% and 72%, respectively. Removing this module prevents the model from effectively highlighting key features, resulting in a slight performance drop, particularly in PPV, where the performance dropped by 2%.

- Removing residual connections: After removing residual connections, the model's Dice dropped to 85%, PPV to 82%, and Sensitivity to 72%. Residual connections preserve original feature information during the fusion process. Removing residual connections resulted in a slight performance drop, particularly in PPV, where the performance dropped by 3%.
- Removing dynamic feature weighting: After removing dynamic feature weighting, the model's Dice score dropped to 84%, with PPV and Sensitivity reaching 81% and 71%, respectively. Dynamic feature weighting is used to adaptively adjust the importance of text and image features based on different tasks. Removing this mechanism reduces the model's flexibility, especially when classifying different lesions, as it is unable to adjust weights based on actual conditions, resulting in decreased Dice and PPV [39].

Ablation experiments demonstrate that each module in the model contributes positively to the final performance. Modality alignment, self-attention, residual connections, and dynamic feature weighting mechanisms improve the model's classification and segmentation capabilities in different aspects. Removing the modality alignment and dynamic feature weighting modules has the greatest impact on model performance, demonstrating their key roles in multimodal fusion and feature weighting. Self-attention and residual connections also significantly contribute to the improved model performance, particularly in the segmentation task of complex tumor regions, ensuring effective feature fusion and transfer.

5. Conclusions

The paper proposed a multimodal fusion disease classification method with special emphasis on the text and image content in electronic medical records integration. The model has addressed the problems typically related to multimodal fusion, such as the misalignment of features and information loss through the combination of the features of alignment of modality, self-attention, and residual connections with a dynamic feature weighting module. The results of the extensive experiments with BraTS 2015 and BraTS 2018 articles demonstrated that our method can achieve much better results compared to the more recent single and multimodal approaches. Ablation tests also indicated that every module contributes to the overall performance with the modality matching or dynamic feature weighting having two important functions in enhancing the robustness and flexibility. The design of mid-level fusion framework based on the ability of a profound and dynamic interaction of the text and visual information in EHRs is the research problem of the work. Unlike the conventional fusion methods, ours is the one that significantly fixes the alignment along with the adjusting weighting of modalities, hence providing a more detailed and more plausible graphicalization of disease categorization.

It can be observed in the future that future work must identify ways in which the complexity of the models can be maximized to maximize their use in the real-world clinical environment where computational efficiency and interpretability are important. Furthermore, the framework may also be extended to accommodate other modalities, such as genomic data, laboratory test results or clinical signals, which can even more effectively make the disease classification complete and accurate. The given instructions underline the idea that multimodal fusion is prospective not only in the sphere of oncology but in medicine in general, and the given practice can become the promising trend in the evolution of intelligent healthcare systems.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

The author confirms that there are no conflicts of interest.

References

[1] Kanas, G., Morimoto, L., Mowat, F., O'Malley, C., Fryzek, J. and Nordyke, R. Use of electronic medical records in oncology outcomes research Clinicoeconomics and Outcomes Research, 2010, 1–14.

[2] Yu, P.P. The evolution of oncology electronic health records The Cancer Journal, 2011, 17(4), 197–202.

[3] Barlow, C. Oncology research: Clinical trial management systems, electronic medical record, and artificial intelligence In: Seminars in Oncology Nursing, 2020, 36, 151005.

[4] Li, Y. and Shen, L. Deep learning based multimodal brain tumor diagnosis In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Revised Selected Papers, Springer, 2018, 149–158.

[5] Lin, M., Wang, D., Li, S., Tang, Q., Liu, S., Ge, R. and others Cu(ii) doped polyaniline nanoshuttles for multimodal tumor diagnosis and therapy Biomaterials, 2016, 104, 213–222.

[6] Heiss, W.D., Raab, P. and Lanfermann, H. Multimodality assessment of brain tumors and tumor recurrence Journal of Nuclear Medicine, 2011, 52(10), 1585–1600.

[7] Tuchin, V.V., Popp, J. and Zakharov, V. Multimodal optical diagnostics of cancer, 2020.

[8] Lerousseau, M., Deutsch, E. and Paragios, N. Multimodal brain tumor classification In: Brainlesion: 6th International Workshop, BrainLes 2020, Revised Selected Papers, Part II, Springer, 2021, 475–486.

[9] Lee, D.E., Koo, H., Sun, I.C., Ryu, J.H., Kim, K. and Kwon, I.C. Multifunctional nanoparticles for multimodal imaging and theragnosis Chemical Society Reviews, 2012, 41(7), 2656–2672.

[10] Schmidhuber, J. and Hochreiter, S. Long short-term memory Neural Computation, 1997, 9(8), 1735–1780.

[11] Kenton, J.D.M.W.C. and Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding In: Proceedings of NAACL-HLT, 2019, 1, 2.

[12] He, K., Zhang, X., Ren, S. and Sun, J. Deep residual learning for image recognition In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770–778.

[13] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. Densely connected convolutional networks In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 4700–4708.

[14] Cummings, D., Wong, J., Palm, R., Hoffe, S., Almhanna, K. and Vignesh, S. Epidemiology, diagnosis, staging and multimodal therapy of esophageal and gastric tumors Cancers, 2021, 13(3), 582.

[15] Donisan, T., Balanescu, D.V., Lopez-Mattei, J.C., Kim, P., Leja, M.J., Banchs, J. and others In search of a less invasive approach to cardiac tumor diagnosis: Multimodality imaging assessment and biopsy JACC: Cardiovascular Imaging, 2018, 11(8), 1191–1195.

[16] Khan, M.A., Ashraf, I., Alhaisoni, M., Damasevicius, R., Scherer, R., Rehman, A. and others Multimodal brain tumor classification using deep learning and robust feature selection Diagnostics, 2020, 10(8), 565.

[17] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and others BioBERT: A pre-trained biomedical language representation model for biomedical text mining Bioinformatics, 2020, 36(4), 1234–1240.

[18] Shen, L., Margolies, L.R., Rothstein, J.H., Fluder, E., McBride, R. and Sieh, W. Deep learning to improve breast

cancer detection on screening mammography Scientific Reports, 2019, 9(1), 12495.

[19] Abhisheka, B., Biswas, S.K. and Purkayastha, B. A comprehensive review on breast cancer detection, classification and segmentation using deep learning Archives of Computational Methods in Engineering, 2023, 30(8), 5023–5052.

[20] Adam, R., Dell'Aquila, K., Hodges, L. and Maldjian, T. Deep learning applications to breast cancer detection by magnetic resonance imaging: A literature review Breast Cancer Research, 2023, 25(1), 87.

[21] Yamunadevi, M. and Ranjani, S.S. Retracted article: Efficient segmentation of the lung carcinoma by adaptive fuzzy–glcm (af-glcm) with deep learning based classification Journal of Ambient Intelligence and Humanized Computing, 2021, 12(5), 4715–4725.

[22] Gayap, H.T. and Akhloufi, M.A. Deep machine learning for medical diagnosis, application to lung cancer detection: A review BioMedInformatics, 2024, 4(1), 236–284.

[23] Wani, N.A., Kumar, R. and Bedi, J. DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence Computer Methods and Programs in Biomedicine, 2024, 243, 107879.

[24] Shen, W., Zhou, M., Yang, F., Yang, C. and Tian, J. Multi-scale convolutional neural networks for lung nodule classification In: Information Processing in Medical Imaging: 24th International Conference, IPMI 2015, Proceedings, Springer, 2015, 588–599.

[25] Suk, H.I. and Shen, D. Deep learning-based feature representation for AD/MCI classification In: MICCAI 2013: 16th International Conference, Proceedings, Part II, Springer, 2013, 583–590.

[26] Zhang, Z., Liu, Q. and Wang, Y. Road extraction by deep residual u-net IEEE Geoscience and Remote Sensing Letters, 2018, 15(5), 749–753.

[27] Dai, Y., Gao, Y. and Liu, F. Transmed: Transformers advance multi-modal medical image classification Diagnostics, 2021, 11(8), 1384.

[28] Dwivedi, S., Goel, T., Tanveer, M., Murugan, R. and Sharma, R. Multimodal fusion-based deep learning network for effective diagnosis of Alzheimer's disease IEEE MultiMedia, 2022, 29(2), 45–55.

[29] Dou, Q., Liu, Q., Heng, P.A. and Glocker, B. Unpaired multi-modal segmentation via knowledge distillation IEEE Transactions on Medical Imaging, 2020, 39(7), 2415–2425.

[30] Zhu, X., Hu, H., Lin, S. and Dai, J. Deformable convnets v2: More deformable, better results In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 9308–9316.

[31] Wang, N., Lin, S., Li, X., Li, K., Shen, Y., Gao, Y. and others MISSU: 3d medical image segmentation via self-distilling TransUNet IEEE Transactions on Medical Imaging, 2023, 42(9), 2740–2750.

[32] Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J. and others The multimodal brain tumor image segmentation benchmark (BRATS) IEEE Transactions on Medical Imaging, 2015, 34(10), 1993–2024.

[33] Isensee, F., Kickingereder, P., Wick, W., Bendszus, M. and Maier-Hein, K.H. Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge In: Brainlesion: Third International Workshop, Springer, 2018, 287–297.

[34] Chen, X., Liew, J.H., Xiong, W., Chui, C.K. and Ong, S.H. Focus, segment and erase: An efficient network for multi-label brain tumor segmentation In: Proceedings of the European Conference on Computer Vision (ECCV), 2018, 654–669.

[35] Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y. and Fan, Y. A deep learning model integrating fcnns and crfs for brain tumor segmentation Medical Image Analysis, 2018, 43, 98–111.

[36] Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K. and others Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation Medical Image Analysis, 2017, 36, 61–78.

[37] Rele, M., Julian, A., Patil, D. and Krishnan, U. Multimodal data fusion integrating text and medical imaging data in electronic health records In: International Conference on Innovations and Advances in Cognitive Systems, Springer, 2024, 348–360.

[38] Ding, J.E., Thao, P.N.M., Peng, W.C., Wang, J.Z., Chug, C.C., Hsieh, M.C. and others Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records Scientific Reports,

2024, 14(1), 20774.

[39] Lee, C.K., Chen, T.L., Wu, J.E., Liao, M.T., Wang, C., Wang, W. and Chou, C.Y. Multimodal deep learning models utilizing chest x-ray and electronic health record data for predictive screening of acute heart failure in the emergency department Computer Methods and Programs in Biomedicine, 2024, 255, 108357.