



# Transformer-Driven Feature Fusion for Robust Diagnosis of Lung Cancer Brain Metastasis Under Missing-Modality Scenarios

Yue Ding<sup>1</sup>, Yunqi Ma<sup>1</sup>, Kuo Jing<sup>1</sup>, Zhansong Shang<sup>1</sup>, Feiyang Gao<sup>1</sup>, Zhengwei Cui<sup>1</sup>, Linyan Xue<sup>1,2,3</sup>, Shuang Liu<sup>1,2,3\*</sup>

<sup>1</sup> College of Quality and Technical Supervision, Hebei University, 071002 Baoding, China

<sup>2</sup> Hebei Technology Innovation Center for Lightweight of New Energy Vehicle Power System, 071002 Baoding, China

<sup>3</sup> National & Local Joint Engineering Research Center of Metrology Instrument and System, Hebei University, 071002 Baoding, China

\* Correspondence: Shuang Liu (wsg6813@163.com)

Received: 11-20-2025

Revised: 01-15-2026

Accepted: 01-27-2026

**Citation:** Y. Ding, Y. Q. Ma, K. Jing, Z. S. Shang, F. Y. Gao, Z. W. Cui, L. Y. Xue, and S. Liu, “Transformer-driven feature fusion for robust diagnosis of lung cancer brain metastasis under missing-modality scenarios,” *Acadlore Trans. Mach. Learn.*, vol. 5, no. 1, pp. 32–43, 2026. <https://doi.org/10.56578/ataiml050104>.



© 2026 by the author(s). Licensee Acadlore Publishing Services Limited, Hong Kong. This article can be downloaded for free, and reused and quoted with a citation of the original published version, under the CC BY 4.0 license.

**Abstract:** Accurate diagnosis of lung cancer brain metastasis is often hindered by incomplete magnetic resonance imaging (MRI) modalities, resulting in suboptimal utilization of complementary radiological information. To address the challenge of ineffective feature integration in missing-modality scenarios, a Transformer-based multi-modal feature fusion framework, referred to as Missing Modality Transformer (MMT), was introduced. In this study, multi-modal MRI data from 279 individuals diagnosed with lung cancer brain metastasis, including both small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), were acquired and processed through a standardized radiomics pipeline encompassing feature extraction, feature selection, and controlled data augmentation. The proposed MMT framework was trained and evaluated under various single-modality and combined-modality configurations to assess its robustness to modality absence. A maximum diagnostic accuracy of 0.905 was achieved under single-modality missing conditions, exceeding the performance of the full-modality baseline by 0.017. Interpretability was further strengthened through systematic analysis of loss-function hyperparameters and quantitative assessments of modality-specific importance. The experimental findings collectively indicate that the MMT framework provides a reliable and clinically meaningful solution for diagnostic environments in which imaging acquisition is limited by patient conditions, equipment availability, or time constraints. These results highlight the potential of Transformer-based radiomics fusion to advance computational neuro-oncology by improving diagnostic performance, enhancing robustness to real-world imaging variability, and offering transparent interpretability that aligns with clinical decision-support requirements.

**Keywords:** Lung cancer brain metastasis; Radiomics; Multi-modal; Modality missing

## 1 Introduction

Lung cancer is one of the most prevalent malignancies globally and a leading cause of cancer-related mortality [1, 2]. This malignant tumor originates from cancerous cells in lung tissue. Pathologically and therapeutically, it is classified into two main categories: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The latter encompasses various histological subtypes, including adenocarcinoma and squamous cell carcinoma, while the former constitutes the remaining category [3]. SCLC accounts for approximately 15% of lung cancer cases, whereas NSCLC makes up 85% [4–6]. Over half of newly diagnosed lung cancer patients present with advanced or metastatic disease [7], with 10%–26% exhibiting brain metastases at diagnosis and another 30% developing them during disease progression [8–10]. Due to the absence of distinct early symptoms, lung cancer is often detected in advanced stages when metastasis has already spread, making treatment and management particularly challenging.

With the rapid advancement of artificial intelligence, machine learning methods have become crucial in diagnosing brain metastases from lung cancer. Radiomics is widely employed to extract high-throughput features from medical images, significantly enhancing the interpretability of image-assisted diagnosis [11–13]. For instance,

Xu et al. [14] utilized radiomics to extract features from preoperative chest CT scans to predict anaplastic lymphoma kinase (ALK) gene molecular signatures. However, while traditional machine learning offers strong interpretability, it often struggles with complex nonlinear problems inherent in multi-modal MRI data. Deep learning, a subset of machine learning, excels at processing such intricate relationships [15, 16]. For example, Guo et al. [17] developed a multi-modal MRI decision fusion network for glioma classification, and Li et al. [18] proposed a Hybrid Graph Convolutional Network (HGCN) for cancer survival prediction. Despite these breakthroughs, existing methods often lack robust diagnostic capabilities when specific modalities are missing.

A dataset of lung cancer brain metastasis patients from the Affiliated Hospital of Hebei University was utilized in this study. Using radiomics, high-throughput feature information was extracted from MRI images [19]. Based on the Transformer architecture, a network was designed for missing-modality scenarios. It employs a modal interaction self-attention mechanism to learn inter-modal correlations and generate multi-level feature representations for missing modalities. This approach enables effective reconstruction and maintains diagnostic accuracy even when modalities are incomplete.

## 2 Methodology

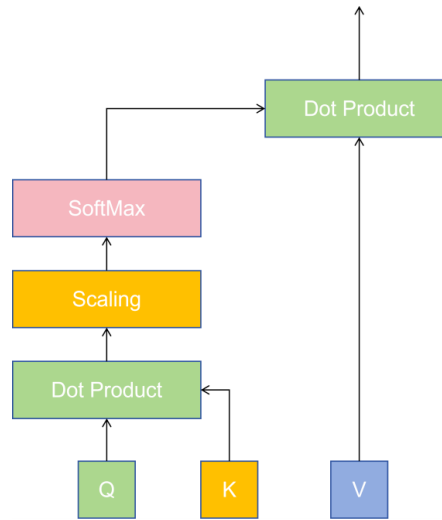
Aiming to address the critical issue of ineffective feature fusion in the diagnosis of brain metastases from lung cancer when MRI image modalities are missing, a Transformer-based feature fusion network called Missing Modality Transformer (MMT) was proposed in this study to achieve effective feature fusion.

### 2.1 Multimodal Fusion Network—Variational Autoencoder Model

Multimodal Fusion Network—Variational Autoencoder (MFN-VAE) is a Variational Autoencoder (VAE) model designed to integrate multi-modal features. It processes multi-modal data by mapping them into a unified latent space. By learning correlations between different modalities, MFN-VAE achieves deep feature integration. The latent representations learned by MFN-VAE serve as robust features for downstream tasks, significantly improving performance. The model consists of three components: an encoder, a decoder, and a predictor. In the lung cancer brain metastasis prediction task, the dataset includes three modalities: T1-weighted imaging (T1WI), fluid-attenuated inversion recovery (FLAIR), and diffusion-weighted imaging (DWI). Each modality is processed through the VAE to learn latent representations, which are then fed into the predictor to forecast brain metastasis outcomes.

### 2.2 Channel Attention Module

Self-attention mechanisms are a cornerstone of Transformer models and a key factor behind their outstanding performance. These mechanisms determine element importance by comparing the similarity between positions in a sequence. Specifically, through self-attention calculations, each position generates an attention weight vector reflecting its relative importance. This adaptive allocation of attention enables the model to process sequence data efficiently. The theory revolves around three core components: Query (Q), Key (K), and Value (V). In Transformer models, input sequences are encoded and mapped to these three embeddings. A classic implementation is the Scaled Dot-Product Attention (SDPA) mechanism, as illustrated in Figure 1.



**Figure 1.** Scaled Dot-Product Attention (SDPA) mechanism

The SDPA mechanism involves performing dot-product operations on  $Q$  and  $K$  elements, followed by scaling the result. The scaling factor is a constant introduced to prevent the dot-product result from becoming excessively

large, ensuring the values remain within a reasonable range. Subsequently, the scaled dot-product result undergoes SoftMax normalization to obtain attention weights. The SoftMax function maps scores to a range between 0 and 1, representing the attention importance of each position relative to others. Finally, the attention weights are used to compute a weighted sum of the  $V$  vectors to generate the context representation for each position. This process effectively integrates information from other positions into the current position's representation. The specific calculation formula is as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where,  $d_k$  denotes the dimension of the  $K$  vector.

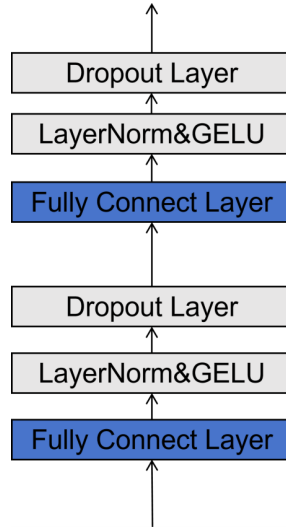
In practice, matrix multiplication enables the simultaneous calculation of similarities between multiple  $Q$  and  $K$  pairs, facilitating efficient batch processing. The SDPA mechanism is favored for its computational simplicity and efficiency, allowing it to effectively capture both local and global dependencies within sequences, which explains its widespread adoption in Transformer models.

### 2.3 Modal Interaction Self-attention

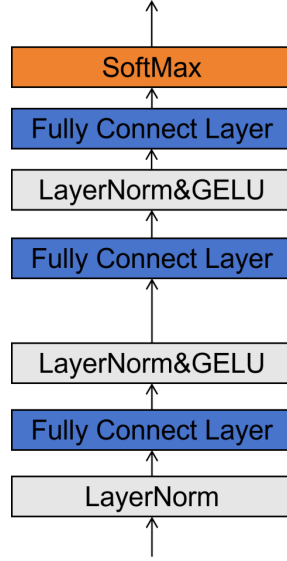
Modal interaction self-attention is the core of MMT's design, enhancing the interaction between different modalities. This enables MMT to maintain high performance even when certain modalities are missing by leveraging existing modalities to generate feature information related to the missing ones. Taking MMT-2M as an example, the four inputs to the modal interaction self-attention mechanism consist of Modal 1 features, Modal 2 features, the Missing Modality Token, and the Classification Token. These inputs first undergo standardization through a Layer Normalization (LayerNorm) layer. Subsequently, the features of Modal 1, Modal 2, and the Missing Modality Token are mapped to a larger feature space and decomposed into  $Q$ ,  $K$ , and  $V$  components. The Classification Token is similarly mapped, but only its  $Q$  component is retained.

Regarding the strategy for cross-modal interaction, the proposed design principle focuses on enhancing interactions between the tokens of Modal 1, Modal 2, and the missing modality, while the Classification Token only interacts with the missing modality. Technically,  $K$  and  $V$  are responsible for fusing with other modalities, whereas  $Q$  solely handles forward propagation. Specifically, the  $K$  vectors from each modality undergo a Hadamard product operation and are summed with their corresponding  $Q$  vectors. The Hadamard product is an element-wise multiplication operation that yields a result matrix of matching dimensions. The  $V$  vectors of each modality are then multiplied with their corresponding  $K$  vectors. Finally, the  $Q$  vector associated with the Classification Token is combined with the  $K$  and  $V$  vectors of the missing modality for propagation.

Through this modality interaction process, the  $Q$  vectors of each modality integrate feature information from all modalities. These four  $Q$  vectors then pass through a LayerNorm layer and a Feedforward Neural Network (FNN) with shared weights. To prevent excessive interference from modal interactions that could obscure original features, these  $Q$  vectors are ultimately combined with the pre-differentiated vectors for output. After passing through two modal interaction attention layers, the four tensors are concatenated and fed into the Multi-Layer Perceptron (MLP). Figure 2 and Figure 3 illustrate the architectures of the FNN and the MLP.



**Figure 2.** Feedforward Neural Network (FNN) structure



**Figure 3.** Multi-Layer Perceptron (MLP) network structure

An FNN consists of two fully connected layers, each followed by a LayerNorm layer and a Dropout layer. As a widely adopted regularization technique, the Dropout layer randomly discards a portion of neurons during training to mitigate overfitting risks. The MLP consists of three fully connected layers. The first layer is preceded by a LayerNorm layer, followed by another LayerNorm layer and a Gaussian Error Linear Unit (GELU) activation function. The second fully connected layer also includes a LayerNorm layer and a GELU activation function. The third fully connected layer is terminated by a SoftMax layer.

## 2.4 Loss Function

The loss function of MMT consists of two components: reconstruction loss and classification loss. The reconstruction loss employs the Mean Absolute Error (MAE) loss. This function calculates the absolute difference between reconstructed and original features, averaging these differences to determine the final loss value. Compared to Mean Squared Error (MSE) loss, MAE demonstrates greater robustness by being less sensitive to outliers. Minimizing MAE helps the model generate reconstructed features that closely resemble the original data. The formula is as follows:

$$L_{\text{mae}} = \frac{1}{m} \sum_{i=1}^m |x_i - \hat{x}_i| \quad (2)$$

where,  $x_i$  and  $\hat{x}_i$  represent the radiomics feature and reconstructed feature of the  $i$ -th sample, respectively, with  $m$  being the number of samples.

The classification loss employs the cross entropy loss function, which quantifies the discrepancy between predicted probabilities and actual labels. When predictions perfectly match the labels, the loss is zero; otherwise, it increases. Minimizing this loss aligns the model's predictions with the true labels. The formula is:

$$L_{CE} = - \sum [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (3)$$

where,  $y_i$  and  $\hat{y}_i$  represent the true label and predicted label of the  $i$ -th sample.

The reconstruction loss and classification loss are then weighted and summed. Due to structural differences between the MMT-2M and MMT-1M models, their loss functions vary slightly, defined by Eq. (4) and Eq. (5):

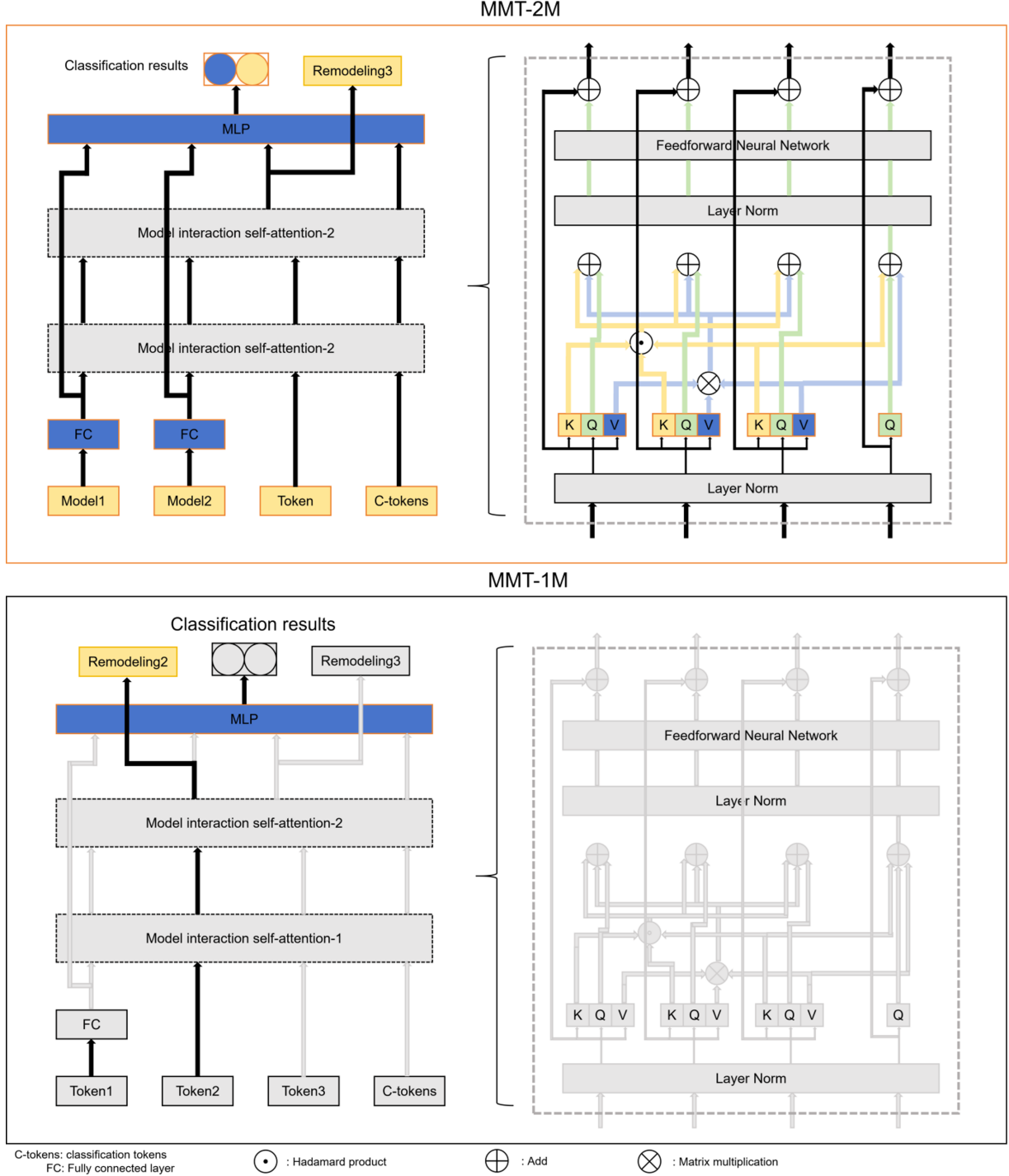
$$L_{MMT-2M} = \alpha L_{\text{mae}} + \beta L_{ce} \quad (4)$$

$$L_{MMT-1M} = \alpha L_{\text{mae}} + \beta L_{ce1} + \gamma L_{ce2} \quad (5)$$

where,  $\alpha$ ,  $\beta$ , and  $\gamma$  are the respective weight parameters.

## 2.5 Missing Modality Transformer Model

The MMT model is a Transformer-based classification network designed for missing-modality scenarios. As shown in Figure 4, two MMT architectures were developed: MMT-2M for single-modality missing and MMT-1M for dual-modality missing. Its core principle is to enhance the model's inter-modal interaction capabilities, enabling the reconstruction of missing modalities.



**Figure 4.** Missing Modality Transformer (MMT) network structure

Taking MMT-2M as an example, it features four inputs: Modal 1 features, Modal 2 features, a Missing Modality Token, and a Classification Token. The Classification Token is a trainable neural network weight parameter updated through backpropagation, facilitating feature reconstruction and classification. The features from Modal 1 and Modal 2 are linearly mapped to a new space via a fully connected layer, then fed into the modal interaction self-attention

mechanism. This layer enhances the network’s feature learning capacity. Since tokens are weight parameters following a normal distribution and contain no prior feature information, they are directly input into the self-attention mechanism.

After two modal interactions, the missing modal features regenerated by the tokens are output. These are then input into the fully connected neural network alongside Modal 1, Modal 2, and the features generated by the Classification Token to output classification results. At its core, MMT is a classification model that utilizes modal interaction self-attention and reconstruction loss to generate missing modal features via inter-modal connections. This capability not only allows MMT to reconstruct missing modalities but, more importantly, enables it to learn deeper-level feature representations. The reconstruction process serves solely as a means to enhance the network’s classification performance rather than focusing on reconstruction quality itself.

### 3 Experimental Research

#### 3.1 Experiment Environment and Workflow

This study constructed a multi-modal MRI dataset encompassing both SCLC and NSCLC brain metastases, which was trained and tested using the MMT network model. The MMT experiments were conducted on an Ubuntu workstation equipped with an NVIDIA GeForce RTX 2080 Ti GPU and Intel® Xeon® Gold 6240 CPU. Training parameters were configured as follows: batch size set to 32, loss weight range between 0.1 and 1, and learning rate range between  $1 \times 10^{-6}$  and  $1 \times 10^{-5}$ . The AdamW optimizer was employed during training. As the MMT network is built on the Transformer architecture, it required substantial memory and computational resources due to its large parameter scale. A total of 200 training epochs were completed with hyperparameter tuning strategies to ensure model convergence and prevent underfitting.

#### 3.2 Data Collection

The dataset, sourced from the Affiliated Hospital of Hebei University, comprises multi-modal MRI images of 279 lung cancer patients with brain metastases. It includes six subtypes: SCLC (100 cases), adenocarcinoma (153 cases), squamous cell carcinoma (17 cases), adenosquamous carcinoma (2 cases), large cell lung cancer (6 cases), and other brain metastases (1 case). Each patient received T1WI, FLAIR, and DWI scans. In clinical practice, specialists classify these subtypes into SCLC (100 cases) and NSCLC (179 cases), as shown in Table 1. To ensure accurate tumor localization, MRI images were annotated by a senior clinical expert, enabling comprehensive analysis of multi-modal data.

**Table 1.** Detection results of ResNet50 with different attention modules

Disease Type	Number of Patients	Age (Years)	Number of Lesions	Gender
SCLC	100	62.4 $\pm$ 9.7	439	72/28
NSCLC	179	60.6 $\pm$ 9.5	679	109/180

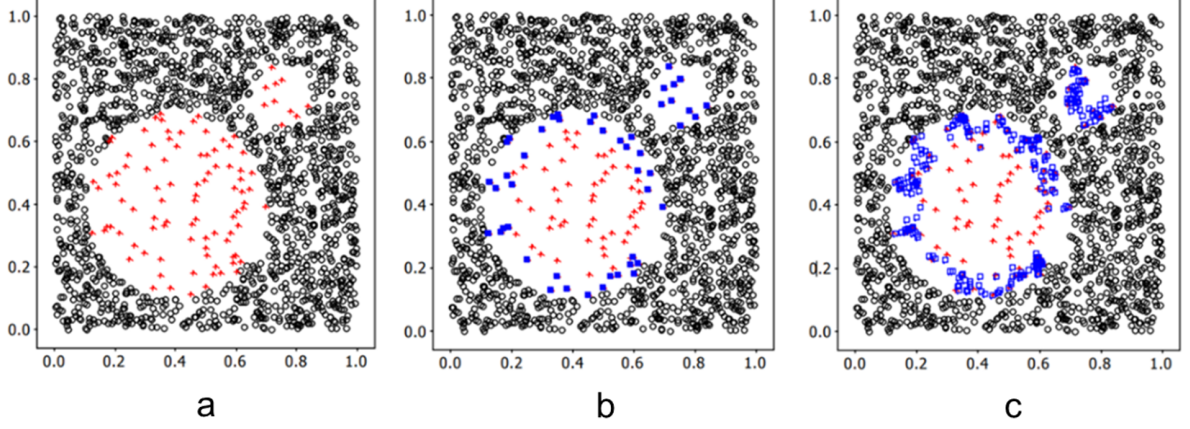
#### 3.3 Radiomics Feature Extraction

Medical imaging contains not only visible structures but also latent information. This study employed radiomics technology to extract and analyze such hidden data. The Python-based open-source package PyRadiomics [20] was utilized to extract morphological, textural, and statistical features from the images. In MRI image processing, feature extraction requires preprocessing filters. Using filters prior to extraction significantly improves the accuracy and robustness of subsequent analyses. The “original” filter refers to unprocessed MRI images retaining complete original information. Wavelet filters are used for noise removal and edge detection. The Laplacian of Gaussian (LoG) filter combines Gaussian filtering with Laplacian operations to enhance edges. Other filters used include square, square root, logarithmic, exponential, gradient, and Local Binary Pattern (LBP), each serving to highlight specific intensity regions or texture details.

#### 3.4 Data Augmentation

Data augmentation expands training datasets by generating new samples through transformations of existing data. To address data imbalance, the Borderline Synthetic Minority Oversampling Technique (Borderline SMOTE) method was employed for oversampling [21]. As an enhanced version of traditional SMOTE, Borderline SMOTE focuses on creating synthetic samples near category boundaries while avoiding the interpolation of noisy points. This approach effectively prevents the generation of inaccurate synthetic samples, thereby minimizing adverse impacts on model performance. Figure 5 shows the distribution of data before and after using Borderline SMOTE for data augmentation.





**Figure 5.** Distribution of data before and after using Borderline Synthetic Minority Oversampling Technique (SMOTE) for data augmentation: (a) Distribution of the original dataset; (b) Borderline minority samples (solid squares); (c) Borderline synthetic minority samples (hollow squares).

## 4 Results

### 4.1 Evaluation Indicators

Experimental results were evaluated using five metrics: accuracy, Area Under the Curve (AUC), precision, sensitivity, and specificity. Accuracy measures the proportion of correct classifications. AUC quantifies the model’s classification performance across different thresholds. Precision indicates the accuracy of positive predictions. Sensitivity reflects the model’s ability to identify true positives, while specificity measures the capacity to distinguish true negatives. The formulas are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (8)$$

where,  $TP$  (True Positives) refers to correctly predicted positive samples,  $TN$  (True Negatives) refers to correctly predicted negative samples,  $FP$  (False Positives) refers to incorrectly predicted positive samples, and  $FN$  (False Negatives) refers to incorrectly predicted negative samples.

### 4.2 Experimental Result

To validate MMT’s superior performance under missing modality conditions, a comparative experiment with MFN-VAE was conducted. The results are presented in Table 2 and Figure 6.

Table 2 results indicate that MMT achieves an accuracy of 0.863–0.877 in single-modality scenarios, surpassing MFN-VAE’s accuracy of 0.838–0.852 in the same setting. This performance level is comparable to MFN-VAE’s results across dual-modal conditions (0.849–0.880). When operating in dual-modal environments (missing one modality), MMT demonstrates an accuracy range of 0.897 to 0.905, outperforming MFN-VAE’s accuracy of 0.888 in full-modal scenarios. These findings demonstrate MMT’s capability to effectively leverage multi-modal feature information. By integrating existing modalities and learning inter-modal relationships through its self-attention mechanism, MMT generates contextual features for missing modalities. This approach enables MMT to achieve performance parity with full-modal models even when data is incomplete.

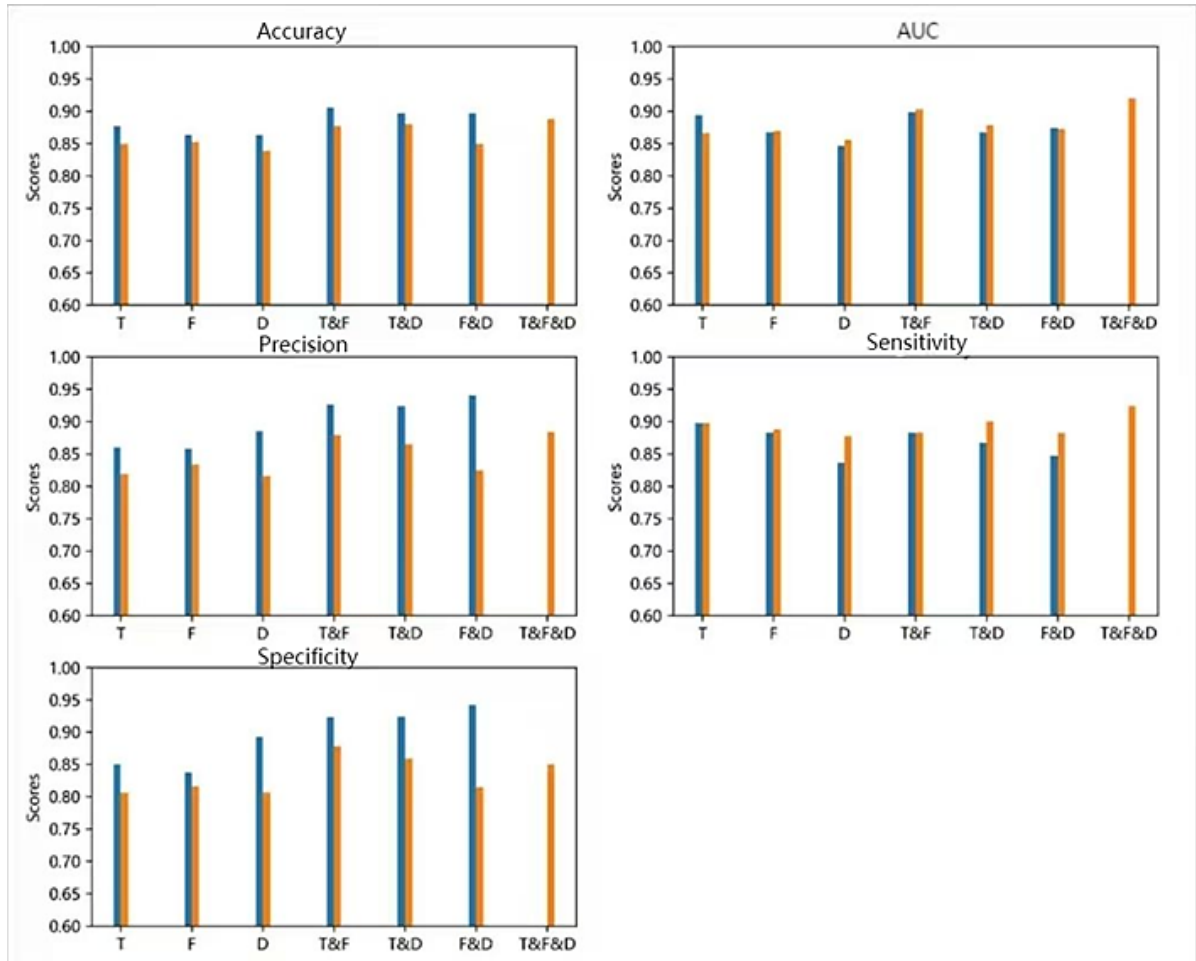
The Receiver Operating Characteristic (ROC) performance of the best-performing MFN-VAE and MMT models is illustrated in Figure 7.

The loss function is a crucial component of MMT optimization. As shown in Figure 8, while the overall MMT loss gradually decreases during epoch iterations, the validation loss consistently remains higher than the training loss, indicating a certain degree of overfitting. The model demonstrates good convergence of reconstruction loss during training, but shows significant fluctuations in validation loss with almost no convergence trend.

**Table 2.** Comparison between Missing Modality Transformer (MMT) under missing-modality scenarios and Multimodal Fusion Network—Variational Autoencoder (MFN-VAE) under different modality combination settings

Model	Mode			Accuracy	AUC	Precision	Sensitivity	Specificity
	T	F	D					
MMT	✓			$0.877 \pm 0.033$	$0.894 \pm 0.026$	$0.860 \pm 0.046$	$0.897 \pm 0.047$	$0.850 \pm 0.068$
		✓		$0.863 \pm 0.047$	$0.867 \pm 0.077$	$0.858 \pm 0.075$	$0.883 \pm 0.056$	$0.837 \pm 0.133$
			✓	$0.863 \pm 0.037$	$0.846 \pm 0.070$	$0.885 \pm 0.038$	$0.836 \pm 0.044$	$0.892 \pm 0.033$
	✓	✓		$0.905 \pm 0.037$	$0.899 \pm 0.016$	$0.926 \pm 0.043$	$0.883 \pm 0.026$	$0.923 \pm 0.057$
	✓		✓	$0.897 \pm 0.032$	$0.867 \pm 0.048$	$0.924 \pm 0.029$	$0.867 \pm 0.022$	$0.924 \pm 0.048$
		✓	✓	$0.897 \pm 0.029$	$0.874 \pm 0.089$	$0.940 \pm 0.054$	$0.847 \pm 0.016$	$0.942 \pm 0.055$
MFN-VAE	✓			$0.849 \pm 0.044$	$0.866 \pm 0.075$	$0.819 \pm 0.076$	$0.897 \pm 0.038$	$0.806 \pm 0.064$
		✓		$0.852 \pm 0.040$	$0.869 \pm 0.063$	$0.833 \pm 0.060$	$0.888 \pm 0.046$	$0.816 \pm 0.085$
			✓	$0.838 \pm 0.040$	$0.855 \pm 0.064$	$0.816 \pm 0.076$	$0.877 \pm 0.063$	$0.806 \pm 0.068$
	✓	✓		$0.877 \pm 0.031$	$0.902 \pm 0.049$	$0.879 \pm 0.066$	$0.883 \pm 0.038$	$0.878 \pm 0.069$
	✓		✓	$0.880 \pm 0.032$	$0.879 \pm 0.051$	$0.865 \pm 0.051$	$0.900 \pm 0.022$	$0.859 \pm 0.050$
		✓	✓	$0.849 \pm 0.022$	$0.872 \pm 0.037$	$0.825 \pm 0.049$	$0.882 \pm 0.031$	$0.815 \pm 0.031$
	✓	✓	✓	$0.888 \pm 0.026$	$0.920 \pm 0.026$	$0.884 \pm 0.038$	$0.924 \pm 0.022$	$0.850 \pm 0.044$

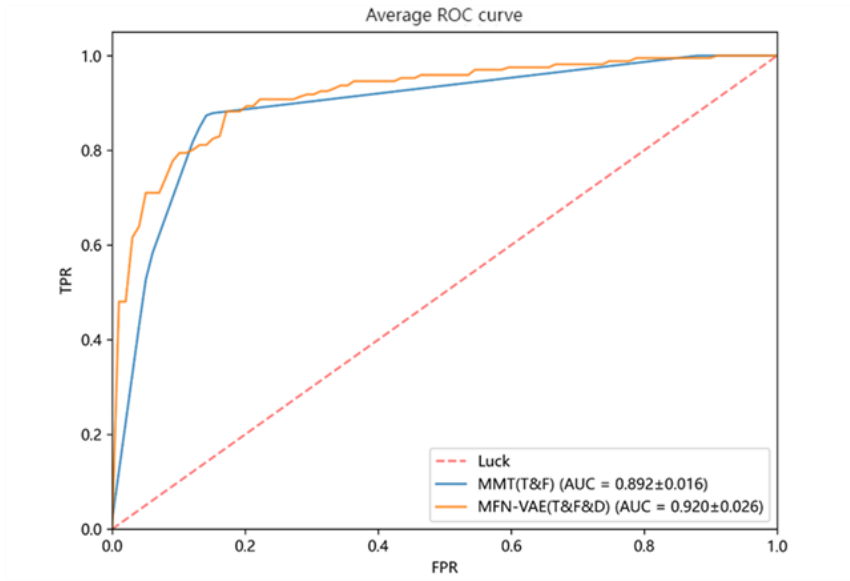
Note: Area Under the Curve (AUC)



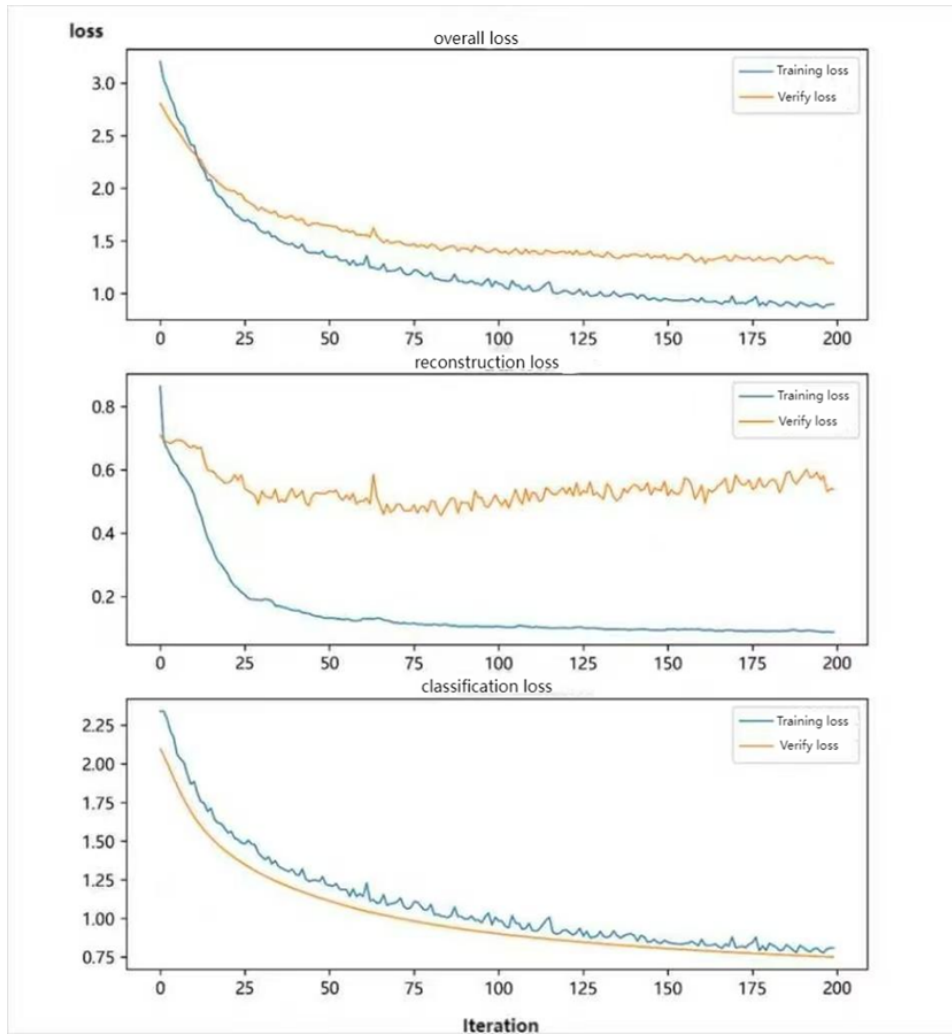
**Figure 6.** Visual comparison of Missing Modality Transformer (MMT) under missing-modality conditions and Multimodal Fusion Network—Variational Autoencoder (MFN-VAE) across different modality combinations

In contrast, the classification loss performs well with only slight overfitting. This discrepancy may stem from MMT directly utilizing the output of the modal interaction self-attention mechanism as the reconstructed feature. This architectural choice prioritizes classification performance over reconstruction fidelity, which explains why the reconstruction loss on the validation set fails to converge.





**Figure 7.** Receiver Operating Characteristic (ROC) curves of the best-performing model weights for Multimodal Fusion Network—Variational Autoencoder (MFN-VAE) and Missing Modality Transformer (MMT)



**Figure 8.** Loss function curve for the optimal modality combination in Missing Modality Transformer (MMT)

Based on the training weights of models under multi-modal combinations, the optimal combination of T1WI and FLAIR was selected, as shown in Table 3. The MMT loss function is a weighted sum of classification loss and reconstruction loss. In this study, the loss function hyperparameters from the study by Hutter et al. [22] were used for evaluation. Overall, the highest score was assigned to the classification loss weight, followed by the reconstruction loss weights for T1WI and DWI. For the FLAIR modality, the reconstruction loss weight consistently scored 0.1, significantly lower than other parameters. When using a single modality, only FLAIR showed a deviation from the overall results, with its classification weight score being only 0.117. Meanwhile, the feature reconstruction scores for both T1WI and DWI exceeded the classification loss score. Overall, it was found that in MMT, the classification loss, T1WI reconstruction loss, and DWI reconstruction loss carry relatively greater weight, while the FLAIR loss holds a secondary position. Therefore, during model optimization, the optimization of these three loss components was prioritized.

**Table 3.** Importance scores of classification and reconstruction loss weights across modality combinations

Model	Mode			Classification Loss	Reconstruction Loss		
	T	F	D		T	F	D
MMT-1M	✓			0.351	—	0.119	0.248
		✓		0.117	0.216	—	0.160
			✓	0.298	0.298	0.126	—
MMT-2M	✓	✓		0.262	—	—	0.378
	✓		✓	0.341	—	0.138	—
		✓	✓	0.292	0.303	—	—
Average				0.277	0.272	0.126	0.262

## 5 Conclusions

This study proposed an MMT network to address the common issue of missing modalities in medical diagnosis. Through comparative validation with the full-modality MFN-VAE network, MMT demonstrates superior performance. MMT's mechanisms were analyzed via loss function visualization and importance assessment, enhancing interpretability. MMT achieves a maximum accuracy of 0.905 with a single missing modality, surpassing the full-modality MFN-VAE by 0.017. This provides a reliable solution for clinical diagnosis in scenarios where data modalities are incomplete.

## Author Contributions

Conceptualization, Y.D., Y.Q.M., and S.L.; methodology, Y.D. and Y.Q.M.; software, Y.D. and Y.Q.M.; validation, K.J., Z.S.S., F.Y.G., and Z.W.C.; formal analysis, Y.D. and Y.Q.M.; investigation, Y.D., Y.Q.M., K.J., and Z.S.S.; resources, S.L. and L.Y.X.; data curation, Y.D., Y.Q.M., F.Y.G., and Z.W.C.; writing—original draft preparation, Y.D. and Y.Q.M.; writing—review and editing, S.L., L.Y.X., and Y.D.; visualization, Y.D. and Y.Q.M.; supervision, S.L. and L.Y.X.; project administration, S.L. and L.Y.X.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

## Data Availability

The data (multi-modal MRI images and extracted radiomics features) supporting our research results are under privacy or ethical restrictions. The data used to support the research findings were sourced from the Affiliated Hospital of Hebei University and are available from the corresponding author upon request for researchers who meet the criteria for accessing confidential medical data.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

- [1] American Cancer Society, *Global Cancer Facts & Figures 4th Edition*. Atlanta, American Cancer Society, 2018. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/global-cancer-facts-and-figures/global-cancer-facts-and-figures-4th-edition.pdf>
- [2] Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018. <https://doi.org/10.3322/caac.21492>

- [3] S. Solmaz and C. I. Gulsever, "Atypical non-enhancing brain metastases from ALK-positive non-small cell lung carcinoma," *Neurocirugía (Engl. Ed.)*, vol. 2026, p. 500758, 2026. <https://doi.org/10.1016/j.neucie.2026.500758>
- [4] General Office of National Health Commission of the People's Republic of China, "Clinical practice guideline for primary lung cancer (2022 version)," *Med. J. Peking Union Med. Coll. Hosp.*, vol. 13, no. 4, pp. 549–570, 2022. <https://doi.org/10.12290/xhyxzz.2022-0352>
- [5] M. C. Rudin, E. Brambilla, C. Faivre-Finn, and J. Sage, "Small-cell lung cancer," *Nat. Rev. Dis. Primers*, vol. 7, no. 1, p. 3, 2021. <https://doi.org/10.1038/s41572-020-00235-0>
- [6] P. Goldstraw, D. Ball, J. R. Jett, T. Le Chevalier, E. Lim, A. G. Nicholson, and F. A. Shepherd, "Non-small-cell lung cancer," *Lancet*, vol. 378, no. 9804, pp. 1727–1740, 2011. [https://doi.org/10.1016/s0140-6736\(10\)62101-0](https://doi.org/10.1016/s0140-6736(10)62101-0)
- [7] National Cancer Institute, "SEER cancer stat facts: Lung and bronchus cancer," 2023. <https://seer.cancer.gov/statfacts/html/lungb.html>
- [8] N. Duma, R. Santana-Davila, and J. R. Molina, "Non-small cell lung cancer: Epidemiology, screening, diagnosis, and treatment," *Mayo Clin. Proc.*, vol. 94, no. 8, pp. 1623–1640, 2019. <https://doi.org/10.1016/j.mayocp.2019.01.013>
- [9] J. S. Barnholtz-Sloan, A. E. Sloan, F. G. Davis, F. D. Vigneau, P. Lai, and R. E. Sawaya, "Incidence proportions of brain metastases in patients diagnosed (1973 to 2001) in the Metropolitan Detroit Cancer Surveillance System," *J. Clin. Oncol.*, vol. 22, no. 14, pp. 2865–2872, 2004. <https://doi.org/10.1200/jco.2004.12.149>
- [10] S. N. Waqar, P. P. Samson, C. G. Robinson, J. Bradley, S. Devarakonda, L. Du, R. Govindan, F. Gao, V. Puri, and D. Morgensztern, "Non-small-cell lung cancer with brain metastasis at presentation," *Clin. Lung Cancer*, vol. 19, no. 4, pp. e373–e379, 2018. <https://doi.org/10.1016/j.clcc.2018.01.007>
- [11] P. Cao, X. Jia, X. Wang, L. Fan, Z. Chen, Y. Zhao, J. Zhu, and Q. Wen, "Deep learning radiomics for the prediction of epidermal growth factor receptor mutation status based on MRI in brain metastasis from lung adenocarcinoma patients," *BMC Cancer*, vol. 25, no. 1, p. 443, 2025. <https://doi.org/10.1186/s12885-025-13823-8>
- [12] P. Tabnak, Z. Kargar, M. Ebrahimnezhad, and Z. HajiEsmailPoor, "A Bayesian meta-analysis on MRI-based radiomics for predicting EGFR mutation in brain metastasis of lung cancer," *BMC Med. Imaging*, vol. 25, no. 1, p. 44, 2025. <https://doi.org/10.1186/s12880-025-01566-8>
- [13] Y. R. Li, Y. Jin, Y. L. Wang, W. Y. Liu, W. X. Jia, and J. Wang, "MR-based radiomics predictive modelling of EGFR mutation and HER2 overexpression in metastatic brain adenocarcinoma: A two-centre study," *Cancer Imaging*, vol. 24, no. 1, p. 65, 2024. <https://doi.org/10.1186/s40644-024-00709-4>
- [14] X. Xu, L. Huang, J. Chen, J. Wen, D. Liu, J. Cao, J. Wang, and M. Fan, "Application of radiomics signature captured from pretreatment thoracic CT to predict brain metastases in stage III/IV ALK-positive non-small cell lung cancer patients," *J. Thorac. Dis.*, vol. 11, no. 11, pp. 4516–4528, 2019. <https://doi.org/10.21037/jtd.2019.11.01>
- [15] Y. Niu, H. B. Jia, X. M. Li, W. J. Huang, P. P. Liu, L. Liu, Z. Y. Liu, Q. J. Wang, Y. Z. Li, S. D. Miao, and et al., "Deep learning radiomics and mediastinal adipose tissue-based nomogram for preoperative prediction of postoperative brain metastasis risk in non-small cell lung cancer," *BMC Cancer*, vol. 25, p. 1133, 2025. <https://doi.org/10.1186/s12885-025-14466-5>
- [16] J. Gong, T. Wang, Z. Z. Wang, X. Chu, T. D. Hu, M. L. Li, W. J. Peng, F. Feng, T. Tong, and Y. J. Gu, "Enhancing brain metastasis prediction in non-small cell lung cancer: A deep learning-based segmentation and CT radiomics-based ensemble learning model," *Cancer Imaging*, vol. 24, no. 1, p. 1, 2024. <https://doi.org/10.1186/s40644-023-00623-1>
- [17] S. Guo, L. Wang, Q. Chen, L. Wang, J. Zhang, and Y. Zhu, "Multimodal MRI image decision fusion-based network for glioma classification," *Front. Oncol.*, vol. 12, p. 8196878, 2022. <https://doi.org/10.3389/fonc.2022.819673>
- [18] X. Li, Y. Xu, F. Xiang, S. Wan, W. Huang, and B. Xie, "Prediction of IDH mutation status of glioma based on multimodal MRI images," in *Proceedings of the 2021 3rd International Conference on Intelligent Medicine and Image Processing*, New York, United States, 2021, pp. 39–44. <https://doi.org/10.1145/3468945.3468952>
- [19] F. Y. Zhu, Y. F. Sun, X. P. Yin, Y. Zhang, L. H. Xing, Z. P. Ma, L. Y. Xue, and J. N. Wang, "Using machine learning-based radiomics to differentiate between glioma and solitary brain metastasis from lung cancer and its subtypes," *Discover Oncol.*, vol. 14, no. 1, p. 224, 2023. <https://doi.org/10.1007/s12672-023-00837-6>
- [20] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, 2017. <https://doi.org/10.1158/0008-5472.CAN-17-0339>
- [21] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data

sets learning,” in *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2005, pp. 878–887.  
[https://doi.org/10.1007/11538059\\_91](https://doi.org/10.1007/11538059_91)

- [22] F. Hutter, H. Hoos, and K. Leyton-Brown, “An efficient approach for assessing hyperparameter importance,” in *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014, pp. 754–762.  
<https://proceedings.mlr.press/v32/hutter14.html>