A Novel Framework for Multimodal Brain Tumor Detection with Scarce Labels

Yanning Ge, Li Xu, Xiaoding Wang, Youxiong Que, and Md. Jalil Piran** (Senior Member, IEEE)

***Abstract*—Brain tumor detection has advanced significantly with the development of deep learning technology. Although multimodal data, such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT), has potential advantages in diagnostics, most existing studies rely solely on a single modality. This is because common fusion methods may lead to the loss of critical information when attempting multimodal fusion. Therefore, effectively integrating multimodal data has become a significant challenge. Additionally, medical image analysis requires large amounts of annotated data, and labeling images is a resource-intensive task that demands experienced professionals to spend a considerable amount of time. To address these challenges, this paper introduces a new unsupervised learning framework named Double-SimCLR. This framework builds on the foundation of contrastive learning and features a dual-branch structure, enabling direct and simultaneous processing of MRI and CT images for multimodal feature fusion. Given the "weak feature" characteristics of CT images (e.g., low soft tissue contrast and low resolution), we incorporated adaptive weight masking technology to enhance CT feature extraction. Moreover, we introduced a multimodal attention mechanism, which ensures that the model focuses on salient information, thereby elevating the precision and robustness of brain tumor detection. Even without substantial labeled data, experimental results demonstrate that Double-SimCLR achieves 93.458% accuracy, 92.463% precision, and a 93.058% F1-score, outperforming state-of-the-art (SOTA) models by 2.871%, 2.643%, and 3.098%, respectively.**

***Index Terms*—Brain Tumor Detection, Multimodal Image Processing, Contrastive Learning, Dual-Branch Network, Adaptive Weight Mask**

I. Introduction

Brain tumors are abnormal masses formed by uncontrolled cell proliferation in the brain, and they are among the leading diseases threatening human health worldwide [1]. Brain tumors can be classified into two types based on severity: benign and malignant. Most benign brain tumors are non-destructive and limited to nearby tissues. In contrast, malignant brain tumors grow at a much faster rate and invade surrounding tissues as they develop [2], [3]. Malignant brain tumors severely impact brain health; consequently, the 5-year survival rate is around 36%, while the 10-year survival rate is just below 31%. Thus, early diagnosis and treatment could significantly impact the survival of brain tumor patients.

Traditional manual detection of brain tumors primarily relies on the expertise of doctors. Given the significant variation in the shape and size of brain tumors, manually detecting and classifying brain tumor images is highly challenging. Moreover, the manual analysis of large volumes of medical images is not only tedious but also takes a considerable amount of time. Errors in brain tumor analysis can have serious consequences, directly affecting patient safety and well-being [4].

In recent years, deep learning (DL) technology has found extensive applications in processing multimodal images, particularly in medical image analysis, where supervised learning techniques are primarily used to enhance diagnostic precision. Supervised methods require very large annotated datasets, using representative images with correct brain tumor labels to train models. These models learn to identify diseased tissues. However, a significant challenge remains: the need for substantial amounts of labeled data for optimal functionality. In many cases, especially with rare diseases and specific types like medical imaging, creating sufficient annotated datasets is difficult. This process is not only expensive but also demands a significant amount of time from experienced professionals. This is a critical issue that needs to be addressed.

Recognizing the limitations of supervised learning due to the scarcity of large annotated datasets underscores the importance of using multiple imaging modalities. Currently, MRI and CT are the most popular imaging techniques used in brain tumor identification because of their respective advantages. MRI provides high-contrast

(*Corresponding Authors: Xiaoding Wang, Md. Jalil Piran)
Yanning Ge and Li Xu are with the College of Computer and Cyber Security, Fujian Provincial Key Laboratory of Network Security and Cryptology, Fujian Normal University, Fuzhou, Fujian 350117 P.R. China. E-mail: ged_mail@163.com, xuli@fjnu.edu.cn.
Xiaoding Wang is with the College of Computer and Cyber Security, Fujian Provincial Key Laboratory of Network Security and Cryptology, Fujian Normal University, Fuzhou, Fujian 350117, P.R. China. He is also with the National Key Laboratory for Tropical Crop Breeding, Institute of Tropical Bioscience and Biotechnology, Sanya Research Institute, Chinese Academy of Tropical Agricultural Sciences, Sanya, Hainan 572024 P.R. China. E-mail: wangdin1982@fjnu.edu.cn.
Youxiong Que is with the National Key Laboratory for Tropical Crop Breeding, Institute of Tropical Bioscience and Biotechnology, Sanya Research Institute, Chinese Academy of Tropical Agricultural Sciences, Sanya, Hainan 572024 P.R. China. He is also with the Key Laboratory of Sugarcane Biology and Genetic Breeding, Ministry of Agriculture and Rural Affairs, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, P.R. China. E-mail: queyouxiong@126.com.
Md. Jalil Piran is with the Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea. Email: piran@sejong.ac.kr.

images of soft tissues and is highly effective in delineating tumor boundaries and visualizing depth and relations to nearby structures. Typically, MRI images have higher resolution and grayscale levels, capturing more details and multi-parameter imaging information. In contrast, CT is a fast imaging technique that excels at showing bone structure and precise tumor location [5], [6]. Compared to MRI, CT exhibits "weak features," including lower soft tissue contrast, high homogeneity, and notably lower resolution, resulting in less information and complexity. Each imaging technique has its limitations in specific cases; for example, MRI has poor sensitivity for calcifications, while CT has limited specificity for distinguishing tumor margins from surrounding edematous tissue [7]. Therefore, utilizing multimodal imaging data is crucial for enhancing the accuracy and sensitivity of brain tumor diagnoses. While techniques for merging multiple images, such as MRI and CT, into a single image to improve diagnostic capabilities have advanced, this technology often involves complex preprocessing that can lead to information loss or mismatched fusion results [8]. This presents another significant issue that needs to be addressed.

To solve these problems, we propose a novel multimodal brain tumor detection framework named Double-SimCLR, which can directly and simultaneously process MRI and CT images. This framework is capable of extracting and fusing information from each modality at the feature level, enabling more accurate brain tumor detection.

The main contributions of this paper are summarized as follows:

- To address the low diagnostic accuracy associated with single-modal images, we developed a novel double-branch framework that combines features from MRI and CT images. By utilizing contrastive learning, we leverage unlabeled medical image data without requiring large annotated datasets, thus addressing the scarcity of medical image labels.
- To accommodate the differences between modalities, we adopted different strategies for handling the MRI and CT branches. Given the "weak features" of CT images, we introduced adaptive weight masking technology in the CT branch, which dynamically adjusts the weights of each layer. This enables the model to better adapt to the weak features of CT images, thereby improving the accuracy of feature extraction.
- To enhance feature extraction capabilities, we employed a multimodal attention mechanism that improves our model's ability to select features, allowing it to focus more on regions potentially affected by a brain tumor.
- We validated our framework on the Harvard Medical Image Fusion Dataset and also used the dataset provided by Cheng et al. [9]. Our framework demonstrated superior performance in brain tumor detection, surpassing state-of-the-art (SOTA) frameworks by 2.871% in accuracy, 2.643% in precision, and 3.098% in F1-score.

The rest of this paper is organized as follows: Section II provides a comprehensive overview of current research on detecting brain tumors. Section III presents the details of our proposed Double-SimCLR model. Section IV describes the design and implementation of our experiments. Section V concludes the paper.

II. Related Work

We discuss the deep learning technologies used in the detection of brain tumors from the following aspects: single-modality image recognition, multimodal image fusion, and unsupervised learning. We will then summarize their advantages and disadvantages.

Single-Modality Image Recognition. Single-modality image recognition technology is more popular because it simplifies data processing and speeds up diagnosis. For example, Lamrani et al. [10] developed a model to detect brain tumors using convolutional neural networks to enhance recognition capacity. Its task is to identify critical features within MRI images, thereby improving their efficiency and accuracy. Combining K-means clustering algorithms and SVM classifiers, Jamberi et al. [11] designed a diagnostic tool for distinguishing between benign and malignant brain tumors based on MRI images, enhancing accuracy and precision in the diagnostic process. Zubair and colleagues proposed an advanced AI-driven model that integrates the strengths of EfficientNetB2 with balanced and homomorphic filtering to further improve MRI image processing methods, maximizing performance in brain tumor detection.

However, single-modality imaging technology presents significant challenges because it cannot provide a comprehensive representation of the various characteristics of brain tumors [12]. Detecting brain tumors using only one imaging modality may lack the full spectrum of features necessary for understanding the disease and its pathogenesis compared to a multimodal approach. Consequently, more researchers are turning to multimodal imaging to gain a more holistic view of diagnosis and prognosis.

Multimodal Image Fusion. Human recognition is limited by the single source of information in vision; thus, research has gradually shifted toward multimodal image fusion. Badal et al. [14] proposed an end-to-end CNN model with multiple layers of convolution and nonlinear activation functions, which aids in automatically extracting multimodal image features by utilizing different information for each layer to achieve optimal interpretation. Guo et al. [15] introduced several imaging modalities, including CT, MRI, and PET, in their study using a CNN framework. This allowed the CNN to learn the complementary information contained in each modality during the feature extraction process. Similarly, Li et al. [16] considered various imaging technologies, such as MRI, CT, and SPECT, to propose multimodal medical image fusion algorithms that address the challenges of integrating features from different imaging modalities. However, important information may be lost during the fusion of

TABLE I: SOTA models' comparison.

Research	Techniques Used	Advantages	Limitations
Lamrani et al.(2022) [10]	Convolutional Neural Networks (CNNs) for MRI images	Improved efficiency and accuracy by automatically identifying key features	Limited to single modality imaging
Jamberi et al.(2024) [11]	K-means clustering algorithms and Support Vector Machine (SVM) classifiers for MRI images	Enhanced accuracy and efficiency of brain tumor diagnosis	Limited to single modality imaging
Zubair et al.(2024) [13]	EfficientNetB2 with balanced and homomorphic filtering for MRI images	Improved accuracy and efficiency through enhanced MRI image processing	Limited to single modality imaging
Sousa et al.(2023) [12]	Analysis of limitations of single-modality imaging	Highlighting limitations of single-modality imaging	Limited comprehensive feature capture
Badal et al.(2018) [14]	End-to-end CNN structure for multimodal image fusion	Effective extraction of multimodal features	Potential loss of critical information during fusion
Guo et al.(2019) [15]	CNN framework for handling different imaging modalities (CT, MRI, PET)	Enhanced feature extraction process	Challenges in combining features from different modalities
Li et al.(2021) [16]	Multimodal medical image fusion technique for MRI, CT, and SPECT	Improved fusion of different imaging technologies	Challenges in effectively combining features from different modalities
Zhou et al.(2019) [17]	Review of challenges in multimodal image fusion	Identifying critical information loss challenges	Critical information loss during feature extraction
Taher et al.(2022) [18]	Transfer learning-based approach combining three unsupervised clustering techniques for MRI images	Improved MRI image processing	Does not incorporate multiple imaging modalities
Saeed et al.(2022) [19]	Improved k-NN algorithm for MRI images	Novel clustering algorithm for better identification and localization	Does not incorporate multiple imaging modalities
Sankareswaran et al. (2022) [20]	Rigid Body Convolutional Neural Network (RBCNN) for MRI registration	Improved MRI registration accuracy and efficiency	Limited to single modality images

multimodal images, complicating effective integration of multi-modality features [17].

Unsupervised Learning. A wide range of unsupervised learning methods is also employed in brain tumor detection. Taher et al. [18] developed a transfer learning-based approach that combines three unsupervised clustering techniques: Gaussian Mixture Model, K-means algorithm, and Agglomerative Hierarchical Clustering, primarily aimed at accelerating brain tumor detection with greater accuracy in MRI image processing tasks. Saeed et al. [19] proposed an enhanced version of the k-NN algorithm, introducing a new unsupervised clustering mechanism that successfully detects and tracks the origins of brain tumors in MRI images. Another approach by Sankareswaran et al. utilized a Rigid Body Convolutional Neural Network for brain tumor MRI registration, presenting an unsupervised end-to-end method for medical image registration [20]. These methods have significantly improved the accuracy and efficiency of MRI image registration for detecting and tracking brain tumors. However, these approaches primarily focus on single-modality images, neglecting the information available from different imaging modalities.

Table I summarizes the comparison of state-of-the-art (SOTA) models for brain tumor detection. Based on a comprehensive analysis of the aforementioned research, our method introduces innovations and improvements by combining multimodal data with unsupervised learning techniques. Our aim is to address the limitations of existing technologies in the field of brain tumor detection.

III. Method

In the medical image analysis of brain tumors, several major challenges exist. Firstly, single-modality information is limited and cannot provide comprehensive diagnostic insights for brain tumors. Secondly, the cost of annotating medical image data is high, and the quantity of available annotated data is limited, making it a pressing issue to train an efficient model under these constraints.

To address these problems, we propose the Double-SimCLR framework, which aims to improve the accuracy

of brain tumor detection through a dual-branch contrastive learning method and an attention mechanism. Our Double-SimCLR model consists of two main modules: the *Feature Extraction Module* and the *Feature Fusion Module*. Table II lists the symbols used in this paper along with their descriptions.

A. Feature Extraction Module

To effectively extract features, we need to address the following two key problems: 1. How can we extract features under conditions of sparse labels? 2. How can we effectively extract features from multimodal data?

1) Solution to the Scarce Label Problem

While most tasks, such as image classification, involve supervised learning that requires a large amount of labeled data, collecting this data can be time-consuming and expensive [21], [22]. In light of this requirement, unsupervised learning has emerged as an alternative that operates without labeled data, learning directly from the structure and features inherent in the data [23].

Contrastive learning [24] (see Fig. 1 for illustration) is a form of unsupervised learning. It aims to increase similarity for samples from the same class while decreasing similarity for samples from different classes. For instance, in the context of image data, positive examples could be two images cropped from the same scene, while negative examples would be from different classes. The model learns to distinguish these through adjusting the distances—minimizing for positive pairs and maximizing for negative pairs using a contrastive loss [25], [26]:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left[y_{ij}\cdot d_{ij}^2 + (1-y_{ij})\cdot \max(m-d_{ij},0)^2\right], \quad (1)$$

where d_{ij} is the distance between sample pairs, (m) is a preset margin value used to distinguish between positive and negative samples, and y_{ij} is an indicator variable, with "1" denoting a positive sample pair and "0" denoting a negative sample pair.

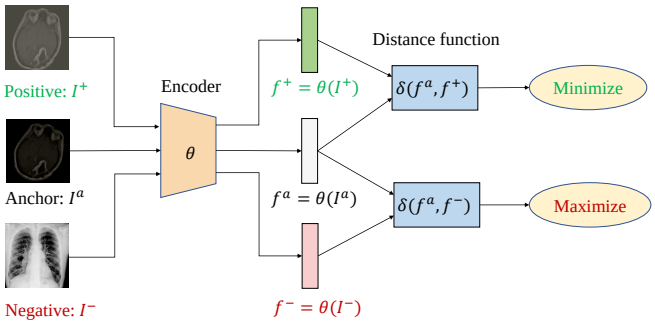

Fig. 1: Illustration of contrastive learning process.

As a contrastive learning model, SimCLR [25] effectively extracts features from single-modal data, making it an important component of the proposed Double-SimCLR model. We summarize the implementation of SimCLR in the following steps:

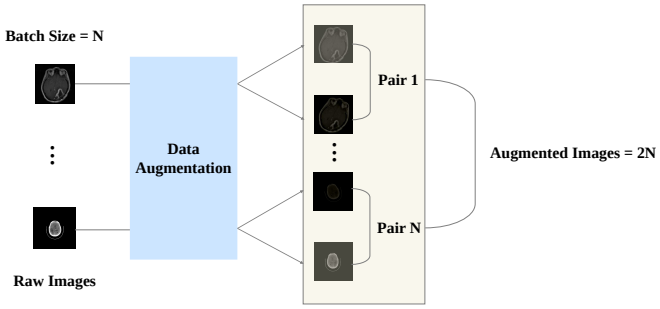

Fig. 2: Sample pair generation.

Input and Data Augmentation. The original instance x undergoes two different random data augmentation transformations from the same family of transformation methods (denoted as T), referred to as t and t', respectively. This method generates two variant images, $\tilde{x}_i$ and $\tilde{x}_j$, which may look different from each other but are effectively alternate views of the same original image. As a result, we obtain positive sample pairs. For any augmented version of an image, the rest of the augmented images in the batch can be considered negative samples. Fig. 2 illustrates the generation process of sample pairs, while Fig. 3 presents the differences in similarity between paired combinations of four images: CT, CT Augmented, MRI, and MRI Augmented.

Feature Extraction. An encoder network $f(\cdot)$ is then applied independently to each of these two augmented images to yield their feature representations, namely:

$$h_i = f(x_i), \; h_j = f(x_j). \quad (2)$$

Projection Head Mapping. Next, the two embeddings h_i and h_j are processed one more time with the projection head $g(\cdot)$ to further reduce their dimensions in a new space, resulting in z_i and z_j, respectively. This step is performed to compute the contrastive loss using these reduced-dimension features. Here, $g(\cdot)$ represents a feedforward neural network (FNN) with one hidden layer, achieving $z_i = g(h_i) = W^{(2)}\sigma(W^{(1)}h_i)$, where W denotes a linear transformation and $\sigma(\cdot)$ is the ReLU nonlinear activation function.

Maximizing Consistency. In the space mapped by the projection head, we maximize the consistency between positive pairs (z_i, z_j) while minimizing their consistency with other samples in the same batch. This encourages representations z_i and z_j from the same original image to be close together, while representations from different original images are pushed apart. To this end, we use the NT-Xent (Normalized Temperature-Scaled Cross-Entropy) loss. Specifically, the NT-Xent loss function introduces a temperature parameter to adjust the similarity distribution, making the differences between positive and negative pairs more pronounced, thereby enhancing the model's discriminative capability. Unlike traditional contrastive loss functions, as shown in Eq. (1), NT-Xent does not require additional threshold settings for positive and negative pairs, has lower computational complexity,

TABLE II: Symbols and meanings.

Symbol	Explanation	Symbol	Explanation
$\mathcal{L}$	Contrastive learning loss function	y_{ij}	Indicator variable
d_{ij}	Distance between sample i and sample j	m	Preset margin value
zi	Projection of the feature representation	$f(\cdot)$	Encoder network
z_c	Compressed feature vector	F_{sq}	Squeeze operation function
u_c	Feature vector of each channel in the input feature map	$u_c(i,j)$	Pixel value
W	Width of the feature map	H	Height of the feature map
s	Excitation weight vector	F_{ex}	Excitation function
z	Compressed feature vector	W	Weight matrix of the fully connected layers
σ	Sigmoid activation function	$g(z,W)$	Intermediate representation in excitation
W_2	Weight matrix of the second fully connected layer	δ	ReLU activation function
$h(\cdot)$	Projection head	W_1	Weight matrix of the first fully connected layer
$\tilde{U}_c$	Reweighted feature map	F_{scale}	Reweight function
U_c	Input feature map	s_c	Weights computed by the excitation function

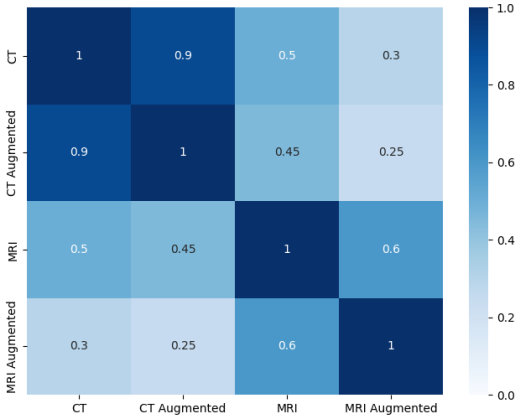

Fig. 3: Pairwise similarity among four images.

and enhances model robustness. For these reasons, we chose NT-Xent as the contrastive loss. Subsequently, we compute the similarity between positive and negative pairs using cosine similarity, defined as $\text{sim}(u,v) = \frac{u \cdot v}{\|u\|\|v\|}$, where u and v are feature vectors generated by the projection head. Formally, let $l_{i,j}$ represent the loss for a given positive pair (z_i, z_j). Then, it can be expressed as:

$$l_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} l_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)}, \tag{3}$$

where $l_{k \neq i}$ denotes an indicator function that is 1 if $k \neq i$ and 0 otherwise; τ is a temperature parameter used to control the scale of similarity scores.

2) Solution to Features Extraction from Multimodal Data

In order to extract features effectively from multi-modal data, we made improvements to the SimCLR model by upgrading the original single-channel input architecture of SimCLR to a more complex dual-branch network structure, thus it can simultaneously process multimodal data, e.g., both MRI and CT images. This dual-branch structure enables the model to capture information from different modalities more comprehensively at the initial stage, thereby enhancing the effectiveness of feature learning.

We performed dedicated optimizations over encoders and projection heads for each branch to enhance feature extraction from MRI and CT images. For the MRI image, considering the high characteristics of details, we set deeper layers with finer convolutional kernels. For the CT image branch, we adjusted the size of the receptive field and the depth of layers to enhance capture of low-contrast features. Furthermore, we considered modality-specific projection heads based on image characteristics to handle their respective structural information effectively.

3) Solution to Multimodal Data Flexibility

To better adapt to changes in data and enhance the overall performance of the model, we introduced an adaptive weight mask mechanism capable of dynamically adjusting the importance of output features at each layer. This adaptive weight mask is generated based on the input data and provides fine-grained control over the outputs from the model's layers. Moreover, we applied the adaptive weight mask mechanism to the CT branch. The following sections will discuss this technology in detail, using the CT branch as an example.

The output of each layer in the CT branch is multiplied by a dynamically generated weight mask. These dynamic weight masks are created by a small network conditioned on the CT input features and are adjusted dynamically during training. The CT branch comprises the following hierarchical structures: the convolutional layer, batch normalization layer, ReLU activation function, max pooling layer, residual block, and fully connected layer. For these layers, we define the following adaptive weight mask matrix M:

$$M = \begin{bmatrix} m_{conv1} \in \mathbb{R}^{N\times 64\times H/2\times W/2} \\ m_{bn1} \in \mathbb{R}^{N\times 64\times H/2\times W/2} \\ m_{relu} \in \mathbb{R}^{N\times 64\times H/2\times W/2} \\ m_{maxpool} \in \mathbb{R}^{N\times 64\times H/4\times W/4} \\ m_{residual} \in \mathbb{R}^{N\times 128\times H/4\times W/4} \\ m_{fc} \in \mathbb{R}^{N\times D} \end{bmatrix}, \tag{4}$$

where $\mathbb{R}$ represents the set of real numbers, N is the batch size, D is the output dimension, H is the height of the image, and W is the width of the image.

To retain more spatial information and simplify the branch structure, we removed the pooling layer and fully connected layer by setting their corresponding mask values to 0, while initializing the other parts to 1. The entire process consists of the following steps:

Weight Adjustment. The output of each layer is masked by

$$H_i' = H_i \odot m_i, \tag{5}$$

where H_i is the output of the ith layer, m_i is the dynamically generated weight mask, and $\odot$ denotes element-wise multiplication.

Mask Generation. The weight mask m_i is generated by a small neural network based on the input features:

$$m_i = \sigma(W_{mask} \cdot H_{i-1} + b_{mask}), \tag{6}$$

where σ is the activation function, and W_{mask} and b_{mask} are the trainable parameters of the mask generation network.

Backpropagation and Update. During backpropagation, we compute the gradient of the loss function with respect to the weight mask and update the parameters of the mask generation network using an optimization algorithm:

$$m_i \leftarrow m_i - \eta \frac{\partial L}{\partial m_i}, \tag{7}$$

where η is the learning rate.

B. Feature Fusion Module

After feature extraction, we introduced a feature fusion module to meticulously combine features from both modalities. Here, we employed a simple concatenation strategy, directly combining the features from the two branches. This method is straightforward to implement and computationally efficient, allowing for full retention of information from different modalities. Although simple concatenation might lose some correlations between the features, it is easy to implement and integrates multimodal features well. This approach is suitable for our current small-scale dataset, thereby mitigating the risk of overfitting that might arise from more complex fusion strategies. Subsequently, we applied an SEBlock [27] attention mechanism to the fused features to more accurately extract and weigh important features, enabling us to fine-tune the model with a small amount of labeled data and thus facilitating effective brain tumor detection.

Algorithm 1 Double-SimCLR

Input: batch size N, constant τ, structure of f_{mri}, f_{ct}, g, augmentation set $\mathcal{T}$
Output: encoder networks $f_{\text{mri}}(\cdot)$ or $f_{\text{ct}}(\cdot)$

for sampled minibatch $\{(x_{k,\text{mri}}, x_{k,\text{ct}})\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{x}_{2k-1,\text{mri}} = t(x_{k,\text{mri}})$
    $h_{2k-1,\text{mri}} = f_{\text{mri}}(\tilde{x}_{2k-1,\text{mri}})$ # extract features
    $z_{2k-1,\text{mri}} = \text{Flatten}(h_{2k-1,\text{mri}})$ #flatten features
    $\tilde{x}_{2k-1,\text{ct}} = t(x_{k,\text{ct}})$
    $h_{2k-1,\text{ct}} = f_{\text{ct}}(\tilde{x}_{2k-1,\text{ct}})$
    $z_{2k-1,\text{ct}} = \text{Flatten}(h_{2k-1,\text{ct}})$
    # the second augmentation
    $\tilde{x}_{2k,\text{mri}} = t'(x_{k,\text{mri}})$
    $h_{2k,\text{mri}} = f_{\text{mri}}(\tilde{x}_{2k,\text{mri}})$
    $z_{2k,\text{mri}} = \text{Flatten}(h_{2k,\text{mri}})$
    $\tilde{x}_{2k,\text{ct}} = t'(x_{k,\text{ct}})$
    $h_{2k,\text{ct}} = f_{\text{ct}}(\tilde{x}_{2k,\text{ct}})$
    $z_{2k,\text{ct}} = \text{Flatten}(h_{2k,\text{ct}})$
    # combine features from MRI and CT
    $z_{\text{combined}} = \text{Concat}(z_{\text{mri}}, z_{\text{ct}})$
    # project combined features into a new space
    $z = g(z_{\text{combined}})$
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \frac{z_i^\top z_j}{\|z_i\|\|z_j\|}$ {pairwise similarity}
  end for
  Calculate $\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f_{mri}, f_{ct}, and g to minimize $\mathcal{L}$
end for

A key component in Squeeze-and-Excitation Networks (SENet), the SEBlock mechanism enhances the representational power of neural networks through a lightweight, computationally efficient attention process. Contrary to more complex attention mechanisms, SEBlock's straightforward implementation only marginally increases computational overhead while significantly improving model performance. Specifically, SEBlock processes features through the following steps:

Squeeze. The Squeeze part of the SEBlock is designed to extract a global information embedding, which means compressing the spatial details of the feature vector of each channel of the feature map U. Thus, a single scalar for each channel captures global information regarding the spatial aspect of the input feature map. In this process, Global Average Pooling (G.A.P.) is performed. Specifically, it calculates the average across the feature maps for each channel c (where $c \in \{1, 2, \ldots, C\}$):

$$z_c = \mathbf{F}_{\text{sq}}(\mathbf{u}_c) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_c(i,j). \tag{8}$$

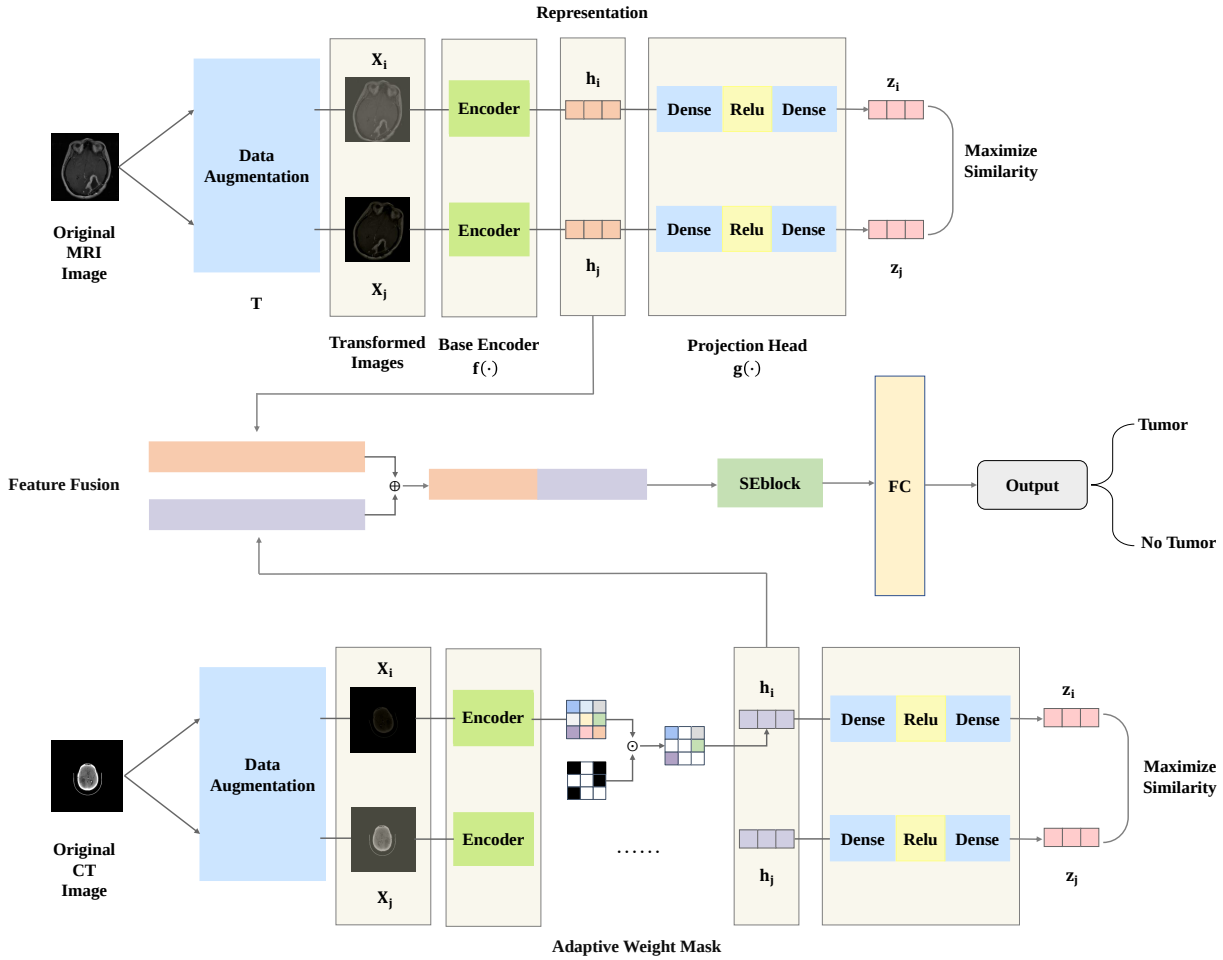

Fig. 4: The framework of Double-SimCLR.

Excitation. The primary function of the Excitation part is to adaptively assign feature weights to each of the C channels by learning z_c. The specific computation is as follows:

$$\mathbf{s} = \mathbf{F}_{\text{ex}}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(g(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z}))), \quad (9)$$

where $\mathbf{z}$ is the compressed feature vector, $\mathbf{W}_1$ and $\mathbf{W}_2$ are weight matrices, δ is the ReLU activation function, and σ is the Sigmoid activation function.

Reweight. This step applies the computed weights to recalibrate each channel in the original feature map, enhancing the representations and recognition of different features by the network. This is implemented for each channel by

$$\tilde{U}_c = F_{\text{scale}}(U_c, s_c) = s_c \cdot U_c, \quad (10)$$

where $\tilde{U}_c$ is the recalibrated feature map and s_c is the weight corresponding to channel c.

C. Overall Design of Double-SimCLR Model

In this section, we elaborate on the overall design of the Double-SimCLR model. This model directly fuses features from MRI and CT images through a two-branch structure. By employing unsupervised contrastive learning, it fully utilizes unlabeled medical image data, thereby avoiding the problem of scarce medical data labels. Additionally, a multimodal attention mechanism and adaptive weight masking technology are introduced to enhance the model's feature selection ability, allowing it to focus on areas that may contain brain tumors. The overall design of the Double-SimCLR model is shown in Fig. 4, while the training process is illustrated in Fig. 5. Specifically, the encoder $f(\cdot)$ and the projection head $g(\cdot)$ are first trained. After training, the projection head $g(\cdot)$ is discarded, and only the output h from the encoder $f(\cdot)$ is retained for feature extraction in downstream tasks. The m within the dashed box represents the features of the CT branch processed through the adaptive weighting mask mechanism. The pseudocode for Double-SimCLR is summarized in Algorithm 1.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3467343

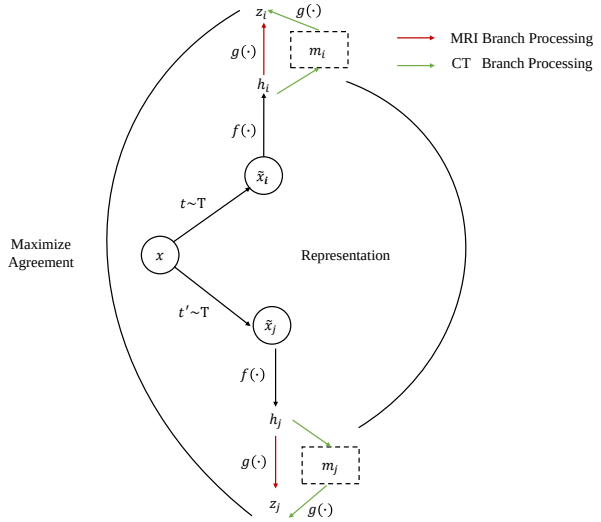

Fig. 5: The training process of MRI or CT branch.

IV. Experiments

In this section, we evaluate the performance of the proposed Double-SimCLR model, while compared with SOTA models SimCLR [25] and ResNet50 [28]. Our code can be accessed from the following link: https://github.com/MohamedAliHabib/Brain-Tumor-Detection.

A. Experimental Equipment and Parameter Description

Our model runs on the following equipment:

- CPU: 16 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz
- GPU: RTX 4090 (24GB) * 1

Software environment:

- Operating System: Ubuntu 20.04 LTS
- Programming Language: Python 3.8
- Dependencies:PyTorch 2.0.0;CUDA 11.8

The detailed configurations of the hyperparameters are given in Table III.

B. Dataset Preparation

The Double-SimCLR model in this paper requires paired MRI and CT images along with their corresponding labeled data. However, public brain tumor datasets do not meet this specific requirement. Therefore, we conducted experimental tests using the model code proposed by MohamedAliHabib to analyze brain tumor recognition in MRI images.

We labeled the data based on the following criteria: if the model predicts the probability of a brain tumor being present as greater than 50%, the corresponding image is labeled as "1"; otherwise, it is labeled as "0". Using this strategy, we collected and labeled a total of 1,178 paired MRI and CT images, with 589 images of each type. These images primarily come from the Harvard Medical Image Fusion Dataset [29]. The Harvard Medical Image Fusion Dataset is a multimodal medical image dataset focused on brain imaging. It contains matched images from CT, MRI, PET, and SPECT scans, mainly sourced from the Whole Brain Atlas at the Harvard Medical School publicly available database. We selected 184 pairs of brain tumor images from different patients, with each pair consisting of one CT image and one MRI image. All images have a resolution of 256x256 pixels, ensuring consistency and comparability in image quality. Below are some examples of the collected image data (see Fig. 6).

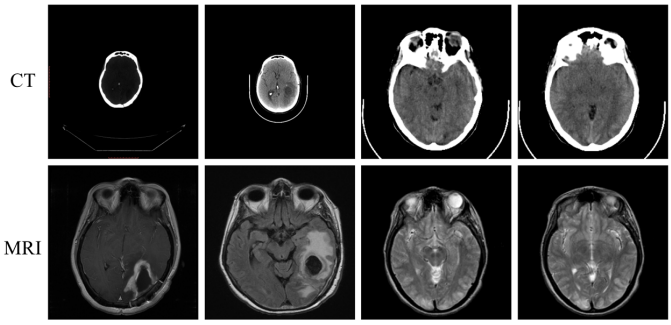

Fig. 6: Paired MRI and CT dataset display.

C. Discussion on the Scale of Dataset

This paper adopts a relatively small-scale dataset containing 589 MRI-CT image pairs. We are aware that the size of the dataset may affect the performance of generalization. The smaller the dataset, the easier it is for the model to overfit and struggle with learning typical variations in tumor characteristics.

However, we have the following reasons for choosing this dataset:

Data Quality. The images are primarily selected from the Harvard Medical Image Fusion Datasets, ensuring high-quality images with good alignment between MRI and CT scans.

Paired Characteristics. Fully matched pairs of MRI and CT images are very difficult to find, while our dataset provides this valuable paired resource.

Data Augmentation. Given the limitations on the size of our dataset, we utilized data augmentation techniques to increase the number of training samples, allowing for better learning of image features.

D. Data Preprocessing

It is evident that the brain sizes in the CT and MRI images in the dataset vary, and there is a significant amount of irrelevant information in the background that can interfere with image analysis. To remove non-target regions and crop areas likely representing the brain, we employed the cropping technique proposed by Dahiwade et al. [30] for preprocessing the dataset, following these steps. To provide a more intuitive demonstration of the results at each processing step, we present the outcome of each step in Fig. 7, while Fig. 8 shows the resulting dataset images processed through following steps:

Grayscale Conversion and Gaussian Blur. First, we converted the MRI and CT images to grayscale to

TABLE III: Hyperparameters for contrastive learning and downstream task phases.

Phase	Hyperparameter	Value	Description
Contrastive Learning	Learning Rate	0.001	Learning rate for the contrastive learning phase.
	Batch Size	32	Batch size for each training iteration.
	Temperature	0.5	Temperature parameter for contrastive loss.
	Projection Head	2 layers, 2048 * 2 neurons	Two-layer projection head, input dimension 2048 * 2, output dimension 128.
	Training Epochs	500	Total number of training epochs.
	Optimizer	Adam	Optimizer used for the contrastive learning phase.
Downstream Task	Learning Rate	0.001	Learning rate for the downstream task phase.
	Batch Size	32	Batch size for each training iteration.
	Training Epochs	300	Total number of training epochs.
	Optimizer	Adam	Optimizer used for the downstream task phase.
	Weight Decay	1×10^{-6}	L2 regularization term to prevent overfitting.

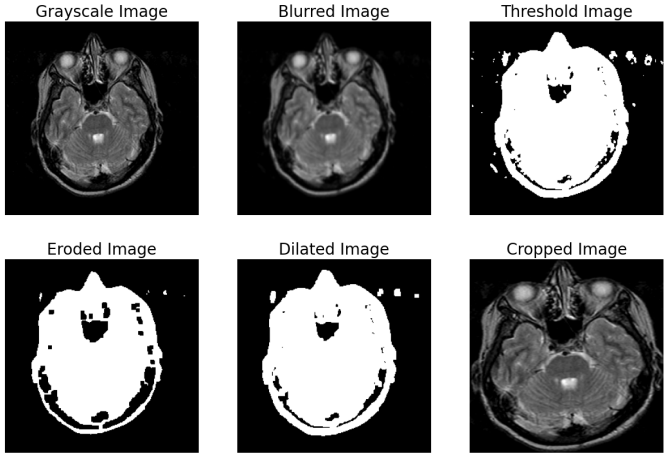

Fig. 7: The visualization of each step's outcome.

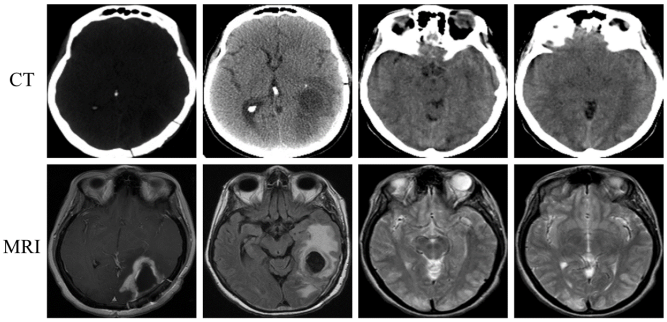

Fig. 8: Display of paired MRl and CT dataset after processing.

simplify them to a single color channel. After this, we applied Gaussian blur to reduce image noise. Gaussian blur smooths the image using a Gaussian function controlled by standard deviation, which is important for removing minor disturbances and providing a solid basis for precise thresholding of brain contours.

Binary Thresholding. Afterward, we applied binary thresholding to the blurred images. This step sets pixels with values less than the threshold to zero (black), while pixels with values greater than the threshold are set to 255 (white). This results in an image with clear differentiation between the background and the region of interest, where the target area is marked white, facilitating further analysis.

Morphological Noise Removal. To enhance the recognition of brain contours, we performed morphological processing on the binary thresholded images, including erosion and dilation operations. This effectively removes small noise points created by thresholding and helps close small gaps within the brain contours, ensuring continuity and completeness for the next contour detection step.

Brain Contour Detection. Using the processed thresholded images, we performed contour detection. Contours are lines that connect all continuous points along a boundary with the same color or brightness. We utilized functions from the "imutils" library to find the main contours based on area size, which usually correspond to the primary regions of interest in the image—the brain.

Extremity Points Calculation. We calculated the extremity points of the detected brain contours, including the leftmost, rightmost, topmost, and bottommost points. These points are derived from the extreme horizontal and vertical coordinates of the contour points, accurately marking the spatial position of the brain in the image.

Image Cropping Based on Brain Contour. Once these extremity points are calculated, a cropping region is defined as a rectangle in the original image, bounded by the furthest extremity points of the contour. The resulting cropped image tightly envelops the identified brain region.

E. Experimental Results

1) Performance Comparison

We conducted experiments based on the aforementioned dataset to perform a comparative analysis of the performance of four models: Double-SimCLR, BTD-CNN [10], EFDL-BTD [18], and ResNet50. After 300 epochs of iterative training, the results are shown in Fig. 9. From the figure, it is evident that the overall accuracy

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3467343

of Double-SimCLR is higher than that of EFDL-BTD and ResNet50. Although its performance is slightly lower than that of BTD-CNN during the initial training phase, Double-SimCLR demonstrates higher accuracy in the later convergence stage, surpassing BTD-CNN. The specific experimental results are presented in Table IV.

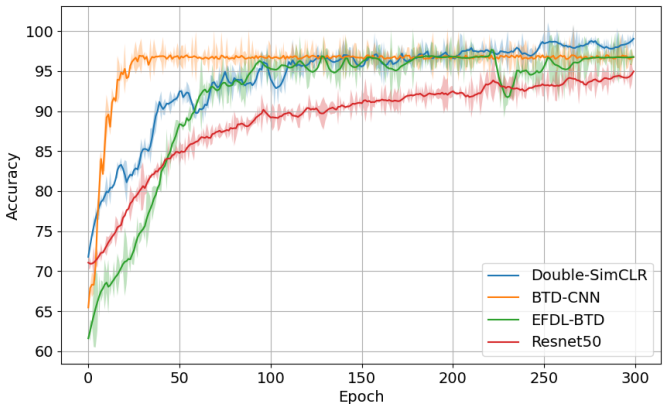

Fig. 9: Comparison of accuracy of model training processes.

TABLE IV: Model performance metrics across test sets.

Test Set	Model	Accuracy	Precision	F1-score
1	Ours	**0.9568**	0.9581	**0.9673**
	BTD-CNN	0.9375	**0.9600**	0.9412
	EFDL-BTD	0.8729	0.8730	0.8593
	Resnet50	0.8715	0.8718	0.8843
2	Ours	**0.9623**	**0.9513**	**0.9567**
	BTD-CNN	0.9483	0.9413	0.9494
	EFDL-BTD	0.9218	0.9184	0.9207
	Resnet50	0.8333	0.8462	0.8462
3	Ours	**1.0000**	**1.0000**	**1.0000**
	BTD-CNN	0.8750	0.8718	0.8843
	EFDL-BTD	0.8333	0.8333	0.8397
	Resnet50	0.8333	0.8571	0.8571
4	Ours	**0.9518**	**0.9493**	**0.9664**
	BTD-CNN	0.8750	0.8235	0.8235
	EFDL-BTD	0.9357	0.9375	0.9302
	Resnet50	0.8750	0.8750	0.8750

We evaluated the trained model on four test sets, which are divided according to image resolution, image type, image quality, and tumor boundary clarity as below.

- Test Set 1: This test set comprises high-resolution MRI images to evaluate the model on high-quality images.
- Test Set 2: This set includes standard-resolution CT images, emphasizing the model's capability to process regular clinical data.
- Test Set 3: A test set consisting of challenging MRI images with high levels of noise or unclear tumor boundaries is used to test the model's robustness.
- Test Set 4: This set contains both MRI and CT images to evaluate the model's integration capability on multimodal datasets.

Our model demonstrated exceptional performance on Test Set 3, which included complex MRI images characterized by high noise and blurred tumor boundaries, achieving an accuracy of 1.0000. This indicates that our model is highly robust under conditions of degraded image quality. The key factor contributing to this achievement lies in the integration of contrastive learning strategies combined with adaptive masking techniques, allowing the model to effectively discern and extract important discriminative features from noisy data.

We further demonstrated the models' accuracy on four different test sets more intuitively using box plots, as shown in Fig. 10. Our method (Ours) indicates the highest median accuracy and the smallest range of accuracy distribution, demonstrating very stable and accurate performance across the different test sets.

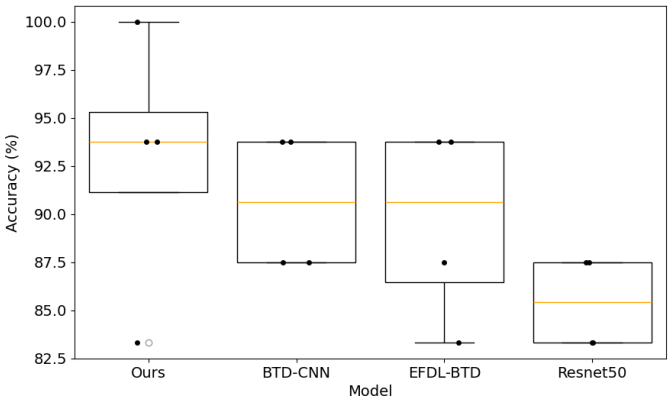

Fig. 10: Comparison of the performance of three models.

2) Ablation Experiment

It can be seen in Fig. 11 that the accuracies of the dual-branch Double-SimCLR are significantly higher compared to those of the single-branch model, which did not improve much and remained around 75%. This result is expected, as contrastive learning typically requires a large amount of data to learn key features from images. Our dataset contains only 589 pairs of MRI and CT images, and the single-branch SimCLR was capable of learning information from just one modality. This limitation greatly affected the learning process. Additionally, the ablation experiment with MSEBlock, represented by six curves, shows that adding the SEBlock to the model stabilizes the learning process and increases accuracy, which is quite encouraging. Furthermore, it incorporates an adaptive weight mask mechanism in the CT branch to accommodate different data. The differences and changes in accuracy between the MRI and CT branches can also be intuitively observed from the figure.

Figures 12 and 13 show the training losses of Double-SimCLR and SimCLR during two training phases. The training losses for both models decreased significantly over time in the contrastive learning phase, but the loss values for Double-SimCLR were lower and converged more quickly. In the initial phase (the first 50 epochs), the loss value of Double-SimCLR dropped rapidly from about 4.0 to 2.5, while SimCLR's loss value decreased from 4.0 to around 2.75. After 100 epochs, the loss value of Double-SimCLR stabilized at about 2.25, while SimCLR's loss value stabilized around 2.5. This indicates that Double-

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3467343

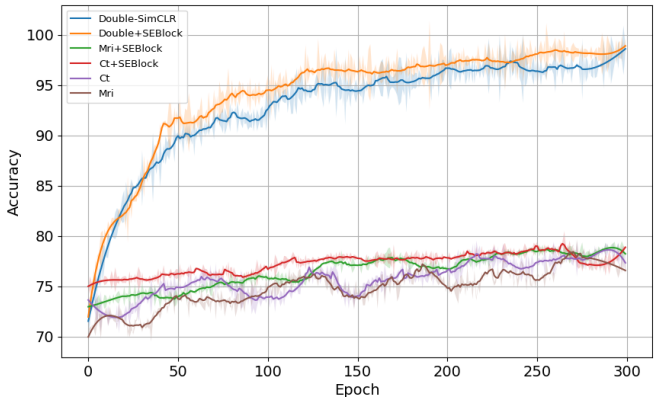

Fig. 11: Performance comparison of the model after ablation.

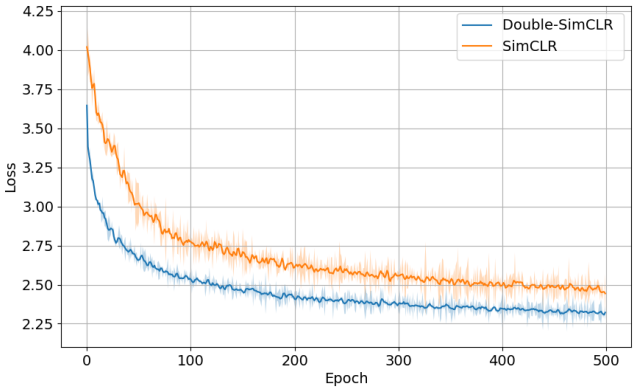

Fig. 12: Comparison of losses in the contrastive learning.

SimCLR was more effective in learning the main features of the data during the contrastive learning phase, primarily because it can integrate multimodal information, thereby improving feature extraction efficiency and representation capability.

In the downstream task phase, both Double-SimCLR and SimCLR quickly reduced their loss values from 3.0 to around 0.5 and 0.75, respectively, during the first phase (the first 10 epochs). After 50 epochs, the loss value of Double-SimCLR decreased further and stabilized at around 0.25, while SimCLR's loss value remained around 0.5. This indicates that Double-SimCLR achieved faster convergence and a lower final loss value in the downstream task phase, which suggests better generalization ability and performance in practical tasks.

From this, it is evident that our proposed dual-branch model, Double-SimCLR, which integrates multimodal information, reached an accuracy of up to 98% during training, surpassing traditional models. This demonstrates that the fusion of multimodal information significantly enhances model performance in cases of limited data. Double-SimCLR excels in extracting and representing image features by processing different modalities in parallel and performing feature fusion, thereby achieving complementary and enriched information.

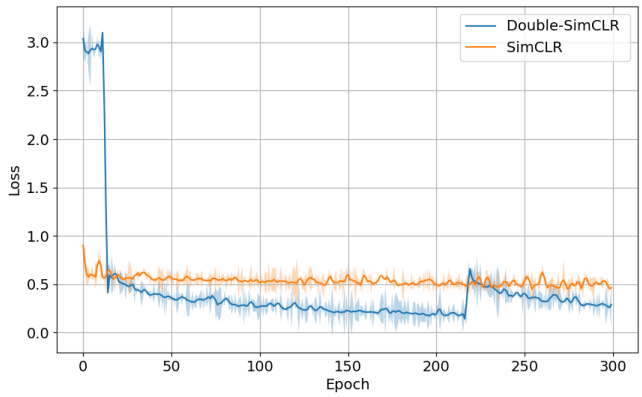

Fig. 13: Comparison of losses in the downstream task.

3) Batch Size Impact on Model Accuracy

In the following section, we discuss the impact of batch size on the model's accuracy and loss. Figures 14 and 15 show the changes in accuracy and loss for different batch sizes.

As shown in Fig. 14, generally, the top-1 accuracy of the model improves with an increase in batch size. Notably, when the batch size is 512, the accuracy after 200 epochs is significantly better than that of other batch sizes. However, the instability associated with larger batch sizes also leads to more significant fluctuations in the accuracy curve.

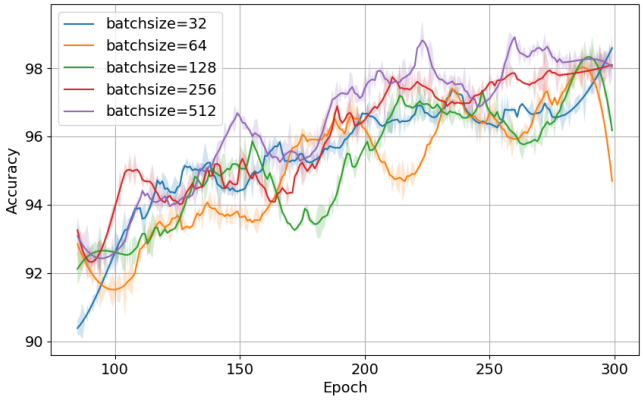

Fig. 14: The accuracy comparison of different batchsize.

Fig. 15 shows the changes in training loss for different batch sizes. The figure clearly demonstrates that batch size has a significant impact on loss. Specifically, smaller batch sizes (like 32 and 64) exhibited a rapid decrease in loss during the early phase of training and maintained lower loss levels throughout the process. In contrast, larger batch sizes (like 256 and 512), while achieving better accuracy, showed higher loss values and slower convergence during training.

Afterward, we tested the top-1 accuracy under various batch sizes using box plots, as shown in Fig. 16. From the analysis of the results, we found that the individual top-1 accuracies were nearly invariant across different batch sizes (32, 64, 128, 256, 512), with medians all close to 95%. The

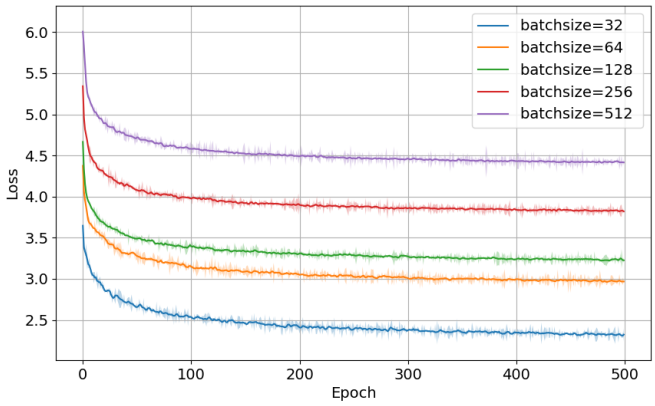

Fig. 15: The loss comparison of different batchsize.

model exhibited lower variability and greater stability with batch sizes of 256 and 512. However, with a batch size of 32, there was more variability, as the accuracy was more spread out. Overall, larger batch sizes resulted in better accuracy and stability as the model trained.

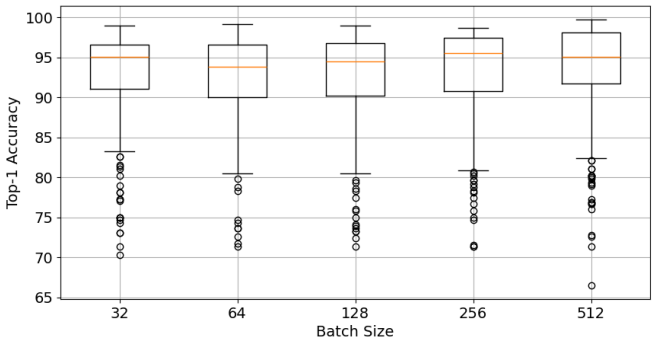

Fig. 16: Top-1 accuracy boxplot for different batchsize.

4) Data Augmentation Impact on Model Accuracy

To further verify the impact of data augmentation methods on the overall performance of the model, we selected four techniques: contrast adjustment, brightness adjustment, flipping, and rotation. We conducted validation analysis on their pairwise combinations. Due to the original image size of 240×240 causing memory overflow during training, we cropped the images to 32×32 for the subsequent experiments. Table V shows the specific parameter values for each data augmentation method.

TABLE V: Data augmentation parameters.

Data Augmentation Methods	Parameters
Random Resized Crop	Crop Size: 32x32
Random Horizontal Flip	Probability: 1
Color Jitter (Brightness)	Adjustment Range: 0.5
Random Rotation	Rotation Angle: ±30°

Fig. 17 and 18 show the heatmap representations of the four transformations. The diagonal values are set to 0 by default, while the values in other regions represent the average top-1 accuracy after 300 training epochs for each combination of data augmentations (rounded to four decimal places). From the combined accuracy and loss value heatmaps, we find that the combination of contrast adjustment and rotation achieves an average accuracy of 93.4699%, surpassing other combinations.

Upon deeper analysis, as observed from the corresponding violin plots (Fig. 19), the Contrast+Rotate combination exhibits the highest mean accuracy and the most concentrated accuracy distribution. Its median and mean also compare favorably to other combinations, with smaller variance and a reasonable range of extreme values. There is no doubt that, in terms of both stability and overall performance, the Contrast+Rotate combination outperforms other data augmentation strategies.

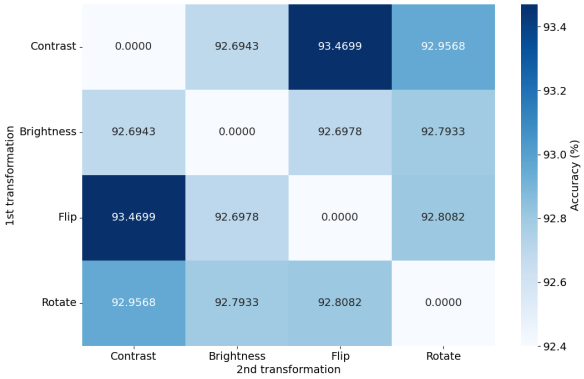

Fig. 17: Accuracy heatmap.

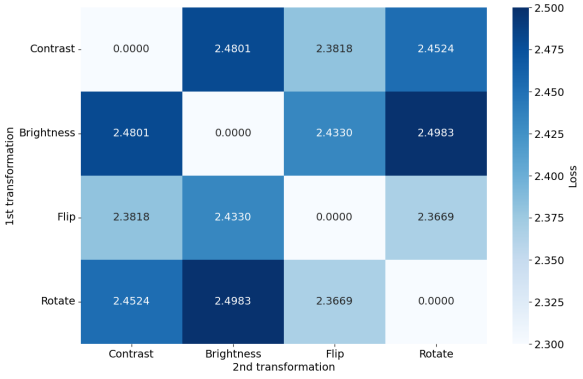

Fig. 18: Loss heatmap.

This result indicates that exploring superior data augmentation methods significantly aids the Double-SimCLR model in learning image features, offering new directions for future model improvements.

V. Conclusion

Detecting brain tumors is a significant challenge due to the brain's complex structure and the diverse morphology of tumors, which complicates the analysis of medical images. Most existing techniques for estimating flow fields in images rely on supervised learning that requires large labeled datasets. Moreover, the substantial benefits of multiple modalities on diagnostic accuracy are often

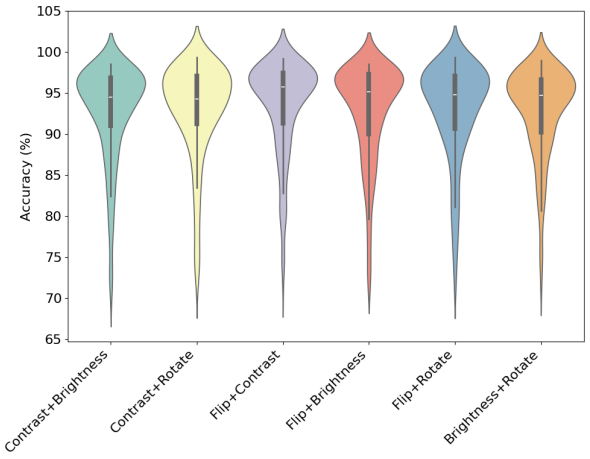

Fig. 19: Accuracy distribution for different data augmentation combinations.

overlooked. In this paper, we propose the Double-SimCLR framework to integrate multimodal imaging data from MRI and CT using unsupervised contrastive learning. This framework offers several advantages over state-of-the-art (SOTA) models. First, by utilizing unsupervised learning, Double-SimCLR circumvents the need for extensive labeled datasets, making it more scalable and cost-effective. Second, the framework's ability to integrate data from multiple modalities enhances its diagnostic accuracy by effectively capturing complementary information from both MRI and CT images, which are critical for accurate tumor detection and characterization. Experimental results demonstrate that Double-SimCLR achieves an accuracy of 93.458%, precision of 92.463%, and an F1-score of 93.058%, outperforming SOTA models by 2.871%, 2.643%, and 3.098%, respectively. These improvements underscore the efficacy of our framework in leveraging multimodal data and unsupervised learning to address the challenges of brain tumor detection.

However, while Double-SimCLR shows promise, it is not without potential challenges. One area for further exploration is the optimization of data augmentation strategies, which play a crucial role in contrastive learning. We studied how different data augmentation schemes affected the model and conducted ablation studies on all possible pairs of the four methods listed above. Additionally, it would be interesting to further explore better combinations of augmentation methods in the future.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62471139). It is also supported by the Open Research Project in Traditional Chinese Medicine Orthopedics at Fujian University of Traditional Chinese Medicine (Nos. XGS2023002, XGS2023003).

References

[1] T. Zhou, S. Canu, and S. Ruan, "Fusion based on attention mechanism and context constraint for multi-modal brain tumor segmentation," *Computerized Medical Imaging and Graphics*, vol. 86, p. 101811, 2020.

[2] P. Shanthakumar and P. Ganeshkumar, "Performance analysis of classifier for brain tumor detection and diagnosis," *Computers & Electrical Engineering*, vol. 45, pp. 302–311, 2015.

[3] Nitish, A. K. Singh, and R. Singla, "Different approaches of classification of brain tumor in mri using gabor filters for feature extraction," in *Soft Computing: Theories and Applications: Proceedings of SoCTA 2018*. Springer, 2020, pp. 1175–1188.

[4] N. Ullah, A. Javed, A. Alhazmi, S. M. Hasnain, A. Tahir, and R. Ashraf, "Tumordetnet: A unified deep learning model for brain tumor detection and classification," *Plos one*, vol. 18, no. 9, p. e0291200, 2023.

[5] J. J. Goud, "Brain tumor detection in medical imaging focusing on mri," *International Journal for Research in Applied Science and Engineering Technology*, 2023.

[6] Y. Luo, S. Zhang, J. Ling, Z. Lin, Z. Wang, and S. Yao, "Mask-guided generative adversarial network for mri-based ct synthesis," *Knowledge-Based Systems*, p. 111799, 2024.

[7] M. Mair, H. Singhavi, A. Pai, M. Khan, P. Conboy, O. Olaleye, R. Salha, P. Ameerally, R. Vaidhyanath, and P. Chaturvedi, "A systematic review and meta-analysis of 29 studies predicting diagnostic accuracy of ct, mri, pet, and usg in detecting extracapsular spread in head and neck cancers," *Cancers*, vol. 16, no. 8, p. 1457, 2024.

[8] F. G. Veshki, N. Ouzir, S. A. Vorobyov, and E. Ollila, "Coupled feature learning for multimodal medical image fusion," *arXiv preprint arXiv:2102.08641*, 2021.

[9] C.-B. Jin, "Mri-to-ct-dcnn-tensorflow," https://github.com/ChengBinJin/MRI-to-CT-DCNN-TensorFlow, note = commit xxxxxxx, 2019.

[10] D. Lamrani, B. Cherradi, O. El Gannour, M. A. Bouqentar, and L. Bahatti, "Brain tumor detection using mri images and convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022.

[11] K. Jamberi, S. Prathap, and A. A. Raj, "Machine learning-based brain tumor detection and classification: A hybrid approach using k-means clustering and support vector machines."

[12] J. V. Sousa, P. Matos, F. Silva, P. Freitas, H. P. Oliveira, and T. Pereira, "Single modality vs. multimodality: What works best for lung cancer screening?" *Sensors*, vol. 23, no. 12, p. 5597, 2023.

[13] A. Zubair Rahman, M. Gupta, S. Aarathi, T. Mahesh, V. Vinoth Kumar, S. Yogesh Kumaran, and S. Guluwadi, "Advanced ai-driven approach for enhanced brain tumor detection from mri images utilizing efficientnetb2 with equalization and homomorphic filtering," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 113, 2024.

[14] C. Badal Regàs, "Multi-modal medical image fusion using convolutional neural networks," B.S. thesis, Universitat Politècnica de Catalunya, 2018.

[15] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, 2019.

[16] Y. Li, J. Zhao, Z. Lv, and J. Li, "Medical image fusion method by deep learning," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 21–29, 2021.

[17] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, p. 100004, 2019.

[18] F. Taher, M. R. Shoaib, H. M. Emara, K. M. Abdelwahab, F. E. Abd El-Samie, and M. T. Haweel, "Efficient framework for brain tumor detection using different deep learning techniques," *Frontiers in Public Health*, vol. 10, p. 959667, 2022.

[19] S. Saeed, A. Abdullah, N. Jhanjhi, M. Naqvi, and A. Nayyar, "New techniques for efficiently k-nn algorithm for brain tumor detection," *Multimedia Tools and Applications*, vol. 81, no. 13, pp. 18 595–18 616, 2022.

[20] S. P. Sankareswaran and M. Krishnan, "Unsupervised end-to-end brain tumor magnetic resonance image registration using rbcnn: rigid transformation, b-spline transformation and convolutional neural network," *Current Medical Imaging*, vol. 18, no. 4, pp. 387–397, 2022.

[21] C. M. Bishop, "Pattern recognition and machine learning," *Springer google schola*, vol. 2, pp. 1122–1128, 2006.

[22] X. Qi, L. E. Brown, and R. X. Hawkins, "Small but mighty: Adversarial perturbations for data-free knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.
[23] P. Dayan, M. Sahani, and G. Deback, "Unsupervised learning."
[24] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
[26] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.
[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[29] "Havard-medical-image-fusion-datasets," https://github.com/xianming-gu/Havard-Medical-Image-Fusion-Datasets?tab=readme-ov-file„ 2022.
[30] D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2019, pp. 1211–1215.