Ullah *et al.*
European Journal of Medical Research (2026) 31:550
https://doi.org/10.1186/s40001-026-04161-x

European Journal
of Medical Research

RESEARCH

Open Access

Multimodal brain tumor segmentation and classification based on optimized DeepLabV3 + and fused fire module with self-attention

Muhammad Sami Ullah[1], Muhammad Attique Khan[2*], Yunyoung Nam[3*], Nouf Abdullah Almujally[4], Areej Alasiry[5], Mehrez Marzougui[5], Juan M. Gorriz[6] and Amir Hussain[7,8]

Abstract

In this work, we propose a novel deep learning architecture for brain tumor segmentation and classification, SMDeepNet, which is based on an optimized DeepLabV3 + + and a Fused Fire Module with Self-Attention. The segmentation framework comprises down- and up-sampling based on a DeepLabV3 + neural network and is optimized by dynamically initializing hyperparameters for training the backbone ResNet-50 architecture. During the down-sampling or encoder stage, the Atrous Spatial Pyramid Pooling (ASPP) module extracted features using convolutional layers with various filter sizes and dilations. These features are then passed to the up-sampling or decoder section for final segmentation. The classification framework comprises two sub-frameworks: a fire-residual bottleneck (Fire-RB) and a Hybrid Efficient Attention (Hybrid-EA). The Fire-RB framework comprises several parallel blocks: one side implements squeeze-and-expand behavior in the fire mechanism, and the other implements residual bottlenecks. The two parallel blocks are concatenated, and features are extracted from Fire-RB. The Hybrid-EA model is a custom variant of the pre-trained EfficientNetB0 model that incorporates a self-attention mechanism. The self-attention mechanism enhanced the functionality of the EfficientNetB0 model. Features from Fire-RB and Hybrid-EA are concatenated channel-wise, and the final modality classification is performed. The BraTS 2023 dataset is used in this work to evaluate the proposed methodologies. Segmentation results indicate that accuracy, Dice Score, and Intersection over Union (IOU) are 0.9871, 0.9420, and 0.8951, respectively. Modality classification results indicate an accuracy of 0.9920, which is improved over the recent state-of-the-art techniques.

Keywords Brain tumor, Multimodal, Neuroscience, Magnetic resonance imaging, Tumor segmentation, Self-attention, Fire mechanism, Networks fusion, Neural network

*Correspondence:
Muhammad Attique Khan
attique.khan@ieee.org
Yunyoung Nam
ynam@sch.ac.kr
Full list of author information is available at the end of the article

Introduction

Brain tumor is a deadly disease, and it affects thousands of people yearly around the globe [1]. The diagnosis of a brain tumor at an early stage may help the patient's survival [2]. Clinical practitioners and radiologists diagnose it with invasive procedures; however, these techniques create health hazards and take a lot of time in patient recovery [3]. Magnetic resonance imaging (MRI) scan for the brain consists of four modalities; these are named as fluid-attenuated inversion recovery (FLAIR), T1 weighted (T1), T2 weighted (T2), and T1 weighted with contrast enhancement (T1CE) [4–6]. These offer complimentary and distinct insights into the anatomy and disorders of the brain [7]. Precise classification guarantees that each modality consists of distinctive attributes or features and can be used to enhance the total precision and accuracy of diagnosis [8]. Considering that every modality emphasizes a distinct facet of brain pathology, it is essential to select those that are optimum for therapy planning; for example, T1CE helps detect abnormalities of the blood–brain barrier, but FLAIR is more sensitive to gliosis and edema, so it is pertinent to choose these images with an accurate classifica*tion mechanism for radiologists* [9].

Accurate classification and segmentation are necessary before developing a diagnostic system [10, 11]. Accurately classifying modalities is crucial in medical research to validate and create new therapeutic approaches and diagnostic tools [12, 13]. Furthermore, in clinical settings, effective classification facilitates the prompt identification and prioritization of images for examination by radiologists and healthcare professionals, guaranteeing fast and correct patient treatment and handling the storage and retrieval of substantial amounts of imaging data [14, 15].

The integration of artificial Intelligence methods maximizes the diagnostic potential of MRI, thereby advancing medical research and improving patient care [16, 17]. Since MRI is a non-invasive imaging method, it is typically the first step in the diagnosis and segmentation of brain tumors [18]. A correct diagnosis depends on the precise identification of the size, location, and extent of brain tumors, which is enabled by segmentation [19]. The ability to distinguish among tumor types and to determine tumor aggressiveness is crucial for developing effective treatment plans [20]. The provision of precise anatomical information and accurate segmentation facilitates improved planning of radiation, chemotherapy, and surgical procedures [21]. It ensures that the tumor is successfully targeted while protecting healthy brain tissue, thereby reducing the risk of adverse effects and increasing patient-reported positive outcomes [22]. Clinicians can assess the response of brain tumors to therapy, identify recurrences early, and modify treatment strategies by comparing segmented images from various time stamps [23]. Segmentation unveils quantitative information about tumors in medical research [24]. It enables the generation of large, structured datasets that are used to train and evaluate machine learning models, thereby improving diagnostic imaging and personalized healthcare [25].

Deep convolutional neural networks (DCNNs) improve the classification of MRI modalities and segmentation of brain tumors [26]. They are very good at extracting complicated characteristics from MRI images for modality classification [27]. Each neural network layer learns modality-specific information that captures varying levels of abstraction, which further allows for discrimination between multiple MRI modalities, including FLAIR, T1, T2, and T1CE [28]. Deep learning models combine features to increase classification accuracy by integrating data from several MRI modalities [29]. Customized deep convolutional neural networks (CNNs) with sophisticated architectural strategies like bottlenecks, inverted residuals, and self-attention mechanisms significantly improve modality classification and segmentation for brain tumor MRI data [30]. To correctly identify and segment tumors, bottleneck layers lighten and speed up the model without sacrificing critical features by reducing the computational cost of deep networks without compromising performance [31]. Richer feature representations are necessary for differentiating between MRI modalities and precisely segmenting tumor areas [32]. Inverted residual architecture is used in the pre-trained MobileNetV2 model, which increases efficiency by balancing depth and computing cost [33]. Self-attention mechanism is used to dynamically focus on different areas of the input image, which improves contextual awareness and allows the network to handle variations in the size of the tumor, appearance, and shape [34, 35]. In a nutshell, customized DCNNs can be developed by employing different neural network architectures that can effectively segment the brain tumor part from an MRI scan and can classify the modalities of MRI [36]. In light of these challenges, it is necessary to contribute to the segmentation and modality classification of brain tumor MRI scans using a deep learning approach. Our significant contributions to this work are as follows:

- An optimized DeepLabV3 + approach is proposed for the segmentation of brain tumors. In this approach, a down-sampling and up-sampling-based architecture, along with Atrous Spatial Pyramid Pooling (ASPP), is employed. The hyperparameters are initialized using the Bayesian optimization method.

- A customized Fire-Residual Block (Fire-RB) framework is proposed for the classification of brain tumor MRI modalities.
- A customized Hybrid Efficient Attention (Hybrid-EA) model is proposed in which the capability of EfficientNet-B0 is enhanced by customizing it with a self-attention mechanism.
- A channel-wise concatenation mechanism is introduced for both customized models for the classification of MRI modalities of brain tumors.
- The proposed classification architecture is further interpreted using explainable AI techniques such as LIME and Grad-CAM. These techniques demonstrate how well the proposed architecture performs in a clinical setting. Additionally, several ablation studies have been conducted to validate the performance of the proposed architecture.

Literature review

It is expected that over 1 million Americans already live with a primary brain tumor, and another 94,000 will receive a diagnosis in 2024. Therefore, it is essential to diagnose brain tumors at an early stage. Several computer vision techniques have been introduced in the literature for the detection and classification of brain tumors from MRI data [37, 38]. Segmentation techniques for brain tumors are based on custom CNN and U-Net architectures, whereas classification techniques are based on pre-trained CNN architectures.

Bouzara et al. [39] presented a two-dimensional U-Net architecture for the segmentation of glioma tumor regions from MRI scans. The authors introduced a CNN using three different encoders to select features from MRI scans. Moreover, a decoder is also applied to upscale feature information after downscaling during the encoding process. Experiments are conducted on the BraTS 2023 dataset. An average Dice score of 0.92 for the whole tumor is observed during the testing phase. Kharaji et al. [40] developed an expanded nnU-Net architecture for segmenting pediatric and adult (glioma) tumors using the BraTS dataset. Their approach included using Hausdorff distance (HD) loss for boundary refinement, attention gates to highlight informative regions, and residual blocks to capture intricate spatial properties. Mean Dice 0.83 for gliomas and 0.71 for pediatric tumors was observed.

Due to small tumor size, irregular brain tumor morphology, and variability in the shape of tumor size, tumor segmentation is a difficult task. Ren et al. [41] developed an optimization methodology based on a 3D U-Net model to segment provided MRI scans in BraTS 2023's Challenges 1, 2, and 3—the methodology employed transfer learning and a variety of pre- and post-processing methods. The methodology yielded an average lesion-wise Dice score of 0.79, the highest among the other challenges on the validation dataset. Zhang et al. [42] utilized deep learning regularization with brain tumor growth Partial Differential Equation (PDE) models, which can be applied to any network model. These PDE models can be incorporated into the presented segmentation methodology, which improves accuracy and Dice score. The presented method employed a periodic activation function to estimate tumor cell density while aligning segmentation with biological behavior. The experiments were conducted on the BraTS 2023 dataset, and the highest Dice score of 0.9234 was achieved using the biophysics-inspired U-Net family of segmentation methods. The authors intended to integrate domain-specific knowledge into medical data to improve performance across different learning modes.

Glioblastoma is a very aggressive brain tumor that needs to be identified quickly and treated right away. It is usually difficult to develop an automatic detection technique due to its variability. The BraTS 2023 dataset is used to conduct experiments. To perform three-dimensional segmentation of Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT), Authors Yazıcı et al. [43] employed a multi-scale, attention-guided, hybrid U-Net model. They name it as GLIMS. Better feature aggregation and extraction are blocked by multi-scale feature extraction and the Swin Transformer. The attention-guided decoder extracted essential features. The presented technique achieved a Dice score of 0.9219 for the whole tumor.

It is essential to detect and identify brain tumors to treat them correctly. Various MRI modalities offer crucial information regarding brain tumors. While conventional fusion methods generate noise and decrease performance, unimodal models have low predictive capability and unpredictable performance. To address this issue, Dunyuan et al. [44] presented a multi-modal learning strategy for MRI brain tumor grading that incorporates dual attention and cross-modality guiding. The presented model employed dual attention to capture semantic interdependencies in spatial and slice dimensions using ResNet Mix Convolution for extraction of features. The classification accuracy, sensitivity, and specificity were 97.9%, 100%, and 97.1%, respectively, for the BraTS 2018 dataset. The same measures were 97%, 95.8, and 97.4, respectively, for the BraTS2019 dataset. In the future, a lightweight model will be adopted to unveil more essential features. Hapsari et al. [45] created a deep learning-based approach named en-CNN, which is based on a pre-trained VGG-16 model and consists of seven convolutional layers, four ReLU layers, and four max pooling

layers. The authors used the en-CNN method on T1, T1CE, T2, and FLAIR MRI sequences after preprocessing and image augmentation. The en-CNN approach on the BraTS 2018 dataset obtained 95.5% accuracy for T1 and T1CE, 94% for T2, and 97% for FLAIR. Content-based image retrieval (CBIR) made image processing easier by automating contrast recognition. However, a technique is needed that can classify the modality of an MR scan based on contrast. In order to classify MR image contrasts, authors Samuel et al. [27] suggested a three-dimensional deep convolutional neural network (CNN) that can distinguish between T1-weighted, T2-weighted, fluid-attenuated inversion recovery (FLAIR) contrasts, and pre-and post-contrast images of MR. The model was evaluated on 1281 images and trained on 2137 images using 3418 image volumes from various sites. Its accuracy was 97.57%. The dataset was collected from multiple sources. The authors indicated an intention to extend their work in the future to examine how it performs across multiple classes.

Zahid et al. [46] presented a deep learning-based method to classify FLAIR, T1, T2, and T1CE modalities. After normalizing the dataset (i.e., BraTS 2018), the Pre-Trained ResNet-101 model was used to refine its classification of brain tumors via transfer learning. The problem was the production of redundant features, which increased computational overhead and reduced accuracy. Particle swarm optimization (PSO) and differential evaluation were employed to identify the optimal characteristics for addressing the problem. After combining these features into a single vector, principal component analysis (PCA) was applied to optimize it. To classify tumors, the optimized vector was fed into multiple classifiers, yielding the highest accuracy of 94.4%. Khan et al. [9] presented a deep learning-based technique that constitutes a contrast stretching mechanism using a histogram-based method and application of discrete cosine transform (DCT), usage of VGG16 and VGG19 to extract the features using transfer learning of pre-trained models, feature selection mechanism using corr-entropy-based joint learning approach using extreme learning machine (ELM), merging of partial least square (PLS)-based covariant-based features into a single matrix, and application of ELM for classification. They used several datasets, such as BraTS2015, BraTS2017, and BraTS2018, and obtained accuracy of 97.8%, 96.9%, and 92.5%, respectively. The authors Thakur et al. [47] presented ED-ViTTL, which is a hybrid ensemble framework that merged a transfer-learned VGG19 CNN with five Vision Transformer variations (R50-ViT-L/16, ViT-L/16, ViT-L/32, ViT-B/16, and ViT-B/32) for global and local feature learning. They classify brain MRIs using feature embeddings from both branches, and they were combined via fully connected layers. Stratified sampling was used to divide the 3,264 MRI scans in the public dataset into training, validation, and testing subsets for the evaluation of the model. The study used extensive augmentation, class-weighted categorical cross-entropy, and stratified fivefold cross-validation to address class imbalance. A class-specific AUC value above 0.99, and the strongest ensemble (ViT-B/32 + VGG19) attained 98.67% accuracy. The confusion matrix results showed very little misclassification. The next section present the proposed methodology in detail.

Proposed methodology

The proposed brain tumor segmentation and classification framework is presented in this section, supported by mathematical and visual illustrations. Figure 1 illustrates the detailed proposed framework, which includes two architectures: the segmentation architecture and the classification architecture. In the first part, an optimized DeepLabV3 + + architecture was designed, and segmentation was performed. In the second part, a novel network-level fused architecture is designed based on the fire modules and is used for classification. The details are provided in the relevant subsection.

Dataset

The Brain Tumor Segmentation (BraTS) Challenge 2023 [48–50] dataset contains four modalities of magnetic resonance imaging (MRI) known as FLAIR, T2 weighted, T1 weighted, and T1CE (T1 with contrast enhancement). The MRI scans are collected from multiple institutions under standard clinical settings. There are seven sub-datasets, which differ in many facets, such as overall brain shape, overall appearance, and morphological aspects of brain tumors. In this study, the BraTS-GLI Glioma dataset is used. The dataset comprises MRI scans of 1251 patients for training, and data from 219 patients are reserved for validation. To perform classification, the dataset is further preprocessed, and 40,062 images are carefully selected from each modality. Only images that provide the most information to the learning model are selected; a total of 160,248 images are selected for classification. The same dataset is used for segmentation purposes. Images from the FLAIR class are used for segmentation, and their corresponding ground truth is selected. Images and ground truth are provided as input to the segmentation model. A few sample images for the segmentation and classification tasks are given in Fig. 2.

The BraTS dataset includes four MRI modalities (i.e., T1, T2, T1CE, and FLAIR). Only the FLAIR modality provides voxel-wise tumor ground truth from the dataset publishers. The other modalities cannot be used for supervised segmentation training or evaluation because

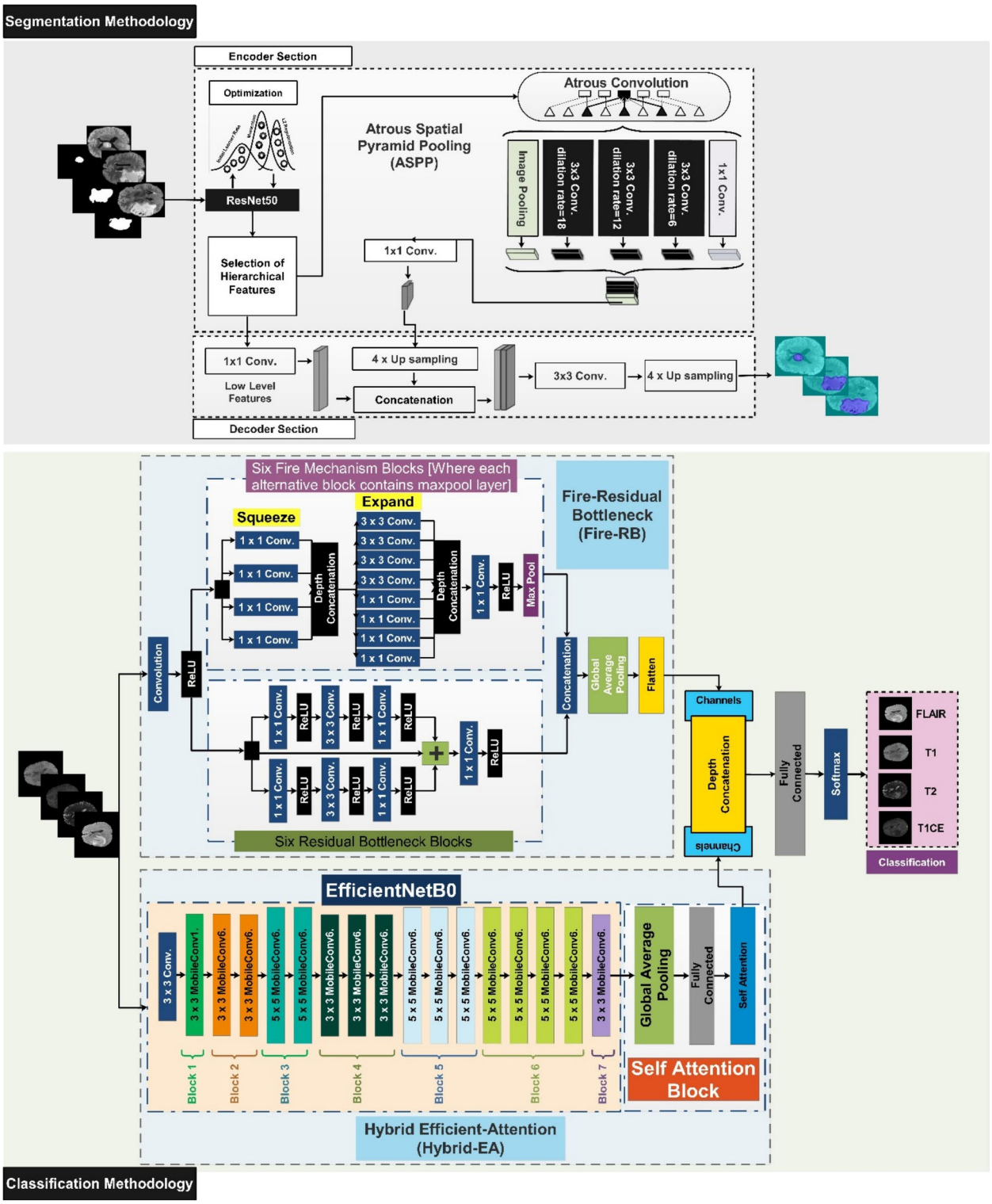


Fig. 1 Detailed framework of proposed segmentation and classification of brain tumor architectures

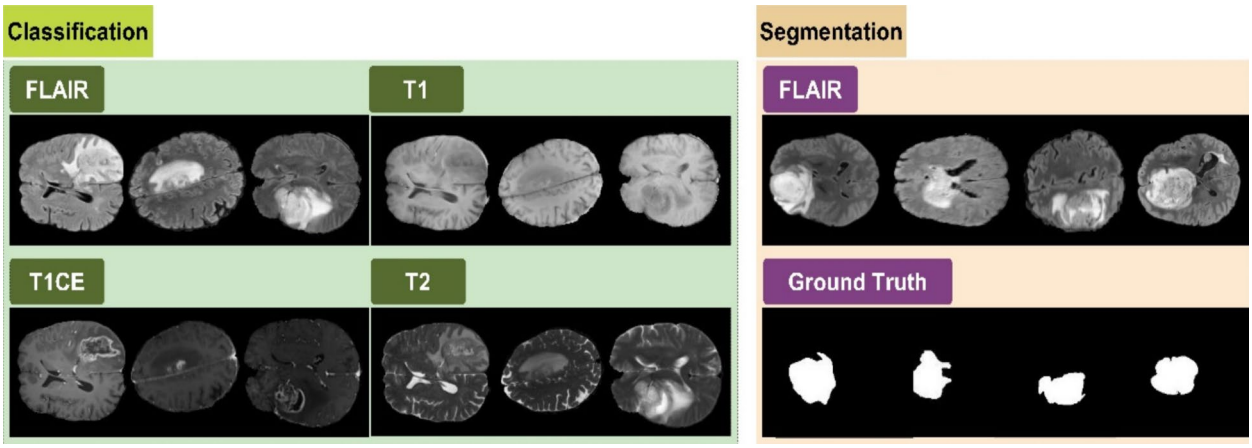


Fig. 2 Sample scans of the BraTS2023 dataset for the classification and segmentation tasks. FLAIR images paired with corresponding ground truth images are selected for the segmentation task

they lack segmentation labels. SO, only FLAIR images are employed for segmentation in this work. However, the classification component uses all four modalities. While classification determines the modality type of MRI slices, segmentation locates tumor areas within labeled FLAIR scans. Both tasks function independently and fulfill distinct therapeutic goals.

Novelty 1: proposed optimized DeepLabV3 + + based tumor segmentation

DeepLabv3 + architecture [51] is used to segment the affected region (i.e., brain tumor). The encoder and decoder are the two components that make up the architecture. The backbone architecture's (i.e., ResNet-50 [30]) training hyperparameters are optimized using Bayesian optimization in the encoder phase of the suggested technique. Following the extraction of hierarchical features for training, the Atrous Spatial Pyramid Pooling (ASPP) module gathers data at various levels and dilation rates [52]. Subsequently, the features are combined. Skip connections and image pooling are employed to further process the combined and hierarchical features produced by the ASPP module. At this point, more precise contextual information has been down-sampled from the image.

The image from the down-sampled encoder step is restored in the decoder via up-sampling. For this, deconvolutions are employed [53], resulting in a semantically segmented mapped image. Further details on the proposed optimized DeepLabv3 + are provided in the subsequent section.

Bayesian-based optimized DeepLabv3 +

The pre-trained ResNet-50 model serves as the foundation for the DeepLabv3 + architecture, which uses the dataset to acquire hierarchical features. The training method determines the quality of the feature extraction process. Bayesian optimization is used to optimize the hyperparameters [54] during the training process of ResNet-50. Initial learner rate, momentum, and L2 regularization are the optimized training hyperparameters. A range of values against each hyperparameter is provided, and selection is done in an optimized way by selecting the best corresponding value against each hyperparameter.

The features extracted from ResNet-50 are then sent to the ASPP module. The ASPP module uses parallel Atrous convolutional modules with different dilation rates to enable the network to collect global as well as local contextual data of an image using its large field of view. This allows the network to acquire information at different scales.

The features are incorporated for additional processing after being obtained at different scales in the previous step. Skip connections are used to combine these with previously extracted features with fine details in order to enhance multi-scale features. Additionally, scaled versions of the dataset's original photos are concatenated with characteristics derived from the ASPP module to perform image pooling. In order to produce finely segmented mapped images later on, fine-grained information is retained at this step by skipping the connection and image pooling process. So far, the image has been down-sampled.

The decoder section performs up-sampling because, in the initial stages of down-sampling, the merged features from the preceding phase possess smaller spatial resolution from the input image. A thorough segmentation map with the exact resolution as the original image can only be created by restoring the missing details. DeepLapv3 + 's decoders use transposed convolutions to boost spatial resolution throughout this procedure.

Using the improved attributes gleaned from earlier phases, a distribution of probabilities for each pixel was generated in the last step, which predicts each pixel's affiliation with a particular class. In order to do this, a convolutional layer with activated softmax is employed. The result is a detailed segmentation map that displays the expected class for each pixel in the original input image. At the end, refined segmented mapped images are returned. The process is pictorially described in upper portion of Fig. 1.

Proposed classification architecture

In this work, we proposed a novel network-level fused deep architecture for the classification of brain tumors into relevant classes such as T1, T2, T1CE, and Flair. The proposed classification architecture consists of two architectures––Customized Fire-Residual Bottleneck (Fire-RB) and Hybrid Efficient Attention (Hybrid-EA). The proposed Fire-RB architecture comprises two separate models: one model implements a fire mechanism, and the second model implements residual bottleneck architecture. These are further concatenated in the second dimension to increase the model capacity to capture complex patterns in the width dimension [31]. The features are then passed to subsequent layers for further processing. The Hybrid-EA model is an enhancement of the EfficientNetB0 pre-trained model. Features extracted from the pre-trained model are passed to the self-attention layer, which applies to its attention mechanism, and the features are then passed to the depth concatenation layer. Features obtained from these two methodologies are concatenated in channel dimension (i.e., Depth concatenation). The numbers of classifiers are employed to classify features, and final classification results are obtained. Lower portion of Fig. 1 constitutes the completion classification methodology.

Novelty 2: proposed customized fire-residual bottleneck (Fire-RB)

The proposed Fire-Residual Bottleneck (Fire-RB) architecture is shown in Fig. 1's classification portion. This proposed Fire-RB architecture consists of 198 layers and a total of 42.8 million learnable parameters. The model takes input of size $240 \times 240 \times 3$. The input is then passed to the convolutional layer; after that, a ReLU layer is used to introduce non-linearity. At this point, the model is subdivided into two portions; first model uses the fire mechanism (explained in the later section). The 1^{st}, 4^{th} and 6^{th} fire modules have *max pool* layers at their ends. The layer performs down-sampling [55] and reduces variability by providing the largest value from a group of T activations and can be represented mathematically as follows:

Let the feature map be $X \in \mathbb{R}^{H \times W \times L}$, the L denotes the number of channels. A kernel of max pooling with size T moves across the spatial dimension with stride represented with S. There are L^{th} related filters for a max pooled operation (g) can be represented with $q_g = \left[q_{1,g}, \cdots, q_{l,g}, \cdots, q_{L,g},\right] \in T^L$:

$$q_{l,g} = max\left(i_{l,(g-1)S+t}\right), \tag{1}$$

where the pooling shift or stride $S \in \{1, \ldots, T\}$ permits overlap between pooling regions where $S <$ T. The dimensionality is now decreased from the exiting number of convolutional bands to. The pooling layer decreases the output dimensionality from U convolutional bands to $M = \frac{(U-T)}{S+1}$ pooled bands, and the resulting layer is $q = [q_1 \ldots, q_m] \in T^{M \cdot J}$. It is observed that if convolution layers are down-sampled early then convolutional maps will be smaller, conversely if down-sampling occurs late in a deep neural network then classification accuracy will be better. The same concept is adopted by the fire module, and more *maxpool* layers with *stride* $= 2$ are added late in the model for better classification accuracy.

Proposed fire modules-based block The fire modules are the core building block of the SqueezeNet architecture. They help lower learnable parameters by maintaining reasonable accuracy thresholds [56]. A fire module consists of a squeeze-and-expansion mechanism. In the squeezing portion of the modules, the convolution layer must have 1×1 filters, which leads to point-wise convolution. The point-wise convolution helps in dimensionality reduction [57], rich and expressive feature representation [58], control over the complexity of learned representations [59], and integration with spatial or channel-wise convolutions [60]. The number of layers in the squeeze portion must be less than the expanded portion. The squeeze portion comprises four 1×1 convolutional layers, which limit the number of input channels to expand the portion, thereby keeping the number of parameters manageable and maximizing overall accuracy. The output of the squeeze layer is concatenated depth-wise. The expanded portion is the combination of 1×1 and 3×3 filters. The proposed methodology comprises eight convolutional layers, with half using 1×1 filters and the other half using 3×3 filters. The output of the expanding layer is also concatenated depth-wise and passed to subsequent layers for further processing. Figure 3 presents the architecture of a single fire module.

Proposed residual bottleneck module The second model uses residual bottleneck architecture (explained in a later section) to process the input feature maps; adaptation of the bottleneck approach with skip connection helps

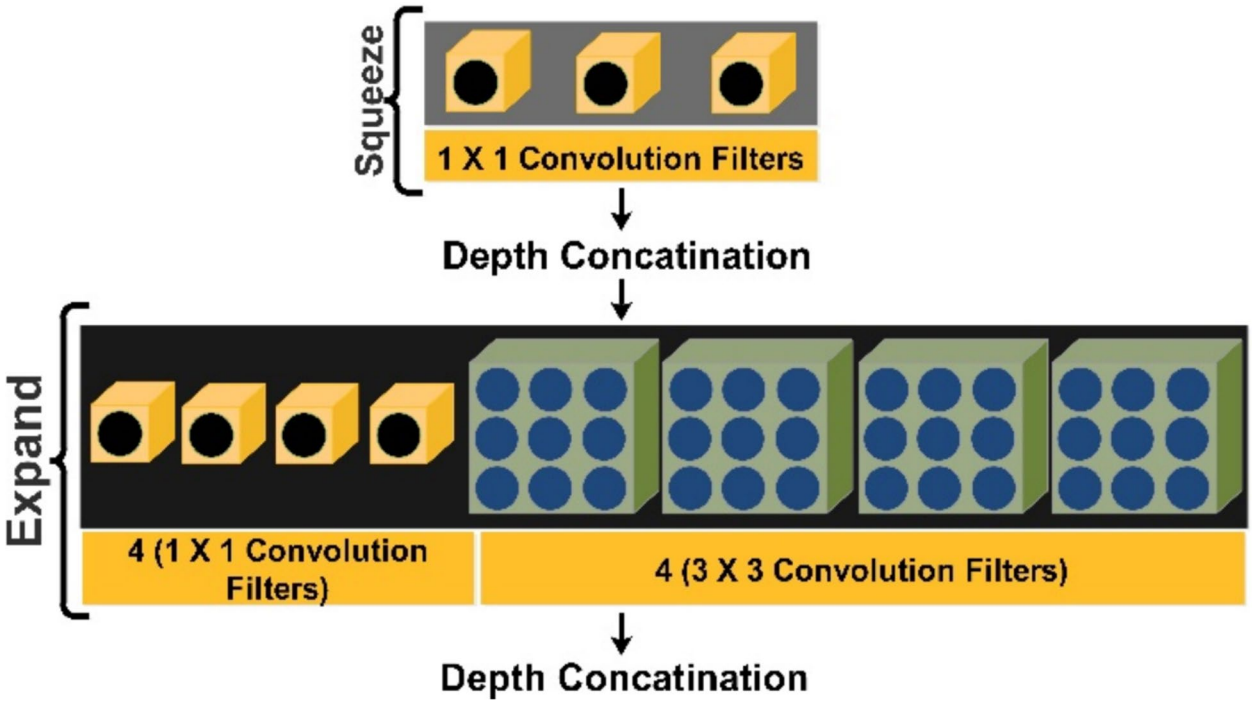


Fig. 3 Illustration of a fire module

avoid vanishing gradients during backpropagation. Due to the presence of multiple hidden layers, deep neural networks face the vanishing gradient problem. Gradients in early layers become very small during training, which makes it difficult for optimization algorithms such as stochastic gradient descent (SGD) to update these layers efficiently. It causes stagnation or slow convergence. Gradients decrease during backpropagation, which leads the weights to approach 0. Hence, early layers cannot update parameters effectively [61].

Residual architectures in deep neural networks solve the vanishing gradient problem [62] by introducing skip connections. They permit gradients to flow directly during backpropagation. By providing alternative gradient flow pathways, these links enable efficient parameter updates in the early stages [25]. Deep neural network training is streamlined by bottleneck architecture, popularized by models such as ResNet. Feature maps are first compressed via 1×1 convolutions, then essential features are extracted via 3×3 convolutions, and finally, the feature maps are expanded to their original dimensions. It also employs a skip connection to mitigate vanishing gradients. This method is the foundation of modern-day deep learning since it maximizes performance and computational resources [30]. In Fig. 1, the lower part of the classification methodology's Fire-RB block represents the Residual Bottleneck architecture.

The outputs of both models are concatenated along the second dimension, yielding a feature vector of size $4\times8\times512$, where the number of channels (i.e., 512) is the best feature map shared by both models. The final feature vector of $1\times1\times512$ is obtained at the pool layer. Classification is performed on features extracted from the pool layer.

Novelty 3: hybrid efficient attention (hybrid-EA)

The Hybrid-EA architecture enhances pre-trained EfficientNetB0. As shown in the classification part of Fig. 1's Hybrid-EA portion. Seven blocks are added from EfficientNet-Bo, each ending with a Global Average Pooling layer and a Fully Connected Layer. The model accepts the dimension $224 \times 224 \times 3$ as an input. There is a total number of 292 layers having 6.9 Million learnable parameters. A self-attention [34] layer is added after the FC layer. The resultant vector at *thepool* layer was $1 \times 1 \times 1280$. The self-attention layer manipulates the features by applying an attention mechanism.

Self-attention Vision transformers use self-attention as their fundamental element [63]. It broadens the receptive field of the convolutional neural network without increasing the computational burden imposed by huge kernel sizes. The model can use a self-attention layer to dynamically focus on relevant regions by varying receptive field

sizes. [64]. Additionally, using self-attention in the final layers has a potent impact on feature learning [65].

The input image from the dataset is divided into three dimensions named height, width, and channel, which are split up into distinct vectors. The height vector is transposed, and the scalar dot product of the width and the transposed height is used to get similarity scores. The highest-probability weights are obtained by applying the softmax function to the scalar dot product, yielding normalized attention weights. Normally, attention weights with the highest probabilities are used to augment the channel vectors. The self-attention feature matrix is obtained in the conclusion. Figure 4 represents functionality pictorially.

The architecture can be depicted mathematically as follows: Suppose the input feature matrix is

$$b \in \mathbb{R}^{C \times N} \tag{2}$$

where $C = Numberofchannels$, $N = H \times W$. Three convolution-based linear projections are applied in order to generate transformed feature spaces, and are represented by $j(b), k(b)\&O(b)$:

$$j(b) = P_j b, \tag{3}$$

$$k(b) = P_k b. \tag{4}$$

In the equations, P_j and P_k represents learnable weight matrices,

$$P = Weights. \tag{5}$$

Projection value is defined as follows:

$$O(b) = P_C b, \tag{6}$$

with

$$P_j, P_k, P_C \in \mathbb{R}^{C^* \times C} \tag{7}$$

After that, the attention scores are calculated by applying a row-wise Softmax in Eq. 11. The normalized attention weights for spatial positions v and w are

$$R_{v,w} = \frac{e^{Y_{v,w}}}{\sum_{v'=1}^{N} e^{Y_{v',w}}}. \tag{8}$$

In the next step, the similarity matrix is obtained by

$$Y_{v,w} = j(b)_v^T k(b)_w, \tag{9}$$

And further, equivalently, using the projection notation,

$$Y_{v,w} = (P_k b)_v^T (P_j b)_w. \tag{10}$$

The aggregated output feature at position w using these normalized attention weights is provided by

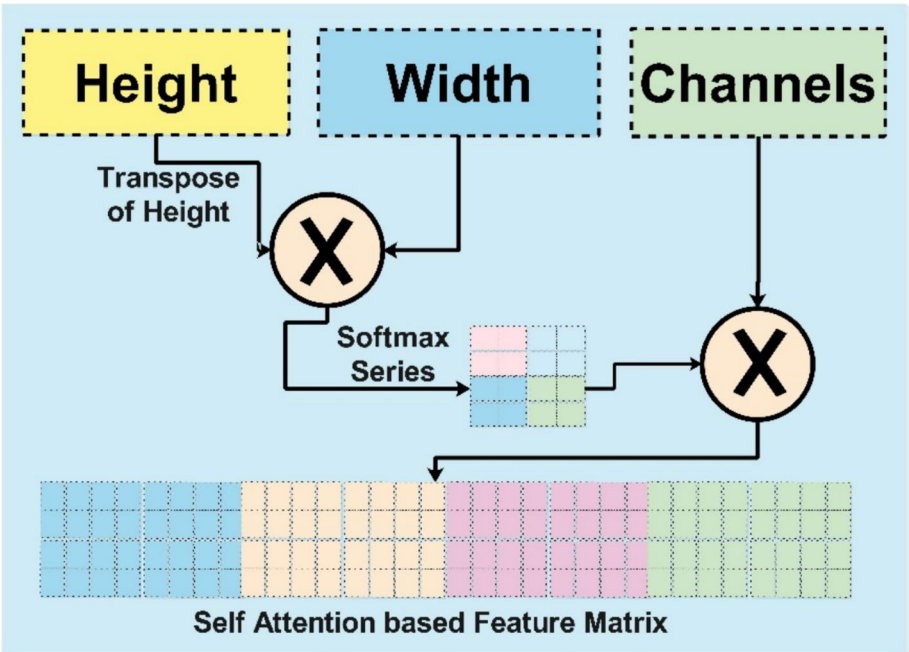


Fig. 4 Self-attention mechanism used in this work

$$U_w = S\left(\sum_{v=1}^{N} R_{v,w}.O(b)_v\right), \tag{11}$$

where Softmax (S) is applied to normalized the weighted combination of value vectors. Hence, the attention-based feature map is defined as follows:

$$S_b = \mathbb{R}^{C\times C^*}, \tag{12}$$

The final output has the same number of channels as the features that were used as an input into the self-attention layer.

Novelty 4: networks fusion using depths

A combination of diverse features and the fusion of information via depth concatenation yields multimodal learning. A variety of tasks can be performed by combining information from other sources [66]. To identify intricate patterns and symbolic power, depth concatenation expands the model's capacity by merging feature maps from different layers of the neural network [67]. Depth concatenation can serve as a regularization technique, yielding resilient and generalizable feature representations from various sources and thereby promoting feature diversity [68]. Original gradients and feature values are preserved, rather than adding or averaging the feature vectors. So an efficient backpropagation is achieved using depth concatenation [69]. The chance of overfitting is decreased by sharing learnable parameters effectively, which ultimately results in increased efficiency [70].

Depth concatenation preserves the whole feature space of both Fire-RB and Hybrid-EA without imposing linear-combination constraints. It is in contrast to the element-wise summation or weighted fusion. Complementary representations may be suppressed using summation-based approaches. It requires both branches to provide feature maps with identical semantic distributions. Weighted fusion introduces additional learnable factors that may reduce generalization and bias the network toward the dominant branch. Depth concatenation maximizes representational diversity, minimizes additional computing load, and preserves the original gradient flow. This decision is vital because Hybrid-EA employs self-attention to capture long-range dependencies, whereas Fire-RB generates fine-grained local descriptors. The classifier can leverage complementary information that cannot be linearly combined without loss by concatenating heterogeneous features.

In this work, the extracted features from the proposed Fire-RB are 512 and 1280 from Hybrid-EA, respectively. The Fire-RB model's features are extracted from the flattening layer, whereas the self-attention layer is employed in the Hybrid-EA architecture for feature extraction. Channel-wise depth concatenation (i.e., fused model) is performed, and a feature vector of dimension Nx1792 channels is obtained. Each channel represents a specific feature map of input data. The mathematical representation of depth concatenation is given as follows:

Suppose the feature vector A can be represented $(.)byheight, width, andchannels$, and feature vector B can be represented$\left(byheight, width, andchannels_{(.)}\right)$. The final feature vector after depth concatenation is represented as$\left(Height, Width, Channels_A + Channels_B\right)$:

$$C = A \parallel B. \tag{13}$$

The final features of this step are passed to the fully connected layer, as shown in Fig. 1's classification step depth fusion. After that, the model training is performed, whereas the hyperparameters are selected using Bayesian Optimization (BO) [54].

Proposed architecture testing

During testing of the proposed CNN architecture, features are extracted from the fused network, including the depth concatenation layer and a feature vector of size Nx1792. Extracted features are passed to classifiers, such as wide neural networks and other models. The best classifier is selected based on accuracy and precision, and its performance is further evaluated using visual output. Figure 5 shows the visual output of the Wide Neural Network classifier based on the returned labeled image. The numerical and interpretation results are discussed in the results section.

Results and discussion

This section discusses the implementation of the proposed architecture, presented in both numerical and visual forms. The proposed segmentation results are presented as both numerical and visual MRI-marked images. The classification results are presented as numerical and confusion matrices. In addition, the explainable AI results are also presented in this section.

Experimental setup

The experiments are conducted using MATLAB 2025a on a NVIDIA GeForce RTX™ GB GPU. Stochastic gradient descent with momentum (SGDM) is employed to accelerate the gradient update and improve convergence. To conduct experiments, k-fold cross-validation is used with $k=10$. The ratio of the training and testing sets is determined. The maximum number of epochs is the number of times a model or algorithm is trained on a dataset. A subset of the dataset is used as the mini-batch size during the algorithm's learning process; it is set to 64, and the initial learning rate is 0.001. For the segmentation methodology, the ResNet-50 backbone is used,

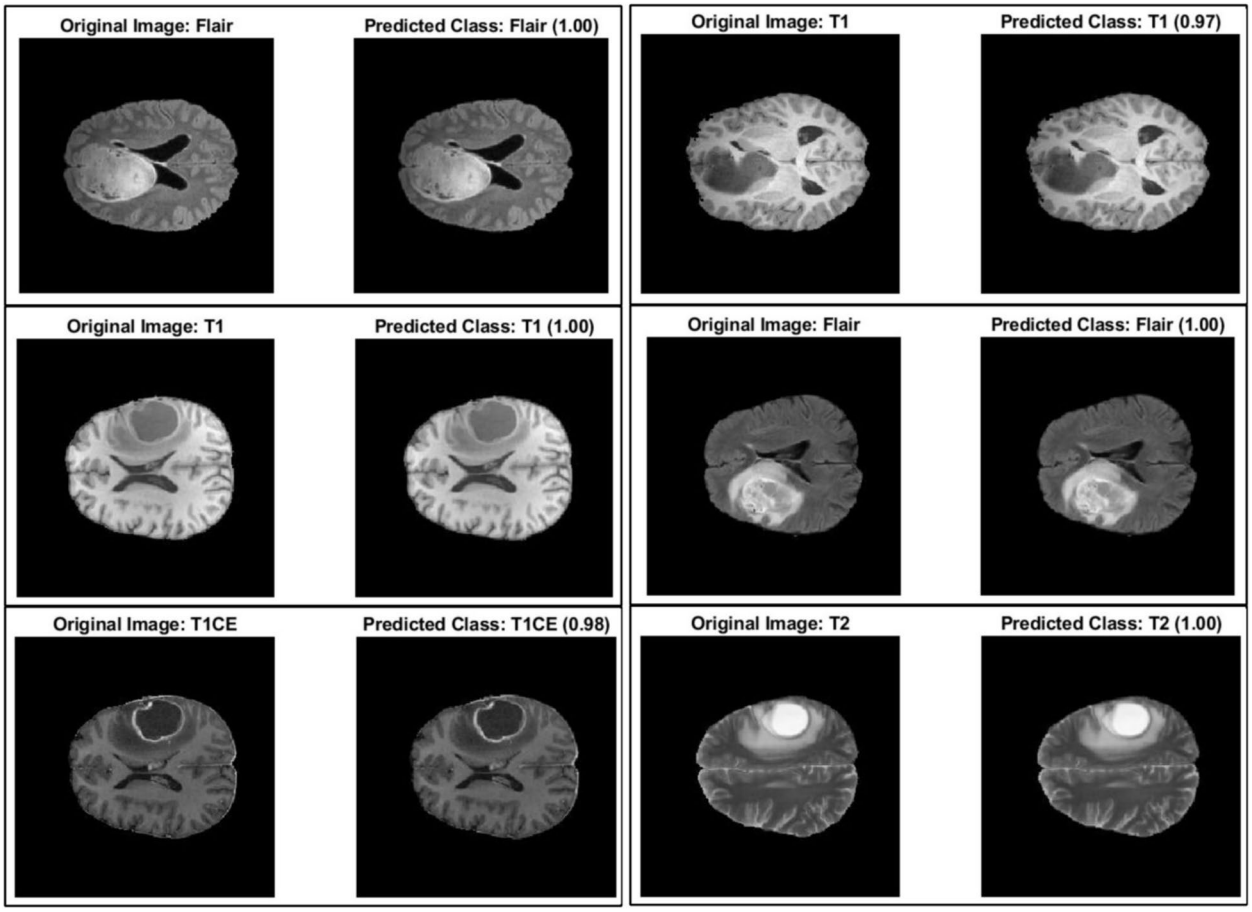


Fig. 5 Proposed SMDeepNet visual prediction using MRI scans

Table 1 Selected performance measures for this work

Name	Performance measure
Percentage accuracy (%)	$\frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative}$
Time	Measured in seconds
Percentage rate of sensitivity (%)	$\frac{TruePositive}{TruePositive+FalsePositive}$
Percentage rate of false (%)	$\frac{FalseNegative}{TruePositive+FalsePositive}$
Percentage rate of precision (%)	$\frac{TruePositive}{TruePositive+FalseNegative}$
Area under curve (AUC)	$\int_{i}^{j} g(x)dx$
DICE score	$DICE = \frac{2\times\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert}$
Intersection over Union (IoU) / Jaccard Index	$DICE = \frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$

with a maximum of 30 epochs and a mini-batch size of 8. The optimal values of the initial learning rate, L2 regularization, and momentum are selected via Bayesian optimization. A 50:50 split is used for the training and validation datasets.

Dataset and performance measures

Experiments are conducted for classification and segmentation using the publicly available MICCAI BraTS 2023 brain tumor MRI datasets. Various performance measures are employed to evaluate the experimental results. These metrics include the area under the curve, accuracy, false negative rate (FNR), and sensitivity. Accuracy, DICE score, and intersection over union (IoU) or the Jaccard index are used to evaluate the segmentation methodology's performance. Table 1 presents the performance measures used to measure the results of this work.

Results of proposed segmentation methodology

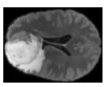
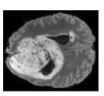
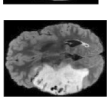
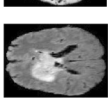
Implementing segmentation as part of the methodology yields results from the experiment. The performance measures used to measure MRI image segmentation for brain tumors are accuracy, DICE score, and Intersection over Union (IoU). Accuracy indicates the overall

correctness of the underlying experiment. The DICE score directly assesses the degree of overlap between the ground truth and the predicted tumor regions. It is beneficial for determining segmentation quality because it is sensitive to both accurate identification of tumor regions and reduction of false positives and negatives.

The IoU provides an alternative measure of segmentation performance by offering an evaluation comparable to the DICE score but using a different penalty mechanism. It is a robust metric because it emphasizes appropriate overlap between the predicted segmentation and the ground truth, penalizing both false positives (i.e., pixels missegmented) and false negatives (i.e., missing pixels). Table 2 presents a few segmentation results for selected performance measures. Image 1 achieves the highest segmentation accuracy (0.9941); image 3 achieves the highest Dice score (0.9401), and the same image also achieves the highest IoU (0.8920). The overall average IoU is 90.16%, the average accuracy is 95.32%, and the Dice Score is 93.54%. A more detailed discussion of the segmentation results is given in Discussion section.

After receiving input such as a FLAIR image and its ground truth, the segmentation process proceeds through several phases. Each phase produces an image as output. Figure 6 represents an original image, its ground truth, a boundary along the perimeter of the ground truth, a predicted semantic segmentation-based image, a model-based prediction around the ground truth, and a combined image where a difference is apparent between the overlapping of the actual and predicted ground truth. The actual ground truth boundary is represented with yellow color, whereas the expected boundary is shown using magenta color.

Table 2 Representation of segmentation results for selected performance measures

Serial	Images	Accuracy	Dice score	Intersection over Union (IoU)
1		**0.9884**	**0.9401**	0.8920
2		0.9871	0.9385	0.8894
3		0.9841	**0.9498**	**0.9074**
4		0.9859	0.9346	0.8831
5		0.9904	0.9472	0.9036

Bold denotes the most significant values

Proposed architecture classification results

The classification results of the proposed fused network are presented in this subsection. The classification results are presented in three different experiments. In the first experiment, features are extracted from the proposed Fire-RB architecture, and classification is performed. In the second experiment, features are extracted from the proposed Hybrid-EA architecture, and the classification performance is obtained. In the last experiment, the entire fused architecture was used for classification.

Fire module architecture and residual bottleneck (Fire-RB)

Input data are provided to the model and processed through many layers of deep neural networks. Resultant features are extracted from the global average pool () layer, and five neural network classifiers are selected to perform the classification of features into relevant classes. The results are provided in Table 3. The results indicate that the wide neural network classifier (WNN) achieves the highest classification accuracy of 98.60%. The Medium NN (MNN) classifier took the lowest time (i.e., seconds) to execute for this experiment. The confusion matrix in Fig. 7 represents the correctly predicted values in the diagonal blue cells against each class. Misclassifications are represented in white cells against each class. The time of each classifier is also noted, and the MNN classifier shows a faster execution than the other classifiers, such as WNN, Tri-layered NN (Tri-NN), Narrow NN (Narr-NN), and Bi-Layered NN (Bi-NN), respectively.

Hybrid Efficient Attention mechanism (Hybrid-EA)

In the second experiment, the dataset is provided as an input; different layers perform relevant processing on the data at various levels of the deep neural network model. The feature vector at the self-attention layer is used to extract features. After feature extraction, a neural network-based classifier was applied to perform classification. The classification task assigns features to relevant classes. The results of this task are provided in Table 4. In this table, the WNN classifier achieves the highest accuracy of 98.50%, whereas the lowest time is 179.31 s for the MNN classifier. The confusion matrix for the WNN classifier is presented in Fig. 8, where correct predictions are shown in the diagonal blue cells, and incorrect predictions are shown in the white cells for each class. The time of each classifier is also noted, and the MNN classifier

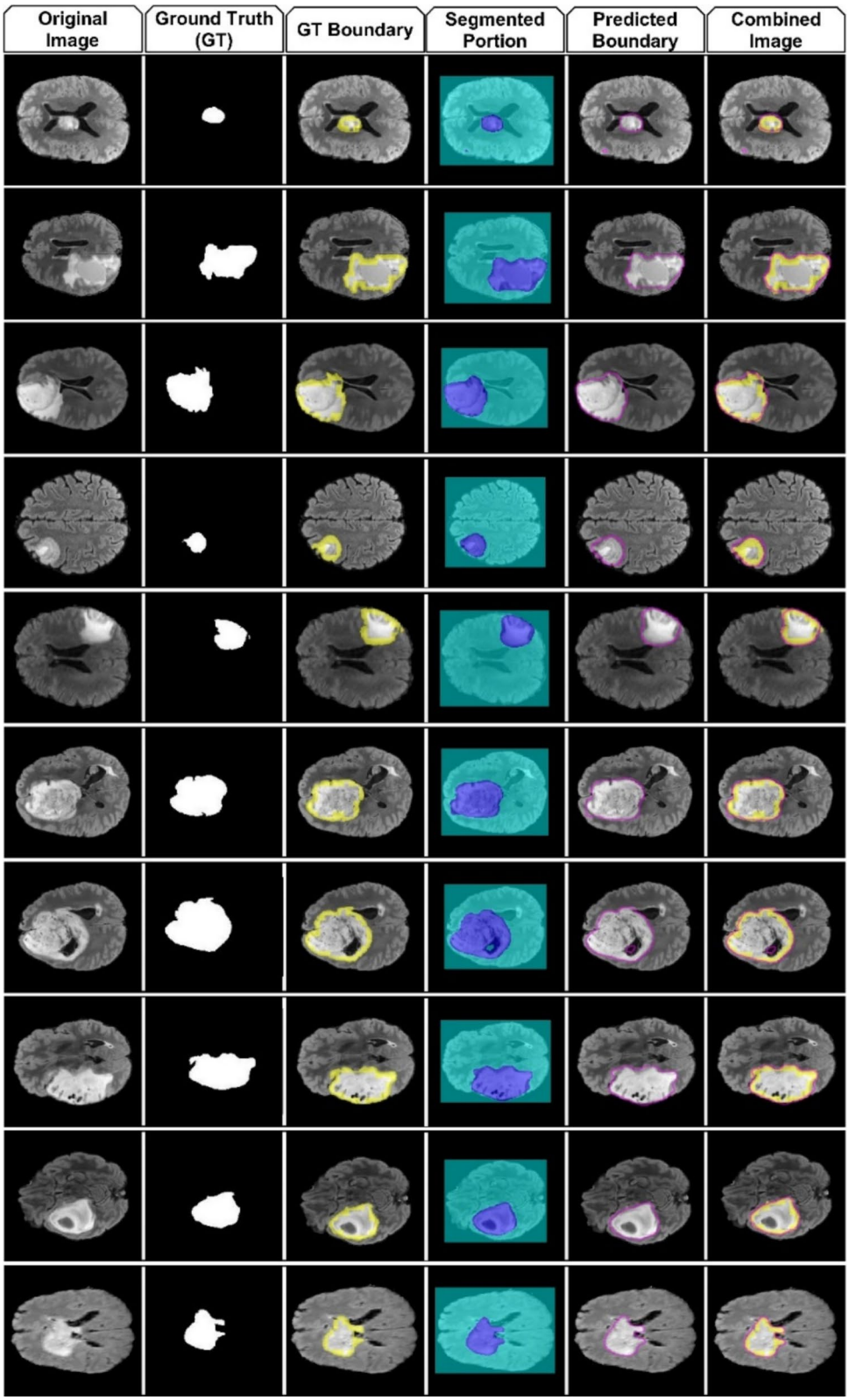


Fig. 6 Representation of original image, its ground truth, ground truth boundary, infected portion after segmentation, predicted boundary of tumor by model, combined image with predicted boundary and ground truth boundary

Table 3 Classification results for features extracted from proposed Fire-RB architecture

Classifier	Accuracy (%)	Time (sec.)	Sensitivity rate (%)	False negative rate (%)	Precision rate (%)	Area under curve (%)
Wide neural network	**98.60**	87.177	**98.63**	**1.38**	**98.63**	0.99
MNN	98.30	**68.89**	98.30	1.70	98.30	0.99
Tri-NN	97.90	205.03	97.85	2.15	97.88	0.99
Narr-NN	97.80	137.74	97.83	2.17	97.85	0.99
Bi-NN	97.70	137.06	97.72	2.28	97.73	0.99

Bold denotes the most significant values

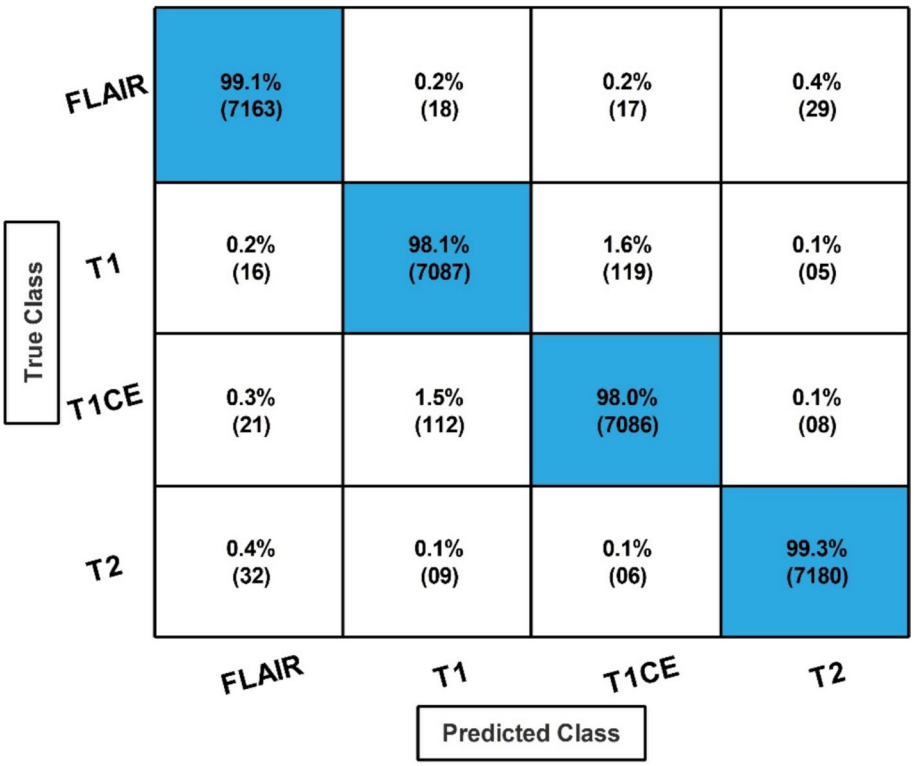


Fig. 7 Confusion matrix for the WNN classifier for features extracted from proposed Fire-RB architecture

Table 4 Classification results for features extracted from Hybrid Efficient Attention (Hybrid-EA) architecture

Classifier	Accuracy (%)	Time (sec.)	Sensitivity rate (%)	False negative rate (%)	Precision rate (%)	Area under curve (%)
Wide neural network	**98.50**	188.97	**98.55**	**1.45**	**98.53**	0.99
MNN	98.30	**179.31**	98.30	1.70	98.33	0.99
Narr-NN	97.80	276.36	97.80	2.20	97.80	0.99
Bi-NN	97.80	217.19	97.83	2.17	97.83	0.99
Tri-NN	97.70	275.28	97.75	2.25	97.75	0.99

Bold denotes the most significant values

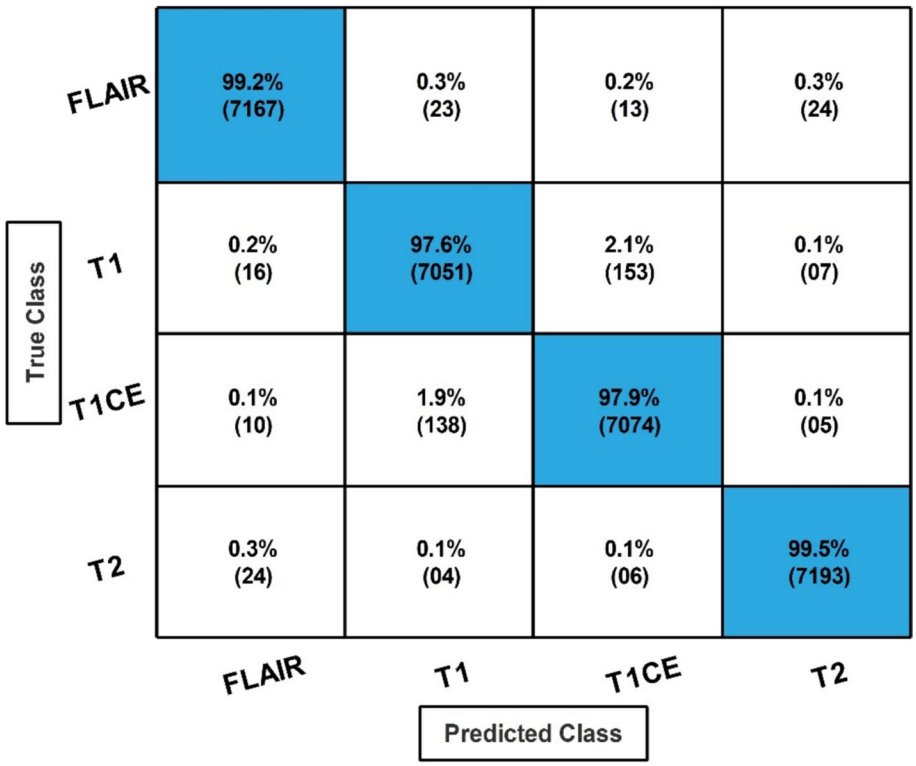


Fig. 8 Confusion matrix for wide neural network for features extracted from hybrid efficient attention (Hybrid-EA)

consumed less time than the other listed classifiers, such as 179.31 (sec). When comparing this architecture's performance with the proposed Fire-RB, the Hybrid-EA consumed more time and showed a minor decrease in accuracy and precision.

Proposed network fusion results

The Fire-RB and Hybrid-EA models are fused using a Network-level fusion technique, and a trained model is used for feature extraction. Features are extracted from the depth concatenation layer, where the channels of both models are fused, yielding numerous feature maps. After feature extraction, neural network (NN) classifiers are employed to perform classification. The results of this experiment are presented in Table 5. Based on the results in this table, WNN achieved an accuracy of 99.20%, which is higher than the pre-fusion model accuracy. However, there has been only a slight increase in the classifiers' computation time. The minimum observed time is 206.52 (sec) for the WNN classifier, whereas in previous experiments it was 87.177 (sec) and 188.97 (sec), respectively. The precision of this fused model is 99.18%; however, it was previously 98.63% and 98.53%, respectively. Hence, the proposed fused architecture improved accuracy and precision. The confusion matrix of this experiment is illustrated in Fig. 9, which depicts the correct and incorrect distribution of observations to relevant classes.

Table 5 Proposed network-level fused CNN architecture classification results

Classifier	Accuracy (%)	Time (Sec.)	Sensitivity rate (%)	False negative rate (%)	Precision rate (%)	Area under curve (%)
Wide neural network	**99.20**	**206.52**	**99.18**	**0.82**	**99.18**	0.99
MNN	99.00	217.55	99.00	1.00	99.00	0.99
Bi-NN	98.90	393.50	98.90	1.10	98.88	0.99
Tri-NN	98.90	371.60	98.85	1.15	99.87	0.99
Narr-NN	98.80	356.50	98.88	1.20	98.83	0.99

Bold denotes the most significant values

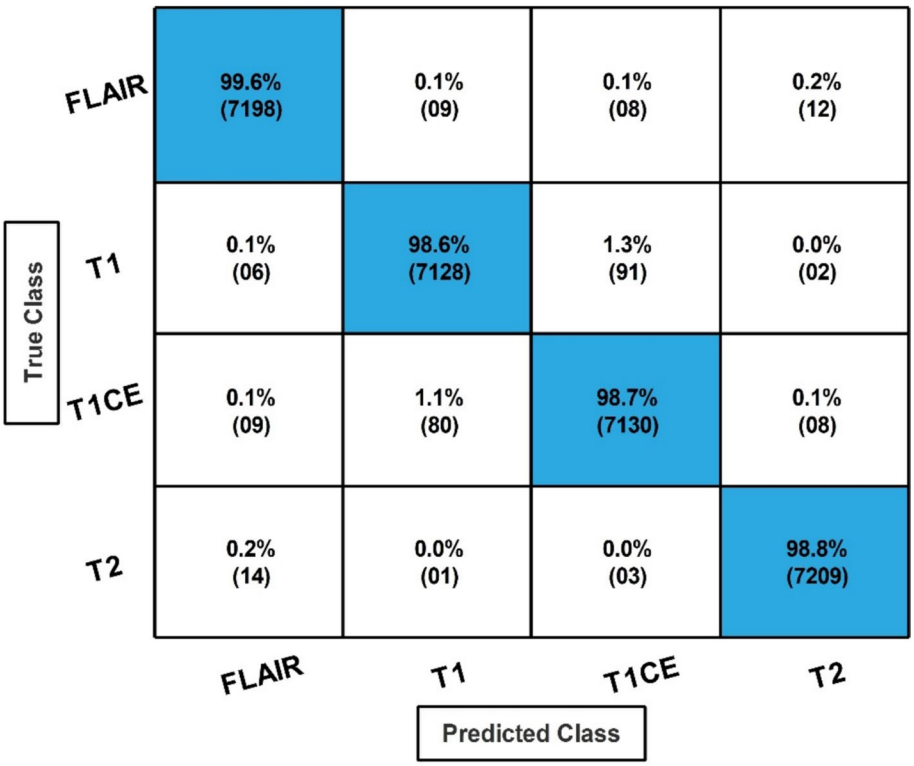


Fig. 9 Confusion matrix for wide neural network for network-level fused CNN architecture

Discussion

A detailed discussion of the proposed tumor segmentation and classification architecture has been presented in this section. Figure 1 shows the proposed segmentation and classification architectures, which comprise a few intermediate steps. The proposed segmentation architecture is illustrated in upper portion of Fig. 1, whereas the visual results are presented in Table 2 and Fig. 6. The proposed classification network-level fused architecture is shown in the lower portion of Fig. 1, whereas the results are presented in Tables 3, 4, 5 and Figs. 7, 8, 9.

Performance in segmentation and classification is affected by several practical issues in clinical brain MRI that extend beyond the architectural design. First, there are significant differences in contrast, tissue visibility, and pathological sensitivity between FLAIR, T1, T2, and T1CE MRI scans. Learning across modalities is more challenging and introduces the possibility of feature mismatches among them. Second, noise, motion artifacts, and intensity non-uniformity are common in MRI images. These factors impair border definition and disproportionately affect diffused tiny tumor patches. Third, domain shifts caused by scanner-to-scanner variability (i.e., acquisition techniques, field strength, and coil designs) reduce deep network generalization. Fourth, due to limited pixel representations and unclear boundaries, irregular tumor regions, especially non-enhancing components, present additional challenges. Each of these problems is addressed by the ASPP-based DeepLabV3 + + encoder, which captures multi-scale contextual cues to mitigate noise-induced detail loss. The hybrid Fire-RB and Hybrid-EA fusion maintains both global attention representations and local texture-level features to stabilize performance across modality disparities, using Bayesian—optimized hyperparameters that enhance robustness to data heterogeneity. Future studies will investigate domain adaptation and noise-aware training to improve real-world applicability further. However, these techniques lessen sensitivity to actual MRI fluctuations. To further analyze the performance of the proposed architectures, we conducted detailed ablation studies.

Ablation studies

In the first ablation study, we compared the proposed tumor segmentation architecture with several pre-trained ResNet backbones, including ResNet18, ResNet101, and ResNet150. We employed ResNet architecture as the backbone and trained it; the results are plotted in Fig. 10. In this figure, segmentation accuracy is shown for two experiments: with BO optimization and without BO. Without BO, hyperparameters are selected, including

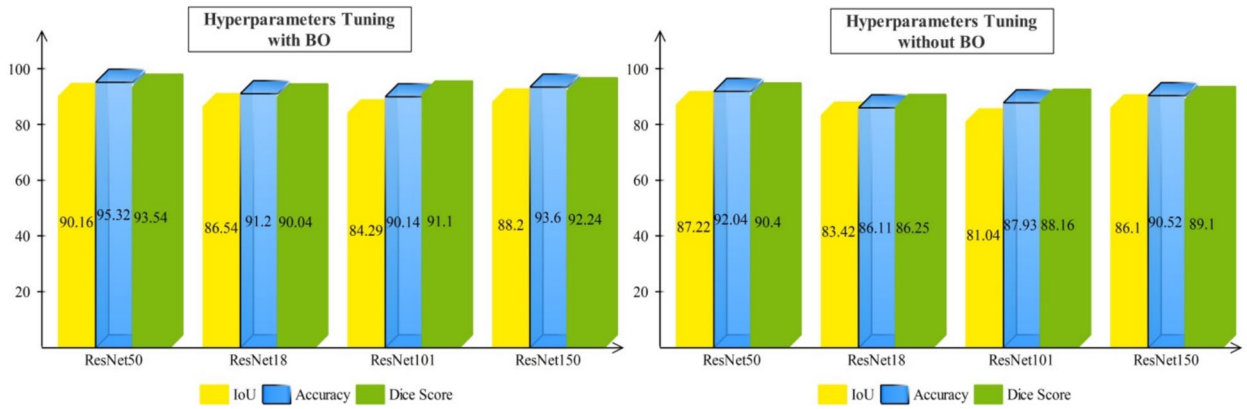


Fig. 10 Ablation studies-based analysis of proposed segmentation results using different backbones

a learning rate of 0.0001 and a momentum of 0.0564. In this figure, it is evident that IoU, accuracy, and Dice score improve when the ResNet50 model is used as the backbone. ResNet101 also achieves better performance, with an IoU of 88.22%, the second highest after ResNet50. In another experiment, models are trained using manual hyperparameter selection, and the accuracy is reduced by approximately 3–4%. Hence, from this ablation study, it is concluded that the ResNet50 architecture performed well as a backbone. In addition, hyperparameter selection via BO increased the final accuracy and IoU by nearly 3%.

In the second ablation study, we compared the proposed classification architecture with individual modules and with pre-trained CNN models, including Fire-RB, Hybrid-EA, ResNet50, ResNet101, InceptionV3, and MobileNetV2. Models are trained using different epochs (i.e., 10, 20, 25, and 30), and results are plotted (minimum and maximum). Figure 11 depicts the visual analysis.

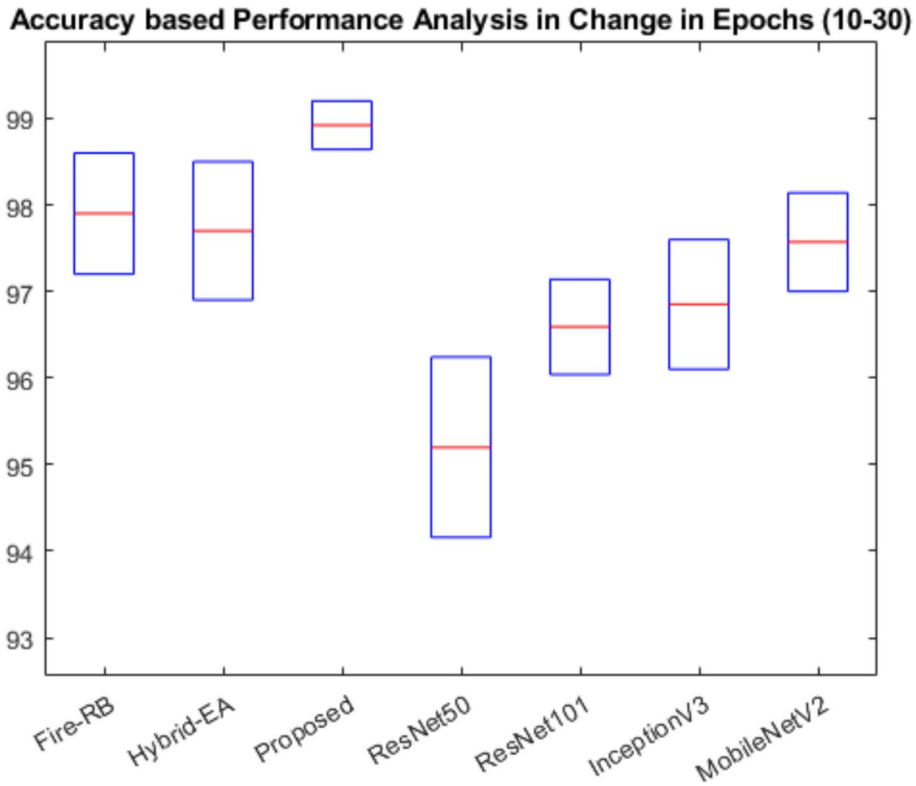


Fig. 11 Ablation study for the analysis of proposed and pre-trained CNN architectures

In this figure, Fire-RB achieved a minimum accuracy of 97.20% after 10 epochs and a maximum of 98.60% after 30 epochs. Similarly, the Hybrid-EA yields minimum and maximum accuracies of 96.90% and 98.50%, respectively. InceptionV3 obtained 97% accuracy after ten epochs; however, it was improved (98.14%) when executed up to 30 epochs. MobileNetV2 is another strong model for brain tumor classification, achieving an accuracy of 97% and 98.14% with 10 and 30 epochs, respectively. The proposed fused network achieved higher consistency, with an accuracy of 98.64% and 99.20% at 10 and 30 epochs, respectively. Hence, the proposed architecture outperforms the other listed deep models.

Generalizability analysis through confidence interval

To further assess the generalizability of the proposed custom CNN architecture, we used the trained BraTS2023 models to evaluate on BraTS2020 and 2021. Test data from BraTs2020 and BraTs2021 were used to obtain classification results. Table 6 presents the results of this experiment. For the BraTS2020 dataset, the proposed Fire-RB obtained an accuracy of 90.20%, whereas the Hybrid-EA and proposed fused obtained an accuracy of 91.94 and 93.50%, respectively. Variance, standard deviation, and standard error of the mean (SEM) are also computed from these values, namely, 1.8168, 1.3478, and 0.7782. These values are plotted in Fig. 12, which shows a 95% confidence level; the resultant margin of error is $91.88 \pm 1.525(\pm 1.66\%)$.

Similarly, testing is performed on the BraTS2021 dataset, yielding accuracy values of 89.16%, 90.80%, and 92.95%. Based on these values, the SEM is 0.8960, which appears somewhat high. Using SEM, the margin of error is computed at a 95% confidence level, and the resultant value is $90.97 \pm 1.756(\pm 1.93\%)$. Based on this analysis, the proposed architecture also performs better on unseen brain MRI datasets.

Table 6 Confidence interval-based analysis of the proposed CNN architectures using cross-datasets

Model and measure value	BraTS2020	BraTS2021
Fire-RB	90.20	89.16
Hybrid-EA	91.94	90.80
Proposed fused	93.50	92.95
Variance	1.8168	2.4084
Standard deviation	1.3478	1.5519
SEM	0.7782	0.8960

Model interpretation and comparison

The proposed fused classification architecture is further interpreted using explainable AI techniques such as LIME. The trained model is applied to the test images and yields labeled results, as shown in Fig. 13. In this figure, the prediction score is shown to be computed from the input image using LIME. Also, the critical region is highlighted with different colors. Based on the highlighted colors, the correct prediction has been made for the selected images. Finally, a comparison is conducted with more recent techniques, as presented in Tables 7 and 8. These tables indicate that the proposed segmentation and classification models achieve improved performance.

Ablation study 1: overlap stability analysis (OSA)

An OSA across five experimental configurations is performed in order to statistically assess the resilience of the suggested segmentation strategy. Key segmentation parameters, such as accuracy, dice coefficient, and intersection over union (IoU), are the focus of the study. The results of these are already presented in Table 2. A mean accuracy of 0.9872 ± 0.0024, a dice score of 0.9420 ± 0.0060, and an IoU of 0.8951 ± 0.0093 are attained by the suggested method. The highest

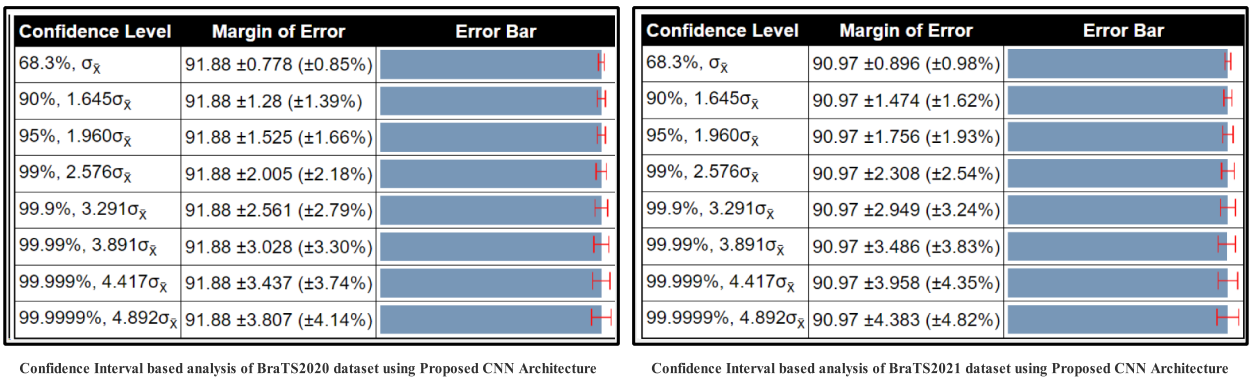


Fig. 12 Confidence Interval (CI)-based analysis of proposed CNN architectures using BraTS2020 and BraTS2021 datasets

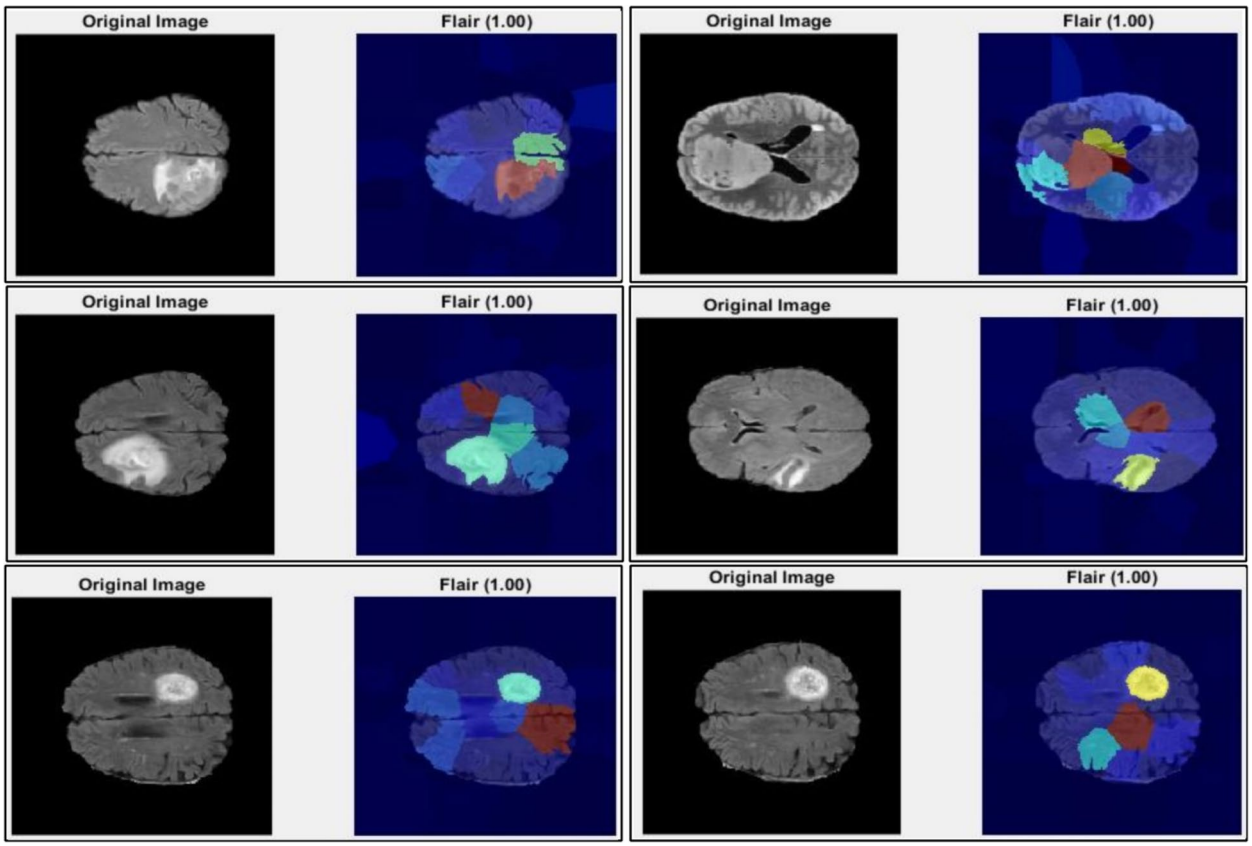


Fig. 13 Proposed SMDeepNet architecture visual interpretation using LIME explainable AI technique

Table 7 Comparison of proposed architecture with existing segmentation techniques

Serial	Reference	Average dice score	Dataset	Year presented
1	Bouzara et al. [39]	0.9200	BraTS-2023	2024
2	Kharaji et al. [40]	0.8300	BraTS-2023	2024
3	Ren et al. [41]	0.7900	BraTS-2023	2024
4	Zhang et al. [42]	0.9234	BraTS-2023	2024
5	Yazıcı et al. [43]	0.9219	BraTS-2023	2024
6	**Proposed Technique**	**0.9354**	**BraTS-2023**	2024

Bold denotes the most significant values

difference between configurations is 0.0063, 0.0152, and 0.0243, for Accuracy, Dice, and IoU, respectively. Notably, IoU showed a little higher sensitivity to configuration modifications because of its rigorous overlap constraint. Nonetheless, the stability and consistency of segmentation model are validated by the comparatively low standard deviation values across all parameters. The results of ablation study are presented in Table 9. These quantitative results offer verifiable numerical proof of the dependability and robustness.

It is significant to note that, even with slight modifications, no configuration led to a significant deterioration in segmentation quality. The model does not show instability or performance collapse under various experimental conditions. It is confirmed by the relative proximity of the minimum and maximum values across all metrics. Regarding robustness, the low standard deviation (SD) indicates that the suggested segmentation architecture retains consistent region overlap and reliable feature extraction even in the configuration modifications. This presents that the learning process of the model is not too

sensitive to initialization or small changes in parameters, which increases the dependability of model for real-world clinical use.

Ablation study 2: fusion strategy

The Fire-RB, Hybrid-EA, summation fusion, weighted fusion, and our suggested depth concatenation were

Table 8 Comparison with existing modality classification methodologies

No	Reference	Accuracy	Core methodology	Parameters (million)	Dataset	Year
1	Samuel et al. [27]	97.57%	Φ-Net (3D CNN); multi-branch architecture with residual modules inspired by ResNet. 7 residual blocks	3.1	T1, T2, FLAIR	2018
2	Khan et al. [9]	97.8%, 96.9% and 92.5%	Transfer learning using VGG16 & VGG19 for deep feature extraction. CML-ELM feature selection. PLS fusion; final classification using Extreme Learning Machine (ELM)	(VGG16 has 138 million, VGG19 has 144 million)	BraTS2015, BraTS2017 and BraTS2018	2020
3	Hapsari et al. [45]	95.50% Average	en-CNN (simplified VGG16). 7 conv layers, 4 max-pool, dropout 0.1. sigmoid classifier; binary LGG vs GBM	Not reported	BraTS-2018	2021
4	Zahid et al. [46]	94.40%	BrainNet framework. ResNet101 deep feature extraction (2048 features). DE + PSO feature optimization. PCA reduction. Cubic SVM classifier	ResNet backbone has 44 million	BraTS-2018	2022
5	Dunyuan et al. [44]	97.9%, 97.00%	Cross-Modality Guidance-Aided Multi-Modal Learning. ResNet Mixed Convolution (3D + 2D CNN). dual attention (spatial + slice-wise). cumulative fusion strategy	Not reported	BraTS-2018, BraTS-2019	2024
6	Thakur et al. [47]	98.67%	ED-ViTTL (ViT + VGG19 ensemble). 5 ViT variants evaluated. ViT-B/32 + VGG19. late feature fusion. fivefold CV. class-weighted loss	200 million	MRI Brain tumor Scans	2025
	Proposed Technique	**99.20%**	**Fire-RB, Hybrid-EA**	**42.8, 6.9**	**BraTS-2023**	**2025**

Table 9 Overlap stability analysis

Metrics	Mean ± SD	Minimum (min)	Maximum (max)	Δ(Max−Min)
Accuracy	0.9872 ± 0.0024	0.9841	0.9904	0.0063
Dice	0.9420 ± 0.0060	0.9346	0.9498	0.0152
IoU	0.8951 ± 0.0093	0.8831	0.9074	0.0243

Table 10 Ablation study on fusion strategy

Serial	Method	Accuracy	Total parameters (millions)	Observation
1	Fire-RB	98.60	42.8	Local features
2	Hybrid-EA	98.50	6.9	Strong global features
3	Summation fusion	98.72	49.7	Compressed complementary features
4	Weighted fusion	98.78	50.3	More parameters with small gain
5	Depth concatenation	99.20	49.8	preserve complimentary representations

examined. It was done to ensure that improvements in accuracy were not solely the result of increased parameters. It is depicted in Table 10. Firm localized texture and boundary cues are captured by the Fire-RB design. The Hybrid-EA architecture uses attention to highlight global contextual information. These heterogeneous feature spaces are compressed via summation fusion, resulting in a partial loss of complementary information. The model size is increased to about 50.3 M parameters using weighted fusion. However, the advantages remain relatively modest. On the other hand, depth concatenation achieves a total parameter count of only 49.8 M while preserving the full representational capacity of both streams. Depth concatenation consistently yields the best classification accuracy, even though it adds fewer parameters than weighted fusion. This demonstrates that successful multimodal feature integration is not just an increase in parameter count. However, it is the reason for the observed performance improvement.

Conclusion

Early diagnosis with a non-invasive strategy is crucial for patients with brain tumor ailments. MRI brain scans are pivotal for early diagnosis and subsequent treatment planning. Brain tumor segmentation precisely segments the infected part of the brain (i.e., brain tumor), and the classification process accurately distinguishes each modality of the MRI scan. This work proposes a fully automated framework that accurately segments and classifies brain tumors from MRI scans. In the segmentation section, an optimized DeepLabV3 + model is presented, in which the backbone network is trained with dynamically initialized hyperparameters via Bayesian optimization. Down-sampling and up-sampling are performed, yielding a precisely segmented image. In the classification stage, two models, Fire-RB and Hybrid-EA, are fused via channel-wise depth concatenation, and the final classification is performed to obtain accurately classified MRI modalities. The segmentation methodology achieves accuracies of 95.32%, 93.54%, and 90.16% for Dice score and IoU, respectively. Experimental results show that the classification methodology achieves an accuracy of 99.20%, which is higher than that reported in similar prior work. Experiments show that both methods improve performance across accuracy, IoU, and precision. In the future, the backbone model (i.e., ResNet50) for segmentation will be replaced with a customized transformer-based model. Moreover, a feature-visualization-based method will be introduced to identify the contributing features for each methodology and assess the effect of each layer on quality feature extraction.

Acknowledgements
This work was supported through Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176) and the Soonchunhyang University Research Fund. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through small Research Project under grant number RGP1/290/46.

Author contributions
Muhammad Sami Ullah, Muhammad Attique Khan, Yunyoung Nam, Juan M Gorriz, Nouf Abdullah Almujally, Areej Alasiry, Mehrez Marzougui, Amir Hussain—All authors contributed equally to this work.

Funding
This work was supported through Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023–00218176) and the Soonchunhyang University Research Fund. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through small Research Project under grant number RGP1/290/46.

Data availability
The datasets used in this work are publicly available on the BraTS platform. ([https://www.med.upenn.edu/cbica/brats2020/data.html] (https://www.med.upenn.edu/cbica/brats2020/data.html) and ([https://www.synapse.org/Synapse:syn25829067/wiki/610863] (https://www.synapse.org/Synapse:syn25829067/wiki/610863)).

Declarations

Ethics approval and consent to participate
Not required.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Computer Science, HITEC University, Taxila, Pakistan. [2]Center of AI, Prince Mohammad Bin Fahd University, Al-Khobar, Saudi Arabia. [3]Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Republic of Korea. [4]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. [5]College of Computer Science, King Khalid University, Abha 61413, Saudi Arabia. [6]Dasci Institute, University of Granada, Granada, Spain. [7]School of Computing, Edinburgh Napier University, Edinburgh, UK. [8]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK.

Received: 10 September 2025 Accepted: 25 February 2026
Published online: 06 March 2026

References

1. Özbay E, Özbay FA. Interpretable features fusion with precision MRI images deep hashing for brain tumor detection. Comput Methods Programs Biomed. 2023;231:107387.
2. Subramaniam EVD, Krishnasamy V. ABES: attention bi-directional ensemble SVM for early detection of brain tumors. Neural Comput Appl. 2024. https://doi.org/10.1007/s00521-024-09688-w.
3. C. Hamel, M. Venturi, R. Margau, and P. Pageau. Canadian association of radiologists diagnostic imaging referral guidelines. SAGE Publications Sage CA: Los Angeles. 2023. 74: 614–615.

4. S. Ghanavati, J. Li, T. Liu, P. S. Babyn, W. Doda, and G. Lampropoulos. Automatic brain tumor detection in magnetic resonance images. 2012 9th IEEE international symposium on biomedical imaging (ISBI), 2012: 574–577.
5. Ullah Z, Usman M, Jeon M, Gwak J. Cascade multiscale residual attention CNNs with adaptive ROI for automatic brain tumor segmentation. Inf Sci. 2022;608:1541–56.
6. X. Xu, J. Yang, D. Hu, L. Y. Por, and C. Li. Diffusion model with relation-aware attention and edge-aware constraint for multi-modal brain tumor segmentation. IEEE J Biomed Health Inform. 2025.
7. S. Ünalan, F. E. Onat, F. Şahin, and B. Keserci. Revealing distinctive insights: machine learning-enhanced ensembled MRI sequences for pediatric posterior fossa tumor classification. 2024.
8. Siddiqah M, et al. DSA: deep self-attention medical transformer neuro-technology for brain tumor segmentation. Int J Imaging Syst Technol. 2025;35(3):e70109.
9. Khan MA, et al. Multimodal brain tumor classification using deep learning and robust feature selection: a machine learning application for radiologists. Diagnostics Basel. 2020;10(8):565.
10. Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. Front Public Health. 2023;11:1273253. https://doi.org/10.3389/fpubh.2023.1273253.
11. Banerjee S, Mitra S, Shankar BU. Automated 3D segmentation of brain tumor using visual saliency. Inf Sci. 2018;424:337–53.
12. Sahoo AK, Parida P, Muralibabu K, Dash S. Efficient simultaneous segmentation and classification of brain tumors from MRI scans using deep learning. Biocybern Biomed Eng. 2023;43(3):616–33. https://doi.org/10.1016/j.bbe.2023.08.003.
13. Zhang X, Ou Q, Wang J. Variable precision fuzzy rough sets based on overlap functions with application to tumor classification. Inf Sci. 2024;666:120451.
14. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17. https://doi.org/10.1038/s41746-020-0221-y.
15. Abbas MJ, Khan MA, Hussain A, Ayouni S, Maddeh M, Alhayan F. XRD-Net: a novel explainable residual dense fusion network for Alzheimer's disease recognition from MRI images. Cogn Comput. 2025;17(6):174.
16. Khalifa M, Albadawy M. AI in diagnostic imaging: revolutionising accuracy and efficiency. Comput Methods Programs Biomed Update. 2024;5:100146. https://doi.org/10.1016/j.cmpbup.2024.100146.
17. Mushtaq M, Khan MA, Hussain Z, Ayouni S, Maddeh M, Alhayan F. A network-level fused DenseInc226 lightweight architecture for Alzheimer's disease prediction from magnetic resonance imaging. Cogn Comput. 2025;17(6):1–25.
18. Ranjbarzadeh R, Caputo A, Tirkolaee EB, Ghoushchi SJ, Bendechache M. Brain tumor segmentation of MRI images: a comprehensive review on the application of artificial intelligence tools. Comput Biol Med. 2023;152:106405. https://doi.org/10.1016/j.compbiomed.2022.106405.
19. Narayana MV, Rao JN, Shrivastava S, Ghantasala GSP, Ioannou I, Vassiliou V. A framework for identification of brain tumors from MR images using progressive segmentation. Heal Technol. 2024;14(3):539–56. https://doi.org/10.1007/s12553-024-00844-9.
20. Abdel-Maksoud E, Elmogy M, Al-Awadi R. Brain tumor segmentation based on a hybrid clustering technique. Egypt Inform J. 2015;16(1):71–81.
21. Sharp G, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. Med Phys. 2014;41(5):050902. https://doi.org/10.1118/1.4871620.
22. Panduri B, Rao OS. A survey on brain tumour segmentation techniques in deep learning. Int J Intell Syst Appl Eng. 2024;12(7s):412–25.
23. Ullah F, Salam A, Abrar M, Amin F. Brain tumor segmentation using a patch-based convolutional neural network: a big data analysis approach. Mathematics. 2023;11(7):1635.
24. Kaifi R. A review of recent advances in brain tumor diagnosis based on AI-based classification. Diagnostics. 2023. https://doi.org/10.3390/diagnostics13183007.
25. Kumar A. Study and analysis of different segmentation methods for brain tumor MRI application. Multim Tools Appl. 2023;82(5):7117–39.
26. Kshatri SS, Singh D. Convolutional neural network in medical image analysis: a review. Arch Comput Methods Eng. 2023;30(4):2793–810.
27. S. Remedios, D. L. Pham, J. A. Butman, and S. Roy, Classifying magnetic resonance image modalities with convolutional neural networks, in Medical Imaging 2018: computer-aided diagnosis. 2018. 10575: 558–563.
28. J. Cho et al. Disentangled multimodal brain MR image translation via transformer-based modality infuser," in Medical Imaging 2024: Image Processing. 2024. 12926: 602–607.
29. Li Y, et al. A review of deep learning-based information fusion techniques for multimodal medical image classification. Comput Biol Med. 2024;177:108635. https://doi.org/10.1016/j.compbiomed.2024.108635.
30. 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. 770–778.
31. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision, in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. 2818–2826.
32. C. Jiao and T. Yang, An overview of multimodal brain tumor MR image segmentation methods, in 3rd international conference on artificial intelligence, automation, and high-performance computing (AIAHPC 2023), 2023. 12717: 725–733.
33. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, Mobilenetv2: inverted residuals and linear bottlenecks, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. 4510–4520.
34. A. Vaswani et al. Attention is all you need, Adv Neural Inf Process Syst. 2017, 30.
35. A. Dosovitskiy et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
36. M. Tan and Q. Le. Efficientnet: rethinking model scaling for convolutional neural networks, in International conference on machine learning, 2019. 6105–6114.
37. Rahman T, Islam MS. MRI brain tumor detection and classification using parallel deep convolutional neural networks. Meas Sensors. 2023;26:100694.
38. Agarwal M, Rani G, Kumar A, Kumar P, Manikandan R, Gandomi AH. Deep learning for enhanced brain tumor detection and classification. Results Eng. 2024;22:102117.
39. A. Bouzara and A. Kermi, Automated brain tumor segmentation in multimodal MRI scans using a multi-encoder U-net model," in 2024 8th international conference on image and signal processing and their applications (ISPA), 2024: IEEE. 1–8.
40. Kharaji M, Abbasi H, Orouskhani Y, Shomalzadeh M, Kazemi F, Orouskhani M. Brain tumor segmentation with advanced nnU-Net: pediatrics and adults tumors. Neurosci Inform. 2024. https://doi.org/10.1016/j.neuri.2024.100156.
41. T. Ren, E. Honey, H. Rebala, A. Sharma, A. Chopra, and M. Kurt, An optimization framework for processing and transfer learning for the brain tumor segmentation, arXiv preprint arXiv:2402.07008, 2024.
42. L. Zhang, Y. Cheng, L. Liu, C.-B. Schönlieb, and A. I. Aviles-Rivero, Biophysics informed pathological regularisation for brain tumour segmentation, arXiv preprint arXiv:2403.09136, 2024.
43. Z. A. Yazıcı, İ. Öksüz, and H. K. Ekenel, Attention-enhanced hybrid feature aggregation network for 3D brain tumor segmentation, arXiv preprint arXiv:2403.09942, 2024.
44. D. Xu, X. Wang, J. Cai, and P.-A. Heng, Cross-modality guidance-aided multi-modal learning with dual attention for MRI brain tumor grading, arXiv preprint arXiv:2401.09029, 2024.
45. Hapsari PAT, et al. Brain tumor classification in MRI images using en-CNN. Int J Intell Eng Syst. 2021;14(4):437–51.
46. Zahid U, et al. BrainNet: Optimal deep learning feature fusion for brain tumor classification. Comput Intell Neurosci. 2022;2022(1):1465173.
47. Thakur A, Patnaik PK, Kumar M, Choudhary C. ED-ViTTL: Ensemble vision transformer and transfer learning approach for brain tumor classification. Mach Vis Appl. 2025;36(6):116.
48. H. B. Li et al., The Brain Tumor Segmentation (BraTS) Challenge 2023: Brain MR image synthesis for tumor segmentation (BraSyn), (in eng), ArXiv. 2023.
49. U. Baid et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, arXiv preprint arXiv:2107.02314, 2021.

50. Menze BH, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging. 2014;34(10):1993–2024.
51. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in Proceedings of the European conference on computer vision (ECCV). 2018. 801–818.
52. L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.
53. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13. Cham: Springer; 2014. p. 818–33.
54. J. Snoek, H. Larochelle, and R. P. Adams, Practical bayesian optimization of machine learning algorithms, Adv Neural Inf Process syst, vol. 25, 2012.
55. H. Gholamalinezhad and H. Khosravi, Pooling methods in deep neural networks, a review, arXiv preprint arXiv:2009.07485, 2020.
56. F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360, 2016.
57. A. G. Howard et al., Mobilenets: efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
58. X. Glorot, A. Bordes, and Y. Bengio, Deep sparse rectifier neural networks, in Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011: JMLR workshop and conference proceedings. 315–323.
59. F. Chollet, Xception: Deep learning with depthwise separable convolutions, in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. 1251–1258.
60. A. Howard, A. Zhmoginov, L.-C. Chen, M. Sandler, and M. Zhu, Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation, in Proc. CVPR. 2018. 4510–4520.
61. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncertain Fuzziness Knowl Based Syst. 1998;6(02):107–16.
62. L. Borawar and R. Kaur, ResNet: Solving vanishing gradient in deep networks. in Proceedings of International Conference on Recent Trends in Computing: ICRTC 2022, 2023: Springer, Cham. 235–247.
63. A. M. Hafiz, S. A. Parah, and R. U. A. Bhat, Attention mechanisms and deep learning for machine vision: a survey of the state of the art. arXiv preprint arXiv:2106.07550. 2021.
64. Araujo A, Norris W, Sim J. Computing receptive fields of convolutional neural networks. Distill. 2019;4(11):e21.
65. P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," Adv Neural Inf Process Syst. 2019. 32.
66. J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, Neural module networks, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. 39–48.
67. C. Szegedy et al. Going deeper with convolutions, in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. 1–9.
68. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58.
69. R. Pascanu, T. Mikolov, and Y. Bengio, On the difficulty of training recurrent neural networks. in International conference on machine learning. 2013. 1310–1318.
70. J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," Adv Neural Inf Process Syst. 1993. 6.

Publisher's Note