



Article

An Explainable Brain Tumor Detection Framework for MRI Analysis

Fei Yan ¹ , Yunqing Chen ¹, Yiwen Xia ¹, Zhiliang Wang ¹ and Ruoxiu Xiao ^{1,2,*} ¹ School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China² Shunde Innovation School, University of Science and Technology Beijing, Foshan 100024, China

* Correspondence: xiaoruoxiu@ustb.edu.cn

Abstract: Explainability in medical images analysis plays an important role in the accurate diagnosis and treatment of tumors, which can help medical professionals better understand the images analysis results based on deep models. This paper proposes an explainable brain tumor detection framework that can complete the tasks of segmentation, classification, and explainability. The re-parameterization method is applied to our classification network, and the effect of explainable heatmaps is improved by modifying the network architecture. Our classification model also has the advantage of post-hoc explainability. We used the BraTS-2018 dataset for training and verification. Experimental results show that our simplified framework has excellent performance and high calculation speed. The comparison of results by segmentation and explainable neural networks helps researchers better understand the process of the black box method, increase the trust of the deep model output, and make more accurate judgments in disease identification and diagnosis.

Keywords: explainable AI; brain tumor detection; deep learning; MRI; re-parameterization



Citation: Yan, F.; Chen, Y.; Xia, Y.; Wang, Z.; Xiao, R. An Explainable Brain Tumor Detection Framework for MRI Analysis. *Appl. Sci.* **2023**, *13*, 3438. <https://doi.org/10.3390/app13063438>

Academic Editor: Sami Bourouis

Received: 16 February 2023

Revised: 7 March 2023

Accepted: 7 March 2023

Published: 8 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Tumors have threatened human health for a long time. Malignant brain tumors, such as glioblastoma, have a very low survival rate and are highly correlated with the death of young patients with brain tumors [1]. Patients with brain tumors often undergo surgery, radiotherapy, chemotherapy, and other treatment methods, but distinguishing brain tumors from the brain parenchyma is difficult visually. This task needs to be carefully treated to locate and remove tumors through surgery. Medical staff usually make diagnosis first through image diagnosis technologies, such as electroencephalogram (EEG), computed tomography (CT), and nuclear magnetic resonance imaging (MRI), and assist in locating the diseased area and boundary through imaging software. Computer-aided diagnosis systems (CADs) based on MRI can generate hundreds of images, which can be used to judge or classify the lesion area in detail [2]. CADs can detect and visualize various structures of the human brain, such as the blood–brain barrier and brain tumor boundary. However, medical researchers should pay attention to a series of problems, such as low contrast and clarity, while processing a large number of medical images due to the complexity of manual retrieval of MRI images and the need for classification. Whether these problems can be well-solved is related to doctors' treatment methods and decision-making plans. If these MRI sequences are automatically analyzed by machine learning, then medical professionals can more quickly confirm whether the tumor exists, then classify or analyze it if necessary.

With the continuous development of artificial intelligence and interdisciplinary technologies, deep models have been increasingly applied to various medical image tasks, especially in brain science research [3]. Compared with natural scene images, medical image data samples are usually less, leading to many limitations in practical applications [4]. Deep models consider the fitting results of data input and output and often pay less attention to the information of middle layers. However, in the medical field, image classification

and segmentation based on MRI are often directly related to a series of important treatment methods, such as surgery, and have high risk. Medical professionals hold some distrust toward the output results due to the opacity of the black box structure [5]. Directly providing a segmentation or classification result may reduce the credibility on the prediction results.

Explainable artificial intelligence (XAI) is necessary in the medical field because it can help people better understand deep models [6]. First, the diagnostic task is vulnerable to errors caused by subjective judgments [7]; second, explainability can increase the transparency of the system, improve the quality of evaluation, and provide help for accountability, fairness, and ethical issues [8]. Therefore, the application of XAI in medical images has practical significance. Through XAI, medical professionals have the opportunity to understand the principles behind medical findings and apply deep learning technology better [9]. The need for explainable tasks related to medical images is increasing. In the European Union, medical-related XAI has become a requirement of general data protection regulation (GDPR) [10].

Therefore, we propose an explainable framework for MRI images analysis of brain tumors to provide a judgment basis for medical researchers in diagnosis and treatment. Our contributions are as follows:

1. Our framework includes two simple and efficient deep models, which can comprehensively analyze the MRI images of brain tumors and explain the classification results;
2. We propose a brain tumor classification network based on re-parameterization. The model has the advantages of simplicity and efficiency and can fully capture space and channel information while considering classification and explainability tasks.

The rest of this paper is organized as follows: Section 2 introduces the related work; Section 3 discusses the materials and methods in detail; Section 4 presents the results and discussion; and Section 5 summarizes the conclusion.

2. Related Work

Explainability based on deep models can be divided into input-dependent explanations (instance-level explanations) and input-independent explanations (model-level explanations). The former can find and explain the features that have the greatest impact on the prediction results, while the latter can directly explain the model without considering the network input. In medical imaging, with the clarity of segmentation and classification, explainability of these tasks usually adopt instance-level explanations, of which the more widely used are perturbation-based methods and gradient-based methods.

In the early stage of medical image explainability work, perturbation-based methods are often used. These methods can study the network by observing output changes under different input disturbances. In the medical field, disturbances can include various forms, such as shape and occlusion [11,12]. However, since the running time depends on the number of input features, more computing resources and time are often needed to achieve better results, so more consideration should be given to the choice between accuracy and explainability.

In gradient-based methods, gradient is the approximate value of the importance of the input feature and is highly related to model parameters. Similar models often have the advantage of post-processing. Explainability can be independent of the model training process, avoiding the balance between accuracy and computational loss. Gradients used in XAI include integral gradient (IG) [13], vanilla gradient (VG) [14], and guided backpropagation (GB) [15]. In recent years, researchers have increasingly used extension methods based on class activation mapping (CAM) [16], such as Grad-CAM [17] and Guided Grad-CAM [17]. These improved methods do not change the model structure nor need to retrain the model.

In the application of deep models in the medical field, researchers mainly use target detection networks based on deep learning, such as U-Net [18], VGG [19], ResNet [20], and GoogLeNet [21], which have important contributions to the diagnosis of a series of diseases. Tian et al. [22] used Fully Convolutional Networks in liver tumor image classification and found that it improved the tumor localization effect. Han et al. [23] proposed an

ACAE model that can be used to classify different parts of the spine and quickly detect the suspected lesion areas. Teixeira et al. [24] used a variety of networks to diagnose and identify COVID-19 on chest X-ray datasets. Khan et al. [25,26] proposed a variety of brain tumor classification models based on deep learning. Based on the pre-trained VGG-like model, tumor classification is performed through migration learning.

In recent years, researchers have used XAI to comprehensively evaluate and explain model results. Yang et al. [27] used 3D convolutional neural networks (CNNs) for the classification and visual explainability of Alzheimer's disease. Wickstrom et al. [28] used GB to analyze the explainability of colon polyps. Esmaili et al. [29] added explainability method based on Grad-CAM to the segmentation task of 2D glioma. Saleem et al. [30] applied similar methods to 3D images. Natekar et al. [31] used Grad-CAM to explain the brain tumor segmentation task. Adebayo et al. [32] found that CAM-based methods are better in classification tasks through the sanity check. Pereira et al. [33] proposed an explainable method combining global and local information for tumor segmentation. They used GB and CAM in brain tumors detection. The experimental results show that GB can distinguish important areas rather than categories, and CAM works well in both tasks. Narayanan et al. [34] used GoogLeNet and ResNet to detect malaria, diabetic retinopathy, brain tumors, and tuberculosis in different imaging modalities. They visualized the class activation mappings to enhance the understanding of these deep networks.

Although XAI has been widely used in natural image tasks, it is insufficient to be applied to medical tasks [9]. The explainability of medical images is important and has great demand and scalability because it is related to the degree of trust of medical professionals in model results and subsequent operations.

3. Materials and Methods

This paper proposes a new framework that uses a composite explainable network to carry out end-to-end technologies of MRI images for better diagnosis and analysis of brain tumors. As shown in Figure 1, our overall framework consists of two models: segmentation and classification. We use 3D nnU-Net [35] segmentation models to complete the segmentation task independently and a new classification model based on re-parameterization to complete the classification task. The classification model can complete the classification task for high-grade gliomas (HGGs) and low-grade gliomas (LGGs). The explainability of the classification model results is also analyzed. Our framework can produce segmentation, classification, and explainability output results by using the MRI input sequences. Although visual analysis can be carried out according to any feature extraction model, we found that explainable heatmaps rendering through the segmentation network are very close to the prediction results. The segmentation result and the generated heatmaps are greatly affected by ground truth (GT). Our main purpose is to conduct detailed analysis and complementary interpretation of the decision process and the results of model segmentation and classification through explainability. The explainability of the classification model can be regarded as the visual result of the output of the classification model. Explainability results will help medical professionals to have better understanding and judgment and carry out more reliable follow-up medical operations. Therefore, our overall framework uses the classification network to generate heatmaps.

3.1. Segmentation Model

Considering the current research situation of medical segmentation model and that our main purpose is to conduct visual analysis rather than research medical segmentation model, we directly adopted nnU-Net for segmentation tasks. This network makes several improvements to U-Net, including using a larger network with Skip Connection, replacing batch normalization (BN) layers with group normalization (GN) layers, and using Axial Attention in decoder network. The indicators are greatly improved compared with those of the original U-Net. The whole network is shown in Figure 2.

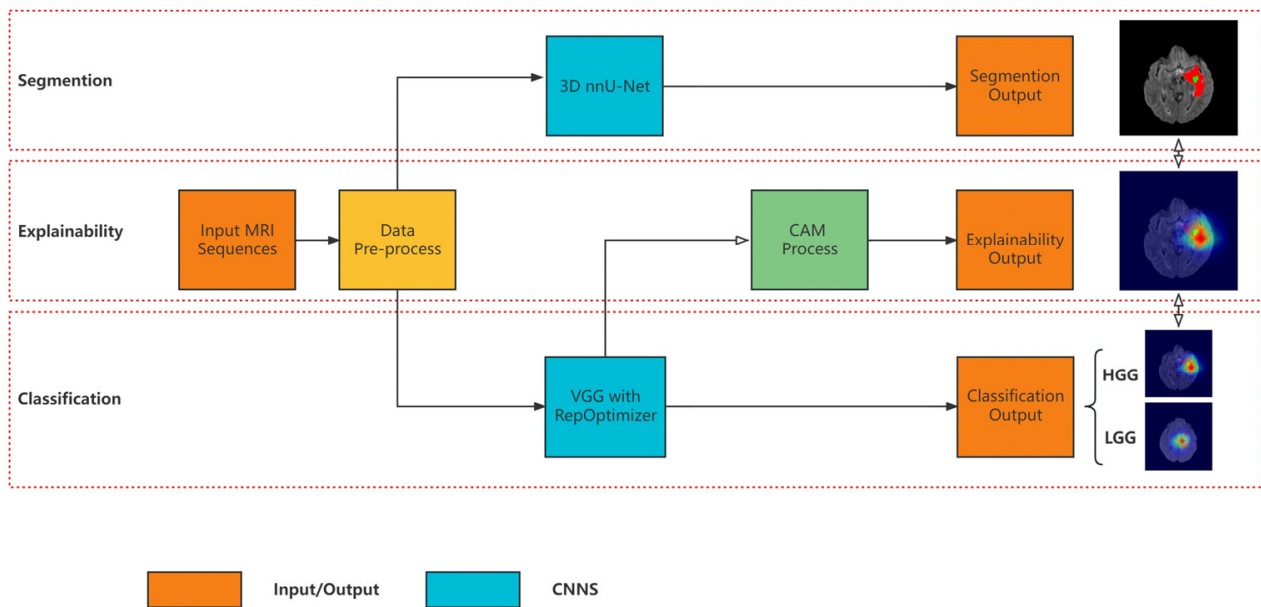
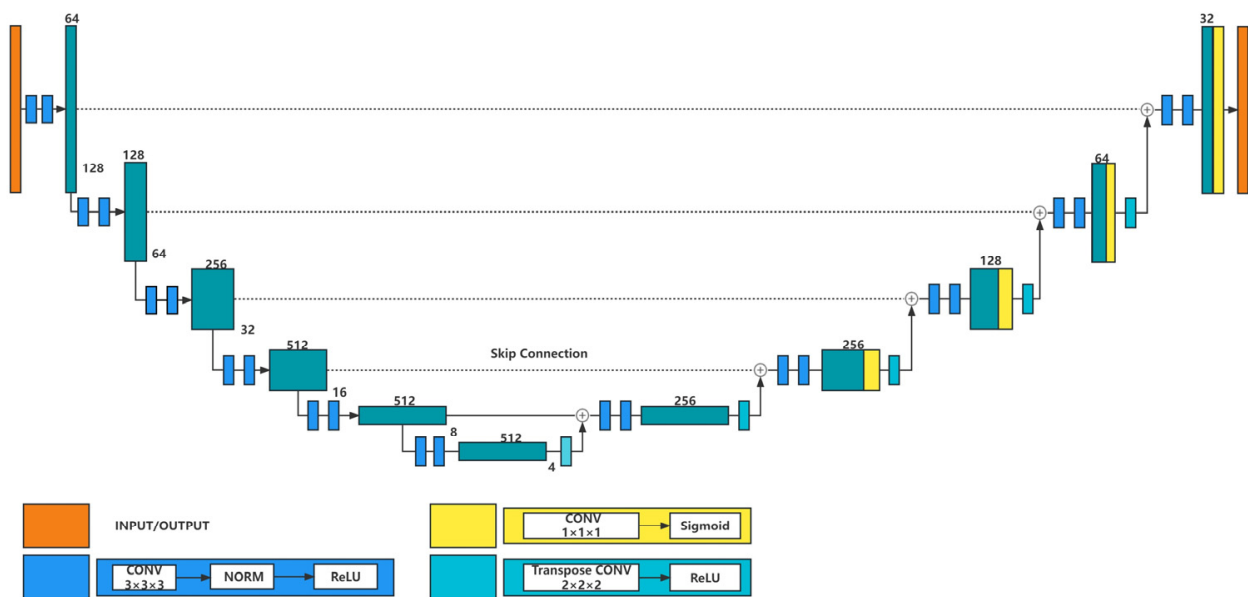


Figure 1. Overall explainable framework of brain tumor detection. The three parts of the framework include: segmentation, classification, and explainability.



of the mode of a specific structure with a general optimizer. The model is trained by adding a priori to the optimizer while maintaining excellent performance and less parameter quantity. Compared with RepVGG, the process of training and reasoning was simplified, and the re-parameterized model can be equivalent to the original architecture, thereby greatly improving the performance of the model while achieving structural sparsity. In the actual model architecture, the BN layer in the RepVGG model is replaced by a constant or trainable channel direction scaling, that is, using Grad Masks to replace the constant-scale linear addition (CSLA). The corresponding optimizer is adjusted for gradient re-parameterization (GR).

When each branch of the multi-branch model contains a linear trainable parameter and a constant scaling value, the model performance can be maintained by reasonably setting the scaling value, that is, CSLA. Assuming that a model contains two convolution branches with linear scaling, the output feature map can be expressed as:

$$M_{CSLA} = \alpha_1 (X_1 * W^{(1)}) + \alpha_2 (X_2 * W^{(2)}) \quad (1)$$

where M_{CSLA} is the output feature map, and α_1 , X_1 , $W^{(1)}$ are the scaling coefficient, input of this channel, and weight of this channel. The corresponding label 2 is the second channel. When the corresponding CSLA is updated using the regular SGD update rule, the gradient equivalent to the GR corresponding item should be multiplied by $(\alpha_1^2 + \alpha_2^2)$, that is, the updated weight is:

$$W'^{(i+1)} \leftarrow W'^{(i)} - \lambda (\alpha_1^2 + \alpha_2^2) \frac{\partial L}{\partial W'^{(i)}} \quad (2)$$

where L is the objective function, and λ is the learning rate. The scale factor $(\alpha_1^2 + \alpha_2^2)$ is the Grad Mask. After the parameter initialization of CSLA, the equivalent replacement can only be achieved by multiplying the parameter and Grad Mask when updating:

$$M_{CSLA} = M_{GR} = X * W' \quad (3)$$

In the actual calculation process, by stacking the middle 3×3 and 1×1 convolution of the RepVGG network at the center point in a fixed proportion, it is equivalent to replacing the BN layer after RepVGG with a Grad Mask to realize replacement on the model. That is, the Grad Mask can be expressed as:

$$G = c_1^2 + c_3^2 + 1 \quad (4)$$

where c_1^2 , c_3^2 , and scalar 1 are the Grad Mask of 1×1 , 3×3 , and identity layers, respectively. Finally, the Grad Mask and the adaptation optimizer are confirmed using the CIFAR-100 dataset [39] for small-scale training through the hyper-parameter search method. In the process of using transfer learning to train other models, the adaptation optimizer can use Grad Masks to multiply the gradient of the corresponding operator to achieve the final equivalent replacement.

The classification network in this paper is similar to the VGG network structure. As shown in Figure 3, the overall structure is built according to the model structure of RepOp-B1, which is a five-layer structure. In the first four blocks, the numbers of layers of RepOpt are 1, 4, 6, and 16. Since the main task of this paper is to obtain better explainable feature maps, under our comprehensive consideration, we modify the last layer to a CNN layer containing three layers of atrous convolution to obtain a larger receptive field. In saliency prediction tasks, better heatmaps can be captured by obtaining a larger receptive field and using a larger resolution. The increased resolution does not significantly affect the network due to the extremely simplified backbone network. Through our modification, our network can achieve better explainability while maintaining network performance and low computational loss. For the classification task, we use the global average pooling (GAP)

layer instead of the full connection (FC) layer to further reduce the amount of parameters. Classification tasks and explainability can be completed independently.

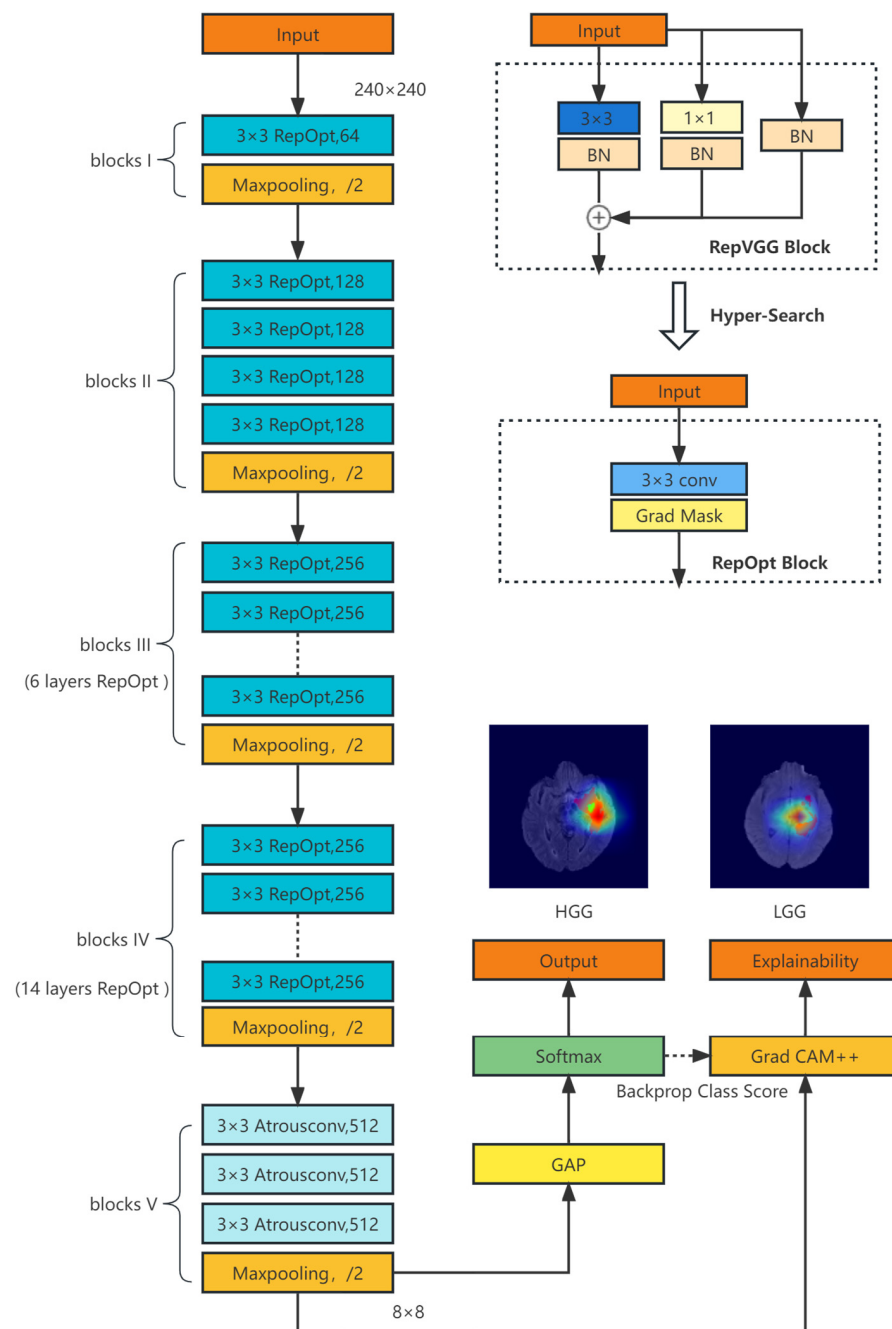


Figure 3. Classification network architecture.

3.3. Explainability

Although segmentation and classification models are widely used, most of the existing deep models in medicine are insufficiently explainable. Our classification model can also completely explain learning tasks.

As described in Chapter 2, explainability uses a variety of methods based on perturbation or gradient. We tested a variety of methods in the framework, including GB, Grad-CAM, and Grad-CAM++ [40]. Grad-CAM++ was chosen in our framework which can better display diseased areas. Compared with GB and CAM, Grad-CAM uses the gradient of the specific class output instead of using the weight of the full connection

layer out for calculation, which has better extensibility and the model does not need to be retrained. Grad-CAM++ uses two steps and three steps gradients, which can be replaced by the square and the third power of one step gradient in calculation to cover comprehensive objects. In VGG networks, all objects can be better covered when multiple instances of the same category exist, so better results can be obtained.

For a feature A of a classification c , its weight α_k^c at ij position of the k channel can be expressed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \quad (5)$$

where Y^c represents the predicted score of category c , and Z equals the product of the width and height of the feature layer. In Grad-CAM++, the weight gradient α_{ij}^{kc} and ReLU functions corresponding to a certain classification c are added. The weight w_k^c can be expressed as:

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot \text{ReLU} \left(\frac{\partial Y^c}{\partial A_{ij}^k} \right) \quad (6)$$

Finally, the explainable heatmaps of this category L_{ij}^c can be expressed as follows:

$$L_{ij}^c = \sum_k w_k^c \cdot A_{ij}^k \quad (7)$$

4. Results and Discussion

4.1. Dataset and Pre-Processing

In this paper, we use the BraTS Challenge 2018 [41] dataset to complete our segmentation and classification tasks. The MRI sequences of this dataset is $4 \times 240 \times 240 \times 155$, where the pixel size of each image is 240×240 , with 155 image sequences in each case. The tumor segmentation tag includes four modalities: T1-weighted (T1W), T2-weighted (T2W), post-contrast T1-weighted (T1Gd), and fluid-attenuated inversion recovery (FLAIR). The dataset also includes 259 images of HGG and 76 images of LGG, which can be used for the classification task. For the segmentation task, we adopt the pre-processing method of z-score normalization, random flip, random rotation, and intensity transformation to better extract effective features and avoid problems, such as over-fitting. The same pre-processing method is applied to the classification input image sequences. To avoid that the final input of the classification network is too small to affect the imaging of the explainable heatmaps, we adjust the clipped input images back to 240×240 . About 70% of the data are used for training, 10% of the data are used for validating, and the remaining 20% are used for testing.

4.2. Experimental Setting

For the segmentation model, we cross-verify all training sequences, with a momentum of 0.99, an initial learning rate (Lr) of 0.001, Lr patience of 30, and decay of 3×10^{-5} . Using Pytorch as the main framework, the image sequence is trained for 1000 epochs on NVIDIA TitanX 3080 GPU, and the batch size (BS) is 64. The training duration is about 2 days.

For the classification model, we use pre-training parameters of REP-OP-B1, with an initial learning rate of 0.01, Lr patience of 30. Random gradient descent (SGD) optimizer and RepOptimizer have the same momentum of 0.9 and decay of 4×10^{-5} . Cross entropy is used as the loss function for training. The image sequence is trained for 200 epochs on NVIDIA TitanX 3080 GPU, and the BS is 64. The training duration is about 1 h.

Our explainable method is completed simultaneously with the training of the classification model. As a post-XAI method, modifying the network structure or training the network again is not needed. Finally, our overall framework can obtain corresponding segmentation results, classification results and explainable heatmaps from training.

4.3. Results and Analysis

In this section, we mainly show the performance of the model and the explainable heatmaps combined with the segmentation outputs. Our main purpose is to explain the role of explainability in medical images, so our network adopts the extremely simplified design and makes modifications to the generation effect of explainable heatmaps. The display result of explainability is equivalent to the visualization result of the classification model.

4.3.1. Segmentation Results

In the segmentation model, flair sequences are used as MRI images to better compare with the explainable output. Figure 4 shows our segmentation results, which are image, segmentation GT (Seg-True) and segmentation outputs (Seg-Pre). The labels of brain tumors are also shown in our model. Different types of tumors are distinguished by different colors, among which the tumor marker edema is red, the enhancing tumor (ET) is blue, and the non-enhancing tumor is green. The location where the three types of regions are concentrated or superimposed is considered as the whole tumor (WT), and the location with relatively concentrated enhancing and non-enhancing regions can be considered as the tumor core (TC).

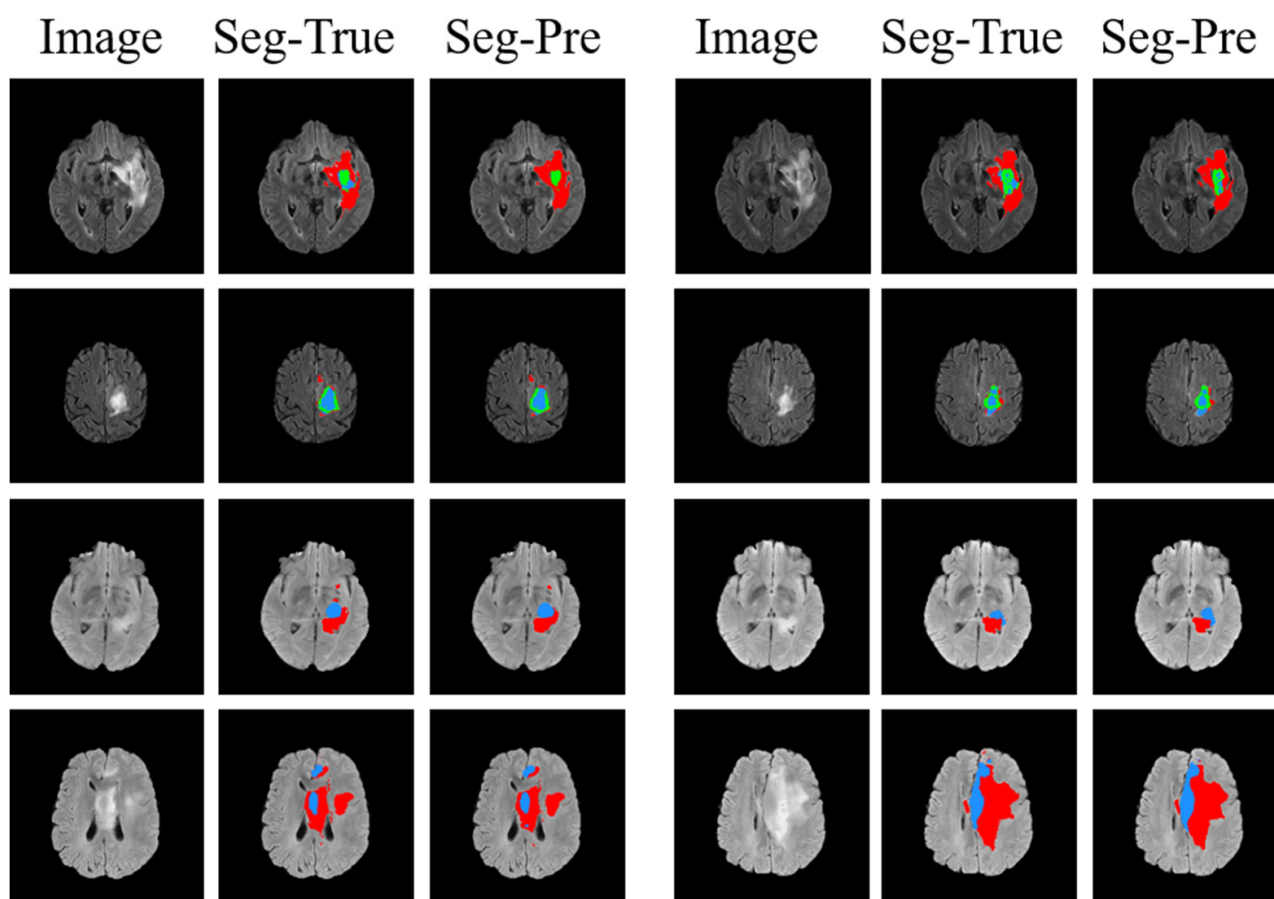


Figure 4. Comparison of brain tumor segmentation outputs and GT.

We compared some U-Net-like segmentation networks. Dice Similarity Coefficients (ET, WT, and TC) are shown in Table 1. Based on the results, 3D nnU-Net can complete the segmentation task well.

Table 1. Quantitative results of segmentation model on BraTS2018 dataset.

Method	ET	TC	WT
U-Net	0.8250	0.8473	0.9005
VAE U-Net	0.8145	0.8041	0.9042
nnU-Net	0.7945	0.8524	0.9119

4.3.2. Classification Results

In the classification model, we use the flair sequences of the dataset, combine it with the classification labels to classify HGG and LGG, and use the classification accuracy as a performance indicator. At present, many kinds of classification models and methods are available for brain tumors in the medical field. Although these classification networks improve the classification accuracy by adding LSTM block, EML block, or adopting multi-steps, multi-branch networks greatly increase the complexity of the model and computational loss. In our framework, the explainable heatmaps produced by the classification network are mainly the explanation and supplement of the framework. These results may be specific to some important MRI slices and regions, which should be used as an information supplement for the segmentation network. Therefore, the original intention of our design of the network is to be as simple and efficient as possible, rather than improving the accuracy rate only. We hope that the model can reduce the calculation loss of the overall frame as much as possible on the basis of maintaining performance. In the classification model, multi-modal reasoning is changed into single modal reasoning, which greatly accelerates the memory latency time and improves the training speed efficiency. As shown in Table 2, the RepOpt-B1 bone net used in this paper has overall advantages over common deep learning models in terms of speed, parameters and floating-point operations per second (FLOPs) on ImageNet dataset. The speed was measured in examples/second. The model with higher speed, higher FLOPs and fewer parameters is considered to be more efficient.

Table 2. Accuracy and training speed on ImageNet dataset (1080Ti BS = 128).

Model	Top-1 Accuracy	Speed	Params (M)	FLOPs (B)
ResNet-50	76.31%	719	25.53	3.9
ResNet-101	77.21%	430	44.49	7.6
VGG-16	72.21%	415	138.35	15.5
RepVGG-B1	78.42%	685	51.82	11.8
RepOpt-B1	78.48%	1254	51.82	11.9

We define the LGG as the positive and the HGG as the negative for evaluation. We use a data of 10 batches to generate indicators with five-fold cross-validation. The indicators of accuracy, precision, F1-score, specificity, and sensitivity are 95.46%, 94.66%, 90.73%, 98.32%, and 87.11%, respectively. The confusion matrix and the ROC curve of the classification model are shown in Figures 5 and 6, respectively.

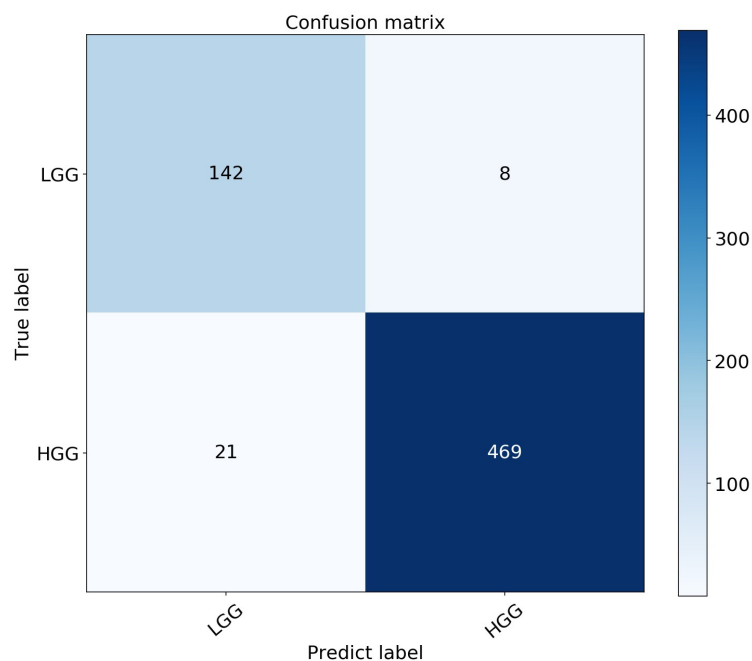


Figure 5. Confusion matrix of HGG and LGG.

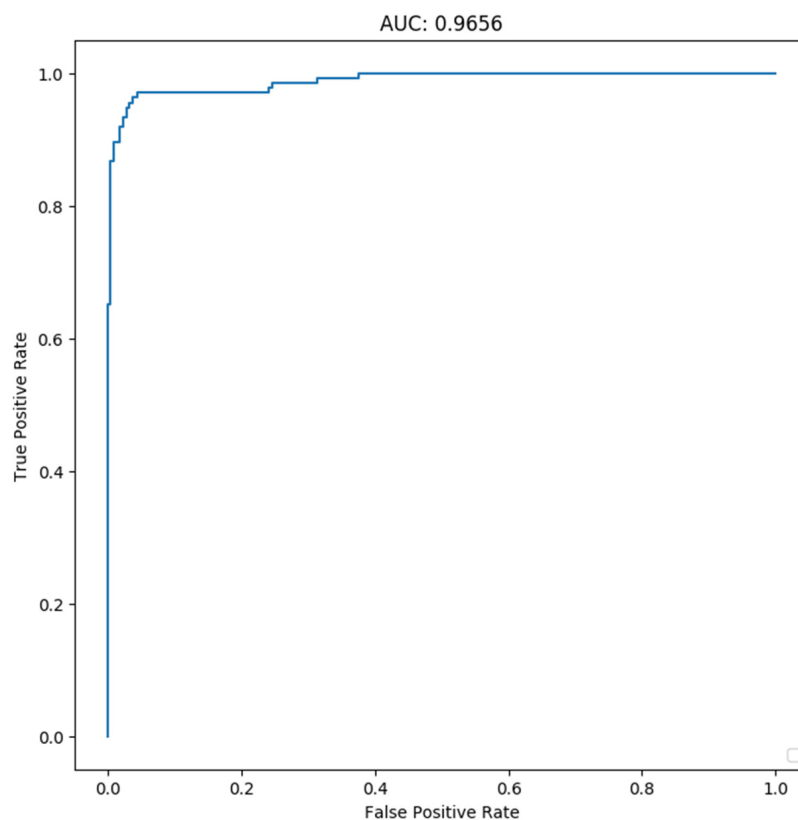


Figure 6. ROC curve of the brain tumor classification model.

We compared the accuracy of some excellent classification networks based on BraTS datasets. As shown in Table 3, although our classification network accuracy did not reach the highest score, it achieved good results.

Table 3. Quantitative results of classification model on BraTS dataset.

Model	Method	DataSet	Accuracy
Ge et al., 2018 [42]	2D CNNs	BraTS 2017	90.87%
Khan et al., 2020 [26]	VGG and EML	BraTS 2018	92.5%
Rehman et al., 2021 [43]	3D CNNs	BraTS 2018	92.67%
Dixit et al., 2022 [44]	FCM-IWOA-RBNN	BraTS 2018	96%
Our Model	RepOpt	BraTS 2018	95.46%

4.3.3. Comparison of Segmentation and Explainability

In this section, we observe and compare the output of the segmentation network and the explainability of the classification network. We adjusted the up-sampling of the explainable heatmaps to the same 240×240 size. Figure 7 shows the images of different layers of HGG and LGG patients, including the original input images (flair sequences), prediction results of segmentation model (Seg-Pre), the GT of segmentation model (Seg-True), explainable heatmaps from Grad-Cam++ (CAM++), and the superposition of segmentation output and output of explainable heatmaps (C+Pre). The color range of the gradient explainable heatmaps is from blue to red, which represents the saliency features from weak to strong. Therefore, the size and extent of the contribution of different regions of MRI to the prediction results of classification model can be observed intuitively, thereby facilitating the understanding of the model. Although the segmentation prediction outputs and explainable heatmaps come from two different deep model, the explainable heatmaps focus on the main lesion areas detected in the segmented image. As there are only classification labels and no segmentation GT, the network can find the lesion areas very well. A similar situation is also reflected in the MRI results of different layers of the same patient. At the same time, visualization results of these layers are also interrelated. As shown in the first and fourth lines (HGG1 and LGG1, respectively), when there are obvious WT and TC regions, the lesion areas detected by the two models are more similar. In the second and fifth lines (HGG2 and LGG2, respectively), certain differences of heatmaps exist between the two models when there are small areas of TC or there are less edemas around the enhancement region. The edema region of LGG3 is larger than HGG3, but the area of the explainable heatmaps is not much. The explainable heatmaps seems to pay more attention to the WT and TC regions. In general, the overall size of explainable heatmaps of HGG seems larger than LGG. The model seems to be able to identify the location of enhanced areas and non-enhanced areas and then classify HGG and LGG by channel features and weights. These observations illustrate the similarities and differences between the two models, which can help medical professionals further understand the model results and make better judgments.

4.4. Discussion

XAI has great potential for widespread application in the healthcare industry. It can help doctors make more informed diagnostic and treatment decisions, avoid bias and data misuse, and increase the credibility of artificial intelligence in healthcare. One important application scenario for explainability is smart terminals. Whether for medical professionals or patients, smart terminal devices need to make quick and accurate decisions. Therefore, explainable maps based on deep learning models need to minimize computation and inference time while ensuring accuracy to meet real-time requirements. In this scenario, the inference speed of deep models is very important.

In order to better verify the role of each part in our model, we performed ablation analysis on the BraTS dataset. We trained with the same loss function and input. In our tests, when changing the overall structure of the model to a 3×3 CNNs similar to VGG, the model accuracy was 87%, while using the VGG16 model had an accuracy of about 86%, with little difference between them. After removing the atrous convolution layers from our network, the model accuracy was about 94.5%. Therefore, we found that RepOpt contributed about 8% to accuracy and atrous convolution layers contributed about 1% to

accuracy. This is consistent with our expectations. RepOpt provides a feature extraction method similar to multi-branch models, effectively improving model accuracy. The atrous convolution layers are mainly used to provide a better receptive field and can generate more accurate heatmaps. At the same time, we tested inference speed for these models and found that they were not significantly different at about 920 samples/second. This shows that although our model has improved accuracy, the speed of our model during the inference stage is not much different from that of simple models.

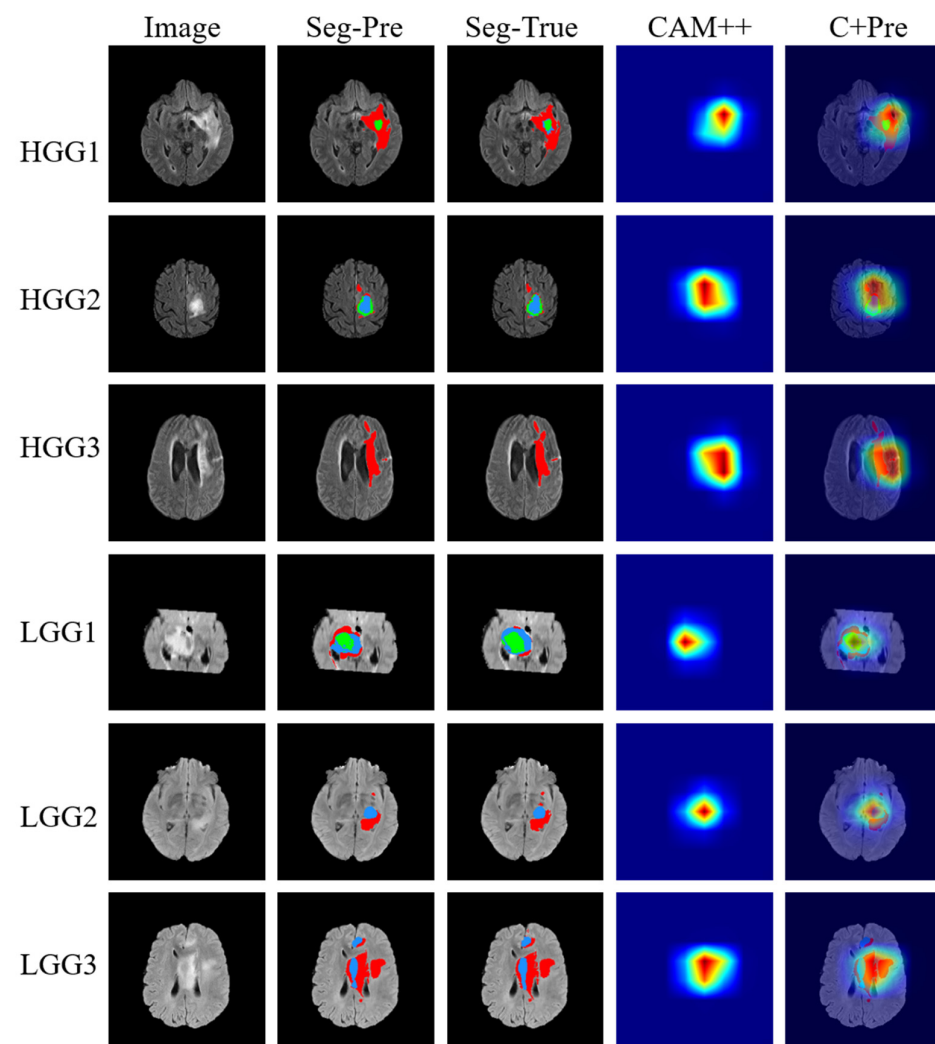


Figure 7. Comparison of brain glioma segmentation and explainability.

Although the explainability of medical images can help medical researchers better understand the deep model results, the current exploration and work of explainability are insufficient. First, explainability lacks datasets and evaluation indicators in medical images field. The explainability of medical images is often accompanied by segmentation or classification models, which cannot be trained independently. The BraTS dataset used in this paper is one of the few datasets that can provide segmentation GT and classification labels at the same time. At present, explainability is usually applied to image classification tasks, because explainability can be seen as a more intuitive visualization output. However, the explainability is similar to the segmentation result, which is a pixel-by-pixel classification result. The indicators of explainability and classification model are not equal. Although some scholars have made useful explorations on evaluation indicators [33], there are no recognized evaluation indicators for evaluating explainability in brain tumor diagnosis. Secondly, deep models depend on gradient information and input images. Therefore, the

models need to balance performance and computational loss, especially when dealing with high-dimensional data, such as 3D data. In addition, an important application of explainability is to shift from professional-oriented to user-oriented, and generate automatic detection reports in combination with natural language processing to provide help for end-to-end medical automatic solutions. Therefore, more expert system support and interdisciplinary knowledge may be needed.

5. Conclusions

XAI can deeply analyze deep models and find hidden information of the model and thus has a special and important role in the field of medical images. In this paper, we propose an explainable brain tumor detection framework, which combines segmentation and classification models and explainable methods for MRI brain tumor diagnosis. Our research focuses on the use of XAI to further explain the medical image results of advanced models, provide a more comprehensive interpretation perspective for the work of brain glioma grading and glioma localization based on deep learning, increase the credibility and enforceability of medical professionals on the results of deep learning models, and provide help for the automatic generation of medical reports combined with natural language processing.

Author Contributions: Conceptualization, F.Y. and R.X.; investigation, F.Y.; resources, Y.C. and Y.X.; writing—original draft preparation, F.Y.; writing—review and editing, R.X.; supervision, Z.W.; project administration, R.X.; funding acquisition, R.X. and Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by grants from National Natural Science Foundation of China (62176268), Beijing Natural Science Foundation-Joint Funds of Haidian Original Innovation Project (L202030), Major Science and Technology Project of Zhejiang Province Health Commission (WKJ-ZJ-2112), and Scientific and Technological Innovation Foundation of Shunde Graduate School of USTB (BK19BF004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jemal, A.; Thomas, A.; Murray, T.; Thun, M. Cancer statistics. *Ca-Cancer J. Clin.* **2002**, *52*, 23–47.
2. Miner, R.C. Image-guided neurosurgery. *J. Med. Imaging Radiat. Sci.* **2017**, *48*, 328–335. [[CrossRef](#)]
3. Isensee, F.; Kickingereder, P.; Wick, W.; Bendszus, M.; Maier-Hein, K.H. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In *International MICCAI Brainlesion Workshop*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 287–297.
4. Yang, G.; Ye, Q.; Xia, J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion* **2022**, *77*, 29–52. [[CrossRef](#)] [[PubMed](#)]
5. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [[CrossRef](#)]
6. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.-Z. XAI—Explainable artificial intelligence. *Sci. Robot.* **2019**, *4*, eaay7120. [[CrossRef](#)] [[PubMed](#)]
7. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608.
8. Tonekaboni, S.; Joshi, S.; McCradden, M.D.; Goldenberg, A. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of the Machine Learning for Healthcare Conference*, Ann Arbor, MI, USA, 8–10 August 2019; pp. 359–380.
9. Messina, P.; Pino, P.; Parra, D.; Soto, A.; Besa, C.; Uribe, S.; Andía, M.; Tejos, C.; Prieto, C.; Capurro, D. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv.* **2022**, *54*, 1–40. [[CrossRef](#)]
10. Temme, M. Algorithms and transparency in view of the new general data protection regulation. *Eur. Data Prot. Law Rev.* **2017**, *3*, 473–485. [[CrossRef](#)]
11. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should i trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.

12. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
13. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3319–3328.
14. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
15. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv* **2014**, arXiv:1412.6806.
16. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
17. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
18. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
22. Tian, J.; Li, C.; Shi, Z.; Xu, F. A diagnostic report generator from CT volumes on liver tumor with semi-supervised attention mechanism. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018; pp. 702–710.
23. Han, Z.; Wei, B.; Leung, S.; Chung, J.; Li, S. Towards automatic report generation in spine radiology using weakly supervised framework. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018; pp. 185–193.
24. Teixeira, L.O.; Pereira, R.M.; Bertolini, D.; Oliveira, L.S.; Nanni, L.; Cavalcanti, G.D.; Costa, Y.M. Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. *Sensors* **2021**, *21*, 7116. [\[CrossRef\]](#)
25. Ramzan, F.; Khan, M.U.G.; Iqbal, S.; Saba, T.; Rehman, A. Volumetric segmentation of brain regions from MRI scans using 3D convolutional neural networks. *IEEE Access* **2020**, *8*, 103697–103709. [\[CrossRef\]](#)
26. Khan, M.A.; Ashraf, I.; Alhaisoni, M.; Damaševičius, R.; Scherer, R.; Rehman, A.; Bukhari, S.A.C. Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists. *Diagnostics* **2020**, *10*, 565. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Yang, C.; Rangarajan, A.; Ranka, S. Visual explanations from deep 3D convolutional neural networks for Alzheimer’s disease classification. In Proceedings of the AMIA Annual Symposium Proceedings, San Francisco, CA, USA, 3–7 November 2018; pp. 1571–1580.
28. Wickstrøm, K.; Kampffmeyer, M.; Jenssen, R. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Med. Image Anal.* **2020**, *60*, 101619. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Esmaeili, M.; Vettukattil, R.; Banitalebi, H.; Krogh, N.R.; Geitung, J.T. Explainable artificial intelligence for human-machine interaction in brain tumor localization. *J. Pers. Med.* **2021**, *11*, 1213. [\[CrossRef\]](#)
30. Saleem, H.; Shahid, A.R.; Raza, B. Visual interpretability in 3D brain tumor segmentation network. *Comput. Biol. Med.* **2021**, *133*, 104410. [\[CrossRef\]](#)
31. Natekar, P.; Kori, A.; Krishnamurthi, G. Demystifying brain tumor segmentation networks: Interpretability and uncertainty analysis. *Front. Comput. Neurosci.* **2020**, *14*, 6. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity checks for saliency maps. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 9505–9515.
33. Pereira, S.; Meier, R.; Alves, V.; Reyes, M.; Silva, C.A. Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 106–114.
34. Narayanan, B.N.; De Silva, M.S.; Hardie, R.C.; Kueterman, N.K.; Ali, R. Understanding deep neural network predictions for medical imaging applications. *arXiv* **2019**, arXiv:1912.09621.
35. Isensee, F.; Jäger, P.; Full, P.; Vollmuth, P.; Maier-Hein, K. nnU-Net for Brain Tumor Segmentation in Brainlesion: Glioma. In Proceedings of the Multiple Sclerosis, Stroke and Traumatic Brain Injuries-6th International Workshop, BrainLes, Lima, Peru, 4 October 2020.
36. Yan, F.; Wang, Z.; Qi, S.; Xiao, R. A Saliency Prediction Model Based on Re-Parameterization and Channel Attention Mechanism. *Electronics* **2022**, *11*, 1180. [\[CrossRef\]](#)

37. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2021; pp. 13733–13742.
38. Ding, X.; Chen, H.; Zhang, X.; Huang, K.; Han, J.; Ding, G. Re-parameterizing Your Optimizers rather than Architectures. *arXiv* **2022**, arXiv:2205.15242.
39. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. Master's Thesis, University of Tront, Toronto, ON, Canada, 2009.
40. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
41. Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R.T.; Berger, C.; Ha, S.M.; Rozycki, M. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv* **2018**, arXiv:1811.02629.
42. Ge, C.; Gu, I.Y.-H.; Jakola, A.S.; Yang, J. Deep learning and multi-sensor fusion for glioma classification using multistream 2D convolutional networks. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 5894–5897.
43. Rehman, A.; Khan, M.A.; Saba, T.; Mehmood, Z.; Tariq, U.; Ayesha, N. Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture. *Microsc. Res. Tech.* **2021**, *84*, 133–149. [[CrossRef](#)] [[PubMed](#)]
44. Dixit, A.; Nanda, A. An improved whale optimization algorithm-based radial neural network for multi-grade brain tumor classification. *Vis. Comput.* **2022**, *38*, 3525–3540. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.