Deep Spatio-Spatial models for Classifying Brain Tumors in MR Images

Dr.Tshetiz Dahal
Medical Doctor /Clinical Researcher and Writer
Lugansk State Medical University, Luhansk Oblast, 93000 Luhansk, Ukraine

Submitted: 10-07-2022 Accepted: 23-07-2022

ABSTRACT
A brain tumor is a mass or cluster of abnormal cells in the brain that has the potential to spread to other tissues nearby and pose a serious threat to the patient's life. For effective treatment planning, a precise diagnosis is necessary, and the main imaging technique for determining the extent of brain tumours is magnetic resonance imaging. The majority of this increase in Deep Learning techniques for computer vision applications may be attributed to the availability of a sizable amount of data for model training and the advancements in model designs that produce better approximations in a supervised environment. The availability of free datasets with trustworthy annotations has significantly improved the classification of cancers using such deep learning techniques. These techniques often use either 3D models that employ 3D volumetric MRIs or even 2D models that take each slice into account separately. However, spatiotemporal models can be used as "spatio-spatial" models for this job by treating each spatial dimension individually or by seeing the slices as a succession of images through time. These models can learn certain spatial and temporal correlations while using less processing power.This study classifies several types of brain tumours using two spatiotemporal models, ResNet (2+1)D and ResNet Mixed Convolution. It was found that both of these models outperformed ResNet18, a model that only used 3D convolutions. It was also shown that pre-training the models on a distinct, even unrelated dataset before training them for the objective of cancer classification enhances performance. As a result of these studies, Pre-trained ResNet Mixed Convolution was shown to be the most effective model, achieving a macro F1-score of 0.9345 and a test accuracy of 96.98% while also being the model with the lowest computing cost.

I. INTRODUCTION

The growth of abnormal brain cells is known as a brain tumour. Based on their rate of growth and likelihood of recurrence following therapy, brain tumours are categorized. They can be broadly classified into two groups: malignant and benign. Noncancerous benign tumours spread slowly and are less likely to come back following therapy. Contrarily, malignant tumours, which are mostly composed of cancer cells, can either locally infiltrate tissues or migrate to other parts of the body through a process known as metastasis1. Mutations in glial cells cause malignancy in normal cells, which results in glioma tumours. They make about 80 % of the total of all malignant tumours and 30 % of all brain and central nervous system tumours, making them the most prevalent forms of astrocytomas (brain or spinal cord tumors)2. Glioma tumours can have Astrocytomas, Oligodendrogliomas, or Ependymomas as its phenotypic makeup. The World Health Organization (WHO) employs the following grading-based methodology to categorize each of these tumours depending on their aggressiveness:

Grade I - This tumours are frequently detected in children and are typically benign tumours, which means they are usually treatable.

Grade II - Three tumour types fall within the grade II category: oligodendrogliomas, oligoastrocytomas, and a combination of both. Adults commonly experience them. All low-grade gliomas have the potential to develop into high-grade tumours over time .

Grade III - Anaplastic Astrocytomas, Anaplastic Oligodendrogliomas, and Anaplastic Oligoastrocytomas are all examples of grade III tumours. They are more sneaky and aggressive than grade II.

Grade IV - Glioblastoma Multiforme (GBM), another name for grade IV glioma, is the most aggressive tumour according to the WHO classification.

In general, grades I and II gliomas are considered low-grade gliomas (LGG), while grades III and IV are known as high-grade glioma (HGG). The LGG are benign tumours, and they can be excised using surgical resection. In contrast, HGGs are malignant tumours that are hard to excise by surgical methods because of their extent of nearby tissue invasion. Figure 1 shows an example MRI of LGG and HGG.

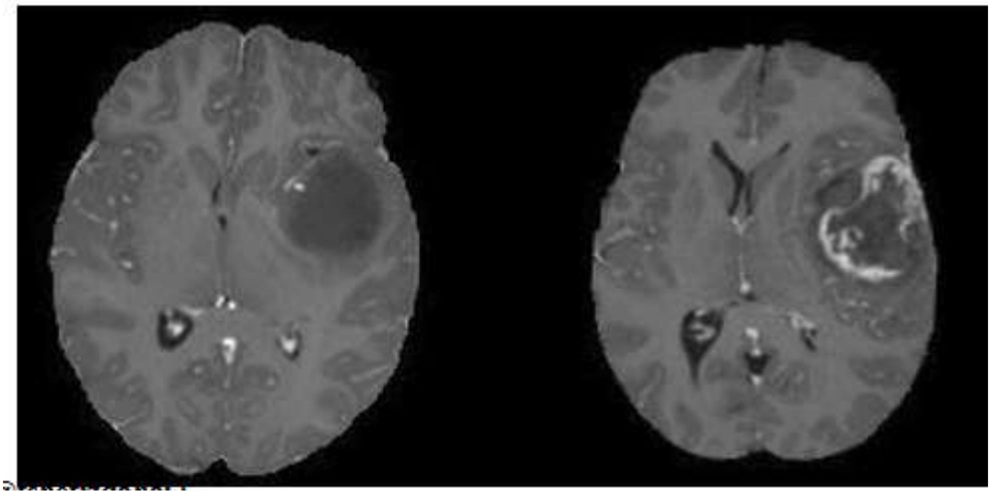


FIGURE 1 : An example MRI of Low-grade glioma (LGG, on the left) and High-grade glioma (HGG, on the right).

A Glioblastoma Multiforme (GBM) typically has the following types of tissues (shown in Fig. 2):

: **The Tumour Core:** This is the region of the tumour that has the malignant cells that are actively proliferating.

: **Necrosis:** The necrotic region is the important distinguishing factor between low-grade gliomas and GBM4. This is the region where the cells/tissue are dying, or they are dead.

: **Perifocal oedema:** The swelling of the brain is caused by fluid build-up around the tumour core, which increases the intracranial pressure; perifocal oedema is caused by the changes in glial cell distribution5.

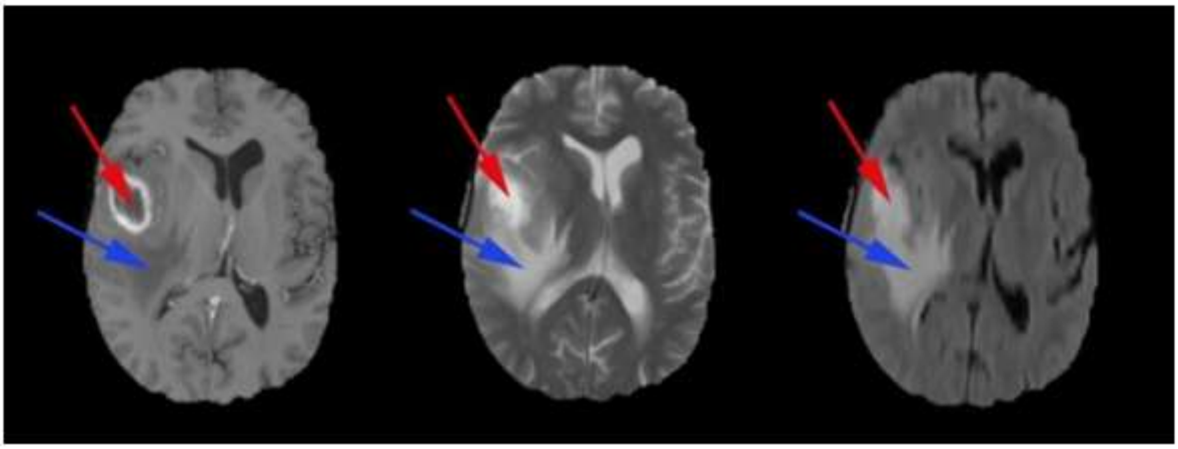


FIGURE 2 : High-grade glioma structure on T1ce, T2 and FLAIR contrast images (from left to right), (red circle) Necrotic core, (blue circle) Perifocal oedema.

The location, histological subtype, and tumour margins are just a few of the variables that affect a brain tumor's prognosis. Even after treatment, the tumour frequently returns and advances to grade IV3. The site of the tumour can be determined using contemporary imaging techniques like MRI, which is then utilised to investigate tumour progression and arrange surgical procedures. Along with its hemodynamics, MR imaging is used to evaluate the anatomy, physiology, and metabolic activity of the lesion. As a result, MR imaging continues to be the major method for diagnosing brain tumours. Early cancer identification in particular has the potential to alter how a patient is treated. Early detection is essential because lesions that are detected earlier are more likely to be treatable; if action is taken, this might be the difference between life and death. By prioritizing only malignant lesions, deep learning techniques might lessen the strain of radiologists reading numerous images and assist in automating the process of finding and classifying brain lesions. This can decrease diagnostic errors6 and eventually increase overall efficiency. Recent research has demonstrated that deep learning techniques in radiography have already surpassed human performance levels for some pathologies7.

BACKGROUND

Recently, a number of deep learning-based approaches for classifying brain tumours have been presented. T1 contrast-enhanced images were used by Mzoughi et al.8 to suggest a method for classifying high-grade and low-grade gliomas. Pei et al.9 performed a similar study on the classification of gliomas based on grading, segmenting the tumour first before classifying it as either HGG or LGG. The majority of the literature on the classification and grading of glioma tumours employed a single MR contrast image at a time, however Ge et al.10 used a fusion framework that simultaneously classifies the tumour using T1 contrast-enhanced, T2, and FLAIR images. The non-subsampled shearlet transform (NSST) was used by Ouerghi et al.11 to transform T1 images into low frequency (LF) and high frequency (HF) sub-images, effectively separating principle information from edge information in the source image. The images were then fused according to predefined rules to include the coefficients, resulting in the fusion of T1 and T2 or FLAIR images. The majority of the literature simply distinguishes between the various grades of tumours and does not include healthy brains as a separate category.

One of the most effective network topologies for image identification tasks, ResNet or residual network, was proposed by He et al.12 and addresses issues with deep networks, such as disappearing gradients. The identity mappings known as residual-links, or "skipped connections," are introduced in this study. Their outputs are appended to the outputs of the other stacked layers. These identification links enhance the training process without increasing network complexity. The spatiotemporal models for action recognition developed by Tran et al.13 are essentially 3D convolutional neural networks built on ResNet. Video data is three-dimensional since it has two spatial dimensions and one time dimension. It is clear that utilizing a network with 3D convolution layers is the best option for processing such data (such as an action detection task). ResNet (2+1)D and ResNet Mixed Convolution are two different types of spatiotemporal models that Tran et al.13 introduced. In the ResNet(2+1)D model, 2D and 1D convolutions are employed, with the 2D convolutions being used for the spatial component and the 1D convolutions being saved for the temporal component. By utilizing non-linear rectification, this provides an advantage of greater non-linearity and makes this type of mixed model more "learnable" than traditional complete 3D models. The ResNet Mixed Convolution model, on the other hand, is built using a combination of 2D and 3D Convolution processes. The model's first layers are constructed using 3D convolution techniques, whereas its subsequent layers use 2D convolutions. The justification for this setup is that since most motion-modelling takes place in the first few layers, using 3D convolution there better captures activity. Transfer learning14 is a method widely employed to boost the performance of the same network architecture in addition to trying to enhance the architecture itself. This method allows you to use a model that has already been trained to perform one task to perform another task entirely. Before beginning the training, the model parameters are typically initialized at random. Transfer learning, on the other hand, trains the model for task two using model parameters learnt from task one as the starting point (referred to as pre-training), rather than random values. Pre-training has proven to be a useful technique for enhancing the initial training process, ultimately leading to higher accuracy. 15,16.

CONTRIBUTION

For three dimensional video classification applications, spatiotemporal models are frequently employed. Their potential for identifying "spatiospatial" models, such as 3D volumetric pictures like MRIs, has not yet been investigated. This examines the potential for using the spatiotemporal models ResNet(2+1)D and ResNet Mixed Convolution as "spatiospatial" models by treating the slice dimension of the three-dimensional volumetric pictures differently from the other two spatial dimensions. Using a single MR contrast, "Spatial" was used to classify brain tumours of various glioma types based on their grade as well as healthy brains from 3D volumetric MR Images. Their performances were compared to a pure 3D convolutional model (ResNet3D). The models will also be contrasted with and without pre-training in order to assess the usefulness of transfer learning for this purpose.

II. METHODOLOGY

The network models utilized in this study are covered in detail in this section along with implementation information, pre-training and training methodologies, data augmentation approaches, dataset details, data pre-processing procedures, and evaluation metrics.

NETWORK MODELS

For tasks using video where there are two spatial and one temporal dimension, spatiotemporal models are typically used. These models, as opposed to pure 3D convolution-based models, handle the spatial and temporal dimensions in various ways. A 3D convolution-based model is frequently used since 3D volumetric image classification tasks lack a time component. They are occasionally cut into 2D slices and subjected to 2D convolution-based models. In order to make the convolution kernels invariant to tissue discrimination in all dimensions and learn more complicated characteristics across voxels, 3D filters are used for the purpose of classifying tumours. 2D convolution filters will be used to capture the spatial representation within the slices.

Spatiotemporal models can either reduce the complexity of the model or provide more non-linearity by combining two different forms of convolution into one model. Consider the spatiotemporal models as "spatiospatial" models in order to take use of these benefits while working with volumetric data; this is the rationale for utilizing such models for a tumour classification task. In this study, in-plane dimensions are taken as the spatial dimensions while slice-dimension is treated as the pseudo-temporal dimension of spatiotemporal models. The work of Tran et al.13 served as the foundation for the spatiotemporal models used here as spatial models. ResNet (2+1)D and ResNet Mixed Convolution are two alternative spatiospatial models that are investigated in this article. Their performances are contrasted with ResNet3D, a model that only uses 3D convolutions.

ResNet (2+1)D

Instead of using a single 3D convolution, ResNet (2+1)D employs a combination of 2D convolution and 1D convolution. As opposed to employing a single 3D Convolution13, this setup has the advantage of allowing an additional non-linear activation unit between the two convolutions. The network's overall number of ReLU units then rises as a result, enabling the model to learn even more complicated functions. The ResNet(2+1)D employs a stem that consists of a 1D convolution with a kernel size of three and a stride of one, followed by a 2D convolution with a kernel size of seven and a stride of two, receiving one channel as an input and producing 45 channels as the output. The following set of blocks includes four convolutional blocks, each of which has two sets of fundamental residual blocks. A 2D convolution with a kernel size of three and a stride of one is found in each residual block, followed by a 1D convolution with a kernel size of three and a stride of one. A 3D batch normalization layer, followed by a ReLU activation function, follows each convolutional layer in the model—both the 2D and the 1D versions. To down-sample the input by half, a pair of 3D convolution layers with a kernel size of one and a stride of two are used to separate the residual blocks inside the convolutional blocks, with the exception of the first convolutional block. The 1D convolutions are applied on the slice dimension, whereas the 2D convolutions are applied in-plane. An adaptive average pooling layer with an output size of one for all three dimensions has been introduced after the last convolutional block.

After the pooling layer, a dropout layer followed by a fully connected layer with n output neurons for n classes were added to obtain the final output. Figure 3(a) portrays the schematic diagram of the ResNet (2+1)D architecture.

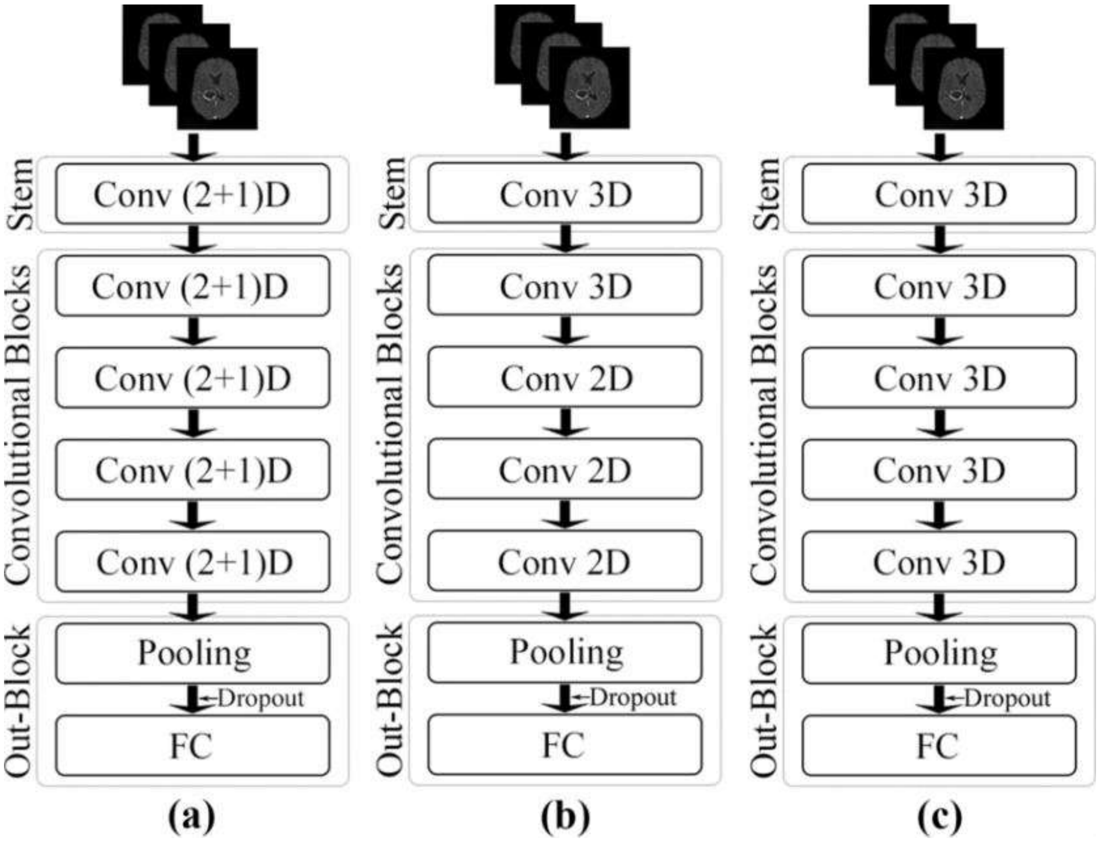


FIGURE 3 : Schematic representations of the network architectures. **(a)** ResNet (2+1)D, **(b)** ResNet Mixed Convolution, and **(c)** ResNet 3D.

ResNet mixed convolution

A combination of 2D and 3D convolutions are used in ResNet Mixed Convolution. This model's stem includes a 3D convolution layer with a kernel size of (3,7,7), a stride of (1,2,2), and a padding of (1,3,3), where the first dimension is the slice dimension and the other two are the in-plane dimensions. This layer accepts a single channel as input and outputs 64 channels. Three 2D convolution blocks come after the stem, then one 3D convolution block. All convolution layers, both 3D and 2D, share the same three-kernel size and one-stride parameters. Each of these residual blocks has two convolution layers, and each of these convolution blocks has two residual blocks. Similar to ResNet (2+1)D, a set of 3D convolution layers with a kernel size of one and a stride of two are used to divide the residual blocks inside the convolutional blocks, with the exception of the first convolutional block, to down-sample the input by half. A 3D batch normalization layer and a ReLU activation function are placed after each convolutional layer in the model, both 3D and 2D. The rationale behind utilizing both 2D and 3D modes of convolution is that although 2D can learn representation inside each 2D slice, 3D filters can learn the spatial properties of the tumour in 3D space. The final pooling, dropout, and fully connected layers follow the convolutional blocks and are the same as those in the ResNet (2+1)D architecture. Figure 3(b) shows the schematic representation of this model.

ResNet3D

A pure 3D ResNet model is used as the benchmark to compare the performance of the spatiospatial models against (c). With the exception of the fact that this model exclusively employs 3D convolutions, the ResNet3D model's design is nearly identical to that of ResNet Mixed Convolution (see "Network models" section). The main variation between these models stems from the usage of four 3D convolution blocks in this model as opposed to one 3D convolution block, followed by three 2D convolution blocks, in

ResNet Mixed Convolution. A 3D ResNet18 model is created using this ResNet3D architectural setup.

III. SUMMARY AND COMPARISON

The input proceeds to the stem, followed by four convolutional blocks, the output block—which has an adaptive pooling layer, then a dropout layer, and lastly a fully connected layer. This is the overall structure of the network models. The stem of ResNet Mixed Convolution and ResNet 3D is identical and consists of a 3D convolutional layer with a kernel size of (3,7,7), a batch normalization layer, and a ReLU. ResNet (2+1)D employs a distinct stem that consists of a 2D convolution layer with a kernel size of seven followed by a 1D convolution layer with a kernel size of three, separating the 3D convolution (3,7,7) used by the other models into a pair of 2D and 1D convolution layers (7,7) and (3,3). (3). A batch normalization layer and ReLU pair are followed by both 2D and 1D convolution inside of this stem. The convolutional blocks in the ResNet3D and ResNet Mixed Convolution designs have the same structure: two residual blocks made up of two sub-blocks, each of which has a 3D convolution with a three-kernel size, followed by a batch normalization layer and a ReLU. As opposed to the 3D convolutional layers used by the other models, the initial convolutional block of the ResNet (2+1)D architecture uses a pair of 2D and 1D convolutions with a three kernel size. The rest of the building is identical. Because the 3D convolutions are divided into a pair of 2D and 1D convolutions, it is noted that this model has more non-linearity than others. Additional pairs of batch normalization and ReLU may have been utilised between the 2D and 1D convolutions. The second, third, and fourth convolutional blocks all contained a downsampling pair, which was composed of a 3D convolutional layer with a kennel size of one and a stride of two, followed by a batch normalization layer. This downsampling pair was included in the first convolutional block, but not in the other three blocks (applicable to all three models). In the first convolutional block, this was absent. The number of input features to the first block is 64, while the number of output features to the fourth (and final) block is 512. The convolution blocks of each of the three models double the input features by two. In the last stage of each of these models, an adaptive average pooling layer imposes a 1×1×1, output form with 512 distinct features.

A dropout with a probability of 0.3 is then applied to introduce regularization to prevent over-fitting before supplying them to a fully connected linear layer that generates n classes as output. The width and depth of these models are comparable, but they differ in terms of the number of trainable parameters depending upon the type of convolution used, as shown in Table 1. It is noteworthy that the less the number of trainable parameters - the less the computational costs. A model with a lesser number of parameters would require lesser memory for computation (GPU and RAM), and also the complexity of the model is lesser—reducing the overall computational costs for both training and inference. Moreover, a lesser number of trainable parameters would also reduce the risk of overfitting.

TABLE 1 Total number of trainable parameters for each model.

Model	**No. of parameters**
ResNet3D	33,150,522
ResNet (2+1)D	31,297,254
ResNet mixed convolution	11,472,963

IV. IMPLEMENTATION AND TRAINNG

The Torchvision models were modified and implemented using PyTorch18. An Nvidia RTX 4000 GPU with 8 GB of memory was used for training with a batch size of 1. Models with and without pre-training were contrasted. On Kinetics-40020, all models with pre-training had been trained, with the exception of the stems and fully connected layers. The 3D volumetric MRIs only have one channel, but the RGB images from the Kinetics dataset have three channels. As a result, the stem that had been trained on the Kinetics dataset was unable to be applied and was initialized at random. The fully linked layer was additionally initialized with random weights because Kinetics-400 offers 400 output classes whereas the task at hand only requires three (LGG, HGG, and Healthy). Trainings were performed using mixed-precision21 with the help of Nvidia's Apex library22. The loss was calculated using the weighted cross-entropy loss function to minimize the under-representation of classes with fewer samples during training and was optimized using the Adam optimizer with a learning rate of 1e−5 and weight decay coefficient λ=1e−3.

Weighted cross-entropy loss

The normalized weight value for each class (**Wc**) is calculated using:

$$W_c = \left[1 - \left(\frac{samples_c}{\Sigma samples_t}\right)\right]$$

Where sample is the number of samples from class c and samplest are the total number of samples from all classes. The normalized weight values from this equation is then used to scale cross-entropy loss of the respective class loss:

$$loss_c = W_c\left[-x_c \log(P(c))\right]$$

Where xc is the true distribution and P(c) is the estimate distribution for class c. The total cross-entropy loss then is the sum of individual class losses.

$$Loss_{total} = loss_{c_1} + loss_{c_2} + loss_{c_2} + \ldots + loss_{c_n}$$

Data Augmentation

Before training the models, different data augmentation techniques were applied to the dataset, and TorchIO23 was utilised for that. Light and heavy augmentation were used in the initial experiments, with light augmentation consisting solely of random affine transformations (scale 0.9-1.2, degrees 10) and random flips (L-R, probability 0.25) and heavy augmentation consisting of the latter two as well as elastic deformation and random k-space transformations (motion, spike, and ghosting). In addition to having poor final accuracy, it was shown that the loss took substantially longer to converge when the network was trained using heavily augmented input. Therefore, in this study, relatively minimal augmentation was applied.

Dataset

In this study, two different datasets were used: the non-pathological images were taken from the IXI Dataset26, and the pathological images were taken from the Brain Tumour Segmentation (BraTS) 2019 dataset, which includes images with four different MR contrasts (T1, T1 contrast-enhanced, T2, and FLAIR). T1 contrast-enhanced (T1ce) MRIs, one of the four types of MRIs available, are most frequently employed for single-contrast tumour classification8,27. Consequently, 332 participants' T1ce images from the BRaTS collection were used in this study: 259 volumes of high-grade glioma (HGG) and 73 volumes of low-grade glioma (LGG). To have the same number of subjects as HGG, 259 T1 weighted volumes were chosen at random from the IXI dataset as healthy samples. The final combined dataset was then randomly divided into 3-folds of training and testing split with a ratio of 7:3.

Data pre-processing

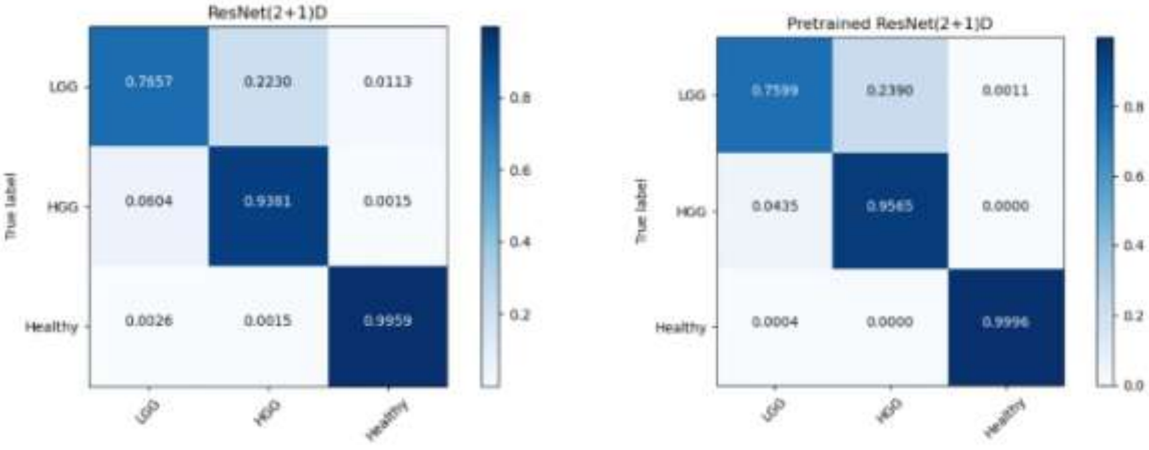
The brain extraction tool (BET2) of FSL28,29 was used as the first step in the pre-processing of the IXI images. As the BraTS photos are already skull stripped, this was done to maintain consistency throughout the input data. As employed by Isensee et al.30, the intensity values of all the volumes from the combined datasets were additionally normalized by scaling intensities to the [0.5,99.5] percentile. Finally, the volumes were re-sampled with the same 2 mm isotropic voxel-resolution.

Evaluation Metrics

The performance of the models was compared using precision, recall, F1 score, specificity, and testing accuracy. Furthermore, a confusion matrix was used to show class-wise accuracy.

V. RESULTS

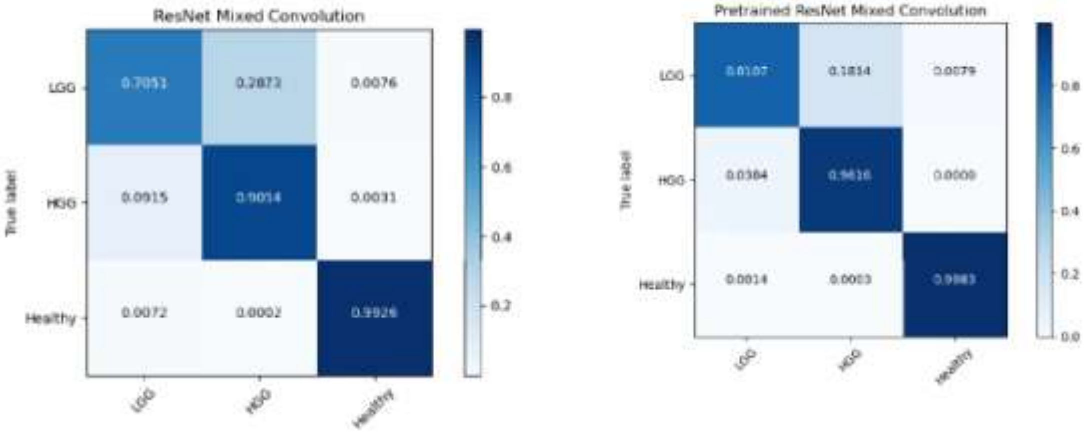
Comparisons were made between the models' performances with and without pre-training. Figures 4, 5, and 6 present, for ResNet (2+1)D, ResNet Mixed Convolution, and ResNet 3D, respectively, the average accuracy over 3-fold cross validation using confusion measures.



(a) Without pre-training

(b) With pre-training

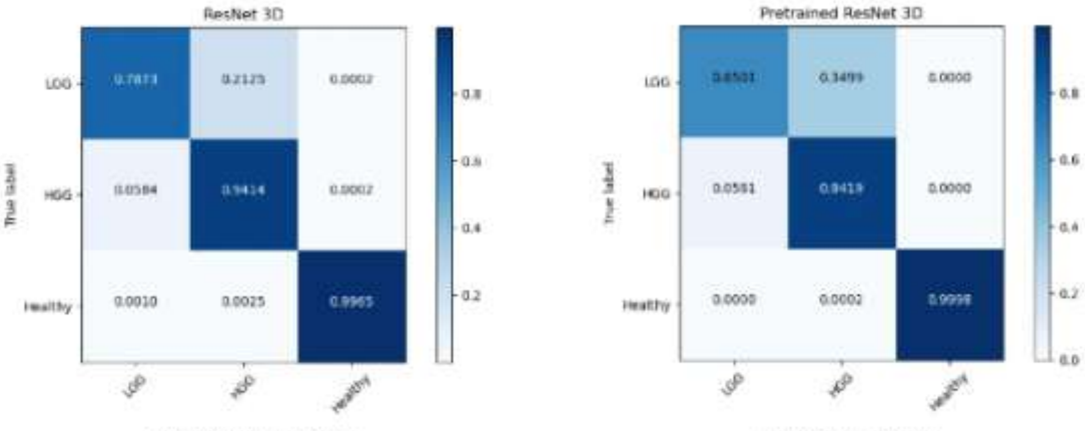
FIGURE 4 : Confusion matrix for 3-fold cross-validation on pre-trained ResNet(2+1)D.



(a) Without pre-training

(b) With pre-training

FIGURE 5 : Confusion matrix for 3-fold cross-validation on ResNet mixed convolution.



(a) Without pre-training

(b) With pre-training

FIGURE 6 : Confusion matrix for 3-fold cross-validation on ResNet3D18.

Figure 7 shows the class-wise performance of the different models, both with and without pre-training, using precision, recall, specificity, and F1-score.

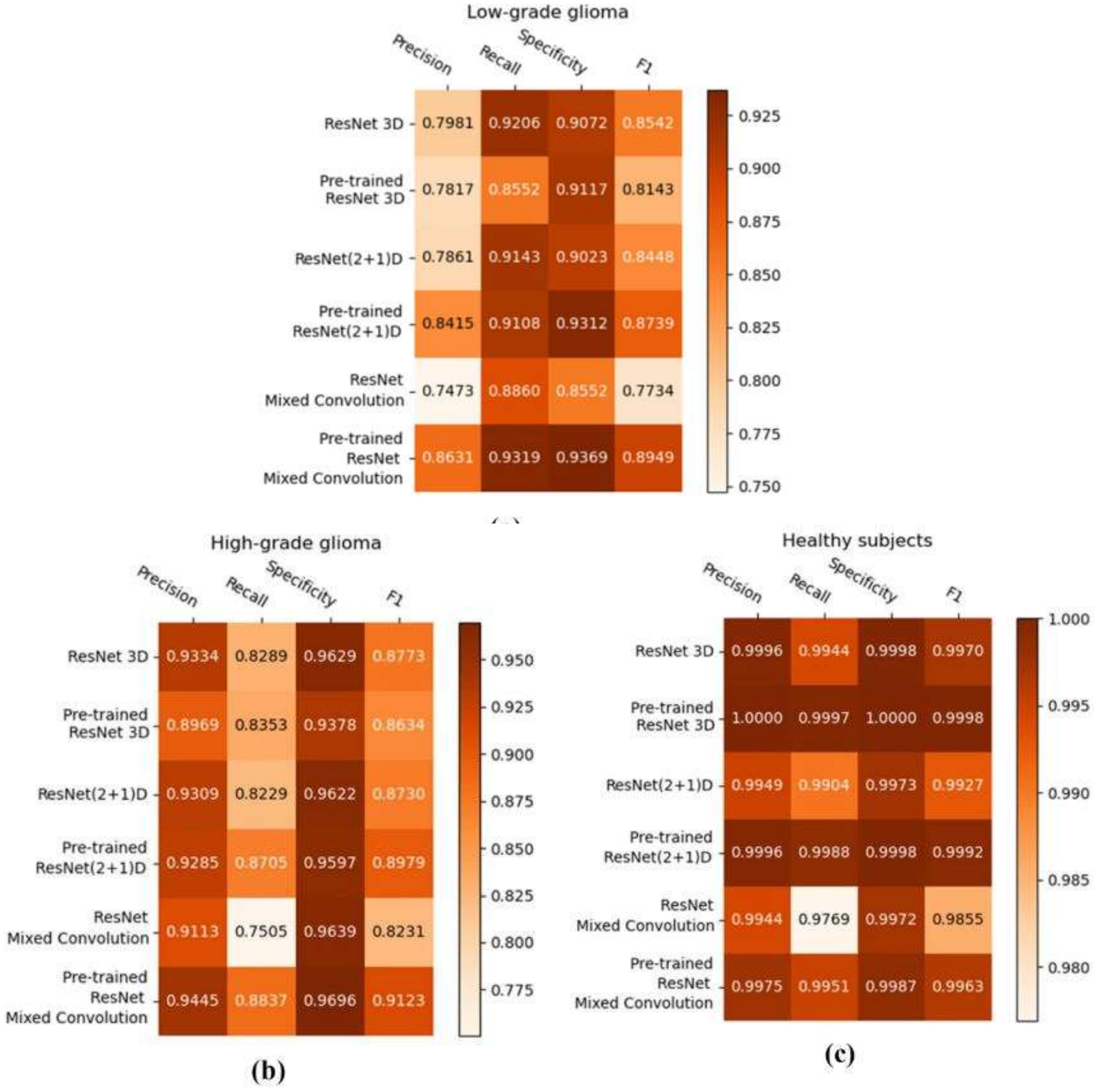


FIGURE 7 : Heat-maps showing the class-wise performance of the classifiers, compared using precision, recall, specificity, and F1-score: **(a)** LGG, **(b)** HGG, and **(c)** healthy.

Comparison of the Models

The mean F1-score over 3-fold cross-validation was used as the metric to compare the performance of the different models. Tables 2, 3 and 4 show the results of the different models for the classes LGG, HGG, and Healthy, respectively; and finally Table 5 shows the consolidated scores.

TABLE 2 : Low-grade glioma model comparison (*denotes the overall winning model).

Low-grade glioma	
Model	**Mean F1 score**
ResNet 3D	0.8542 ± 0.049
Pre-trained ResNet 3D	0.8143 ± 0.048
ResNet(2+1)D	0.8448 ± 0.019
Pre-trained ResNet(2+1)D	0.8739 ± 0.041

ResNet mixed convolution	0.7734 ± 0.031
Pre-trained ResNet mixed convolution	0.8949 ± 0.033

For low-grade glioma (LGG), ResNet Mixed Convolution with pre-training achieved the highest F1 score of 0.8949 with a standard deviation of 0.033. The pre-trained ResNet(2+1)D is not far behind, with 0.8739 ± 0.033.

TABLE 3 : High-grade glioma model comparison (*denotes the overall winning model).

High-grade glioma	
Model	**Mean F1 score**
ResNet 3D	0.8773 ± 0.034
Pre-trained ResNet 3D	0.8634 ± 0.042
ResNet(2+1)D	0.8730 ± 0.022
Pre-trained ResNet(2+1)D	0.8979 ± 0.032
ResNet mixed convolution	0.8231 ± 0.027
Pre-trained ResNet mixed convolution	0.9123 ± 0.029

For the high-grade glioma (HGG) class, the highest F1 was achieved by the pre-trained ResNet Mixed Convolution model, with an F1 score of 0.9123 ± 0.029. This is higher than the best model's F1 score for the class LGG. This can be expected because of the class imbalance between LGG and HGG. As with low-grade glioma, the second-best model for HGG is also the Pre-trained ResNet(2+1)D with the F1 score of 0.8979 ± 0.032.

TABLE 4: Healthy brain model comparison (*denotes the overall winning model).

Healthy brain	
Model	**Mean F1 score**
ResNet 3D	0.9970 ± 0.005
Pre-trained ResNet 3D	0.9998 ± 0.0002
ResNet(2+1)D	0.9927 ± 0.009
Pre-trained ResNet(2+1)D	0.9992 ± 0.001
ResNet mixed convolution	0.9855 ± 0.004
Pre-trained ResNet mixed convolution	0.9963 ± 0.002

The healthy brain class achieved the highest F1 score of 0.9998 ± 0.0002, with the pre-trained ResNet 3D model, which can be expected because of the complete absence of any lesion in the MR images making it far less challenging for the model to learn and distinguish it from the brain MRIs with pathology. Even though the pre-trained ResNet 3D model achieved the highest mean F1 score, all pre-trained models achieved similar F1 scores, i.e. all the mean scores are more than 0.9960—making it difficult to choose a clear winner.

TABLE 5 : Consolidated comparison of the models (*denotes the overall winning model).

Consolidated scores		
Model	**Macro F1 score**	**Weighted F1 score**
ResNet 3D	0.9095	0.9269
Pre-trained ResNet 3D	0.8925	0.9171
ResNet(2+1)D	0.9035	0.9220
Pre-trained ResNet(2+1)D	0.9237	0.9393
ResNet mixed convolution	0.8607	0.8881
Pre-trained ResNet mixed convolution	0.9345	0.9470

ResNet Mixed Convolution with pre-training came up as the best model for both classes with pathology (LGG and HGG) and achieved a similar score as the other models while classifying healthy brain MRIs, as well as based on macro and weighted F1 scores - making this model as the clear overall winner. It can also be observed that the spatiospatial models performed better with pre-training, but ResNet 3D performed better without pre-training.

Comparison against Literature

TABLE 6 : Comparisons against other published works *s=specificity † cross-validated ‡ State of the art.

Study	Method	Contrast	Dimension	Test Accuracy
Shahzadi et al.31	CNN with LSTM	T2-FLAIR	3D	84.00 %
Pei et al.9	Similar to U-Net for Segmentation, Regular CNN for classification	T1, T1ce, T2, T2-FLAIR	3D	74.9%
Ge et al.10	Deep CNN	T1, T2, T2-FLAIR	2D	90.87%
Yang et al.27	Pre-trained GoogLeNet	T1ce	2D	94.5% †
Mzoughi et al.8	Deep CNN	T1ce	3D	96.49%
Zhuge et al.32	Deep CNN	T1, T1ce, T2, T2-FLAIR	3D	97.1% s*=0.968
Ouerghi et al.11 ‡	Random forest	T1, T2, T2-FLAIR	2D	96.5%
This paper	Pre-trained ResNet mixed convolution spatiospatial model	T1ce	3D	$96.98\%^{\dagger}s = 0.9684$

This subsection compares seven more research papers that classified LGG and HGG tumours against the top model from the preceding subsection (ResNet Mixed Convolution with pre-training). Since mean test accuracy was the most often used statistic in those articles, it was utilised as the metric to compare the outcomes.

Beginning with Shahzadi et al.31, who employed T2-FLAIR images from the BraTS 2015 dataset and LSTM-CNN to distinguish between HGG and LGG. Their research focused on employing a smaller sample size, and they were successful in achieving an accuracy rate of 84.000%31. Pei et al.9, who used all of the contrasts in the BraTS dataset and segmented their data using a model akin to the U-Net before doing classification, nonetheless only managed to reach a classification accuracy of 74.9 percent. Ge et alstrategy .'s of simultaneously training several streams utilizing multiple contrasts is new. On all the contrasts, their model had an overall accuracy of 90.87 percent, and on T1ce, it had an accuracy of 83.73 percent.

Deep convolutional neural networks were used by Mzoughi et al.8 to reach 96.59 percent on T1ce images. It is challenging to compare their conclusions to other research because their study only provides the overall accuracy of their model as a metric for their findings. Using pre-trained GoogLeNet on 2D images, Yang et al.27 carried out subsequent research, attaining an overall accuracy of 94.5 percent.

Although they did not use the BraTS dataset, the goal of their work was to categorize glioma tumours according to LGG and HGG grading. In comparison to our research, their dataset contained less samples of the LGG and HGG classes, with the former having 52 samples and the latter having 61 samples27. In their article, Ouerghi et al.11 used a variety of machine learning techniques to train on fusion images, including the random forest technique, on which they were able to classify high-grade and low-grade gliomas with an accuracy of 96.5 percent.

Finally, Zhuge et al.32 surpassed the suggested model by 0.12 percent and reached an outstanding 97.1 percent utilizing Deep CNN for classification of glioma based on LGG and HGG grading. This discrepancy can be attributed to two factors: 1) their use of BraTS 2018 in conjunction with an extra dataset from The Cancer Imaging Archive, and 2) their use of four different contrasts, both of which greatly expand the training set. Furthermore, their publication has no reports of cross-validation. The complete comparison data are displayed in Table 6.

VI. DISCUSSION

As separating healthy brains from brains with pathology is, in comparison, a simpler task than determining the grade of the tumour, all of the models' F1 scores for classifying healthy brains were extremely close to one. Furthermore, MRIs may have caused a dataset bias by employing two distinct datasets for healthy and sick brain. The pre-trained ResNet Mixed Convolution model fared best at grading tumours, whereas all three pre-trained models performed similarly at grading healthy brains. Macro and weighted F1 scores were utilised to compare the models based on aggregated scores. However, because the dataset was unbalanced, the macro F1 score should be given more weight. Both of the metrics declared the pre-trained ResNet Mixed Convolution as the clear winner.

The classification performance of the models for the LGG class has been less successful than the other two classes, which is an intriguing finding from the confusion matrices. Even the top model could only achieve an accuracy of 81 percent for LGG, 96 percent for HGG, and almost perfect results for healthy. This could be explained by the dataset's extreme imbalance, which included 259 volumes for HGG and healthy while only having 73 volumes for LGG (see "Dataset" section).

Even though weighted cross-entropy loss (“Weighted cross-entropy loss” section) was used in this research to deal with the problem of class imbalance, increasing the number of LGG samples or employing further techniques to deal with this problem further and might improve the performance of the models for LGG33.

Notably, despite having the fewest trainable parameters of any model, the pre-trained ResNet Mixed Convolution produced the best classification performance (see Table 1). It should also be highlighted that both spatiospatial models outperformed the pure 3D ResNet18 model, despite the fact that they had less trainable parameters. Less trainable parameters can result in lower computational costs and a lower likelihood of overfitting. The increase in non-linearity brought about by the additional activation functions between the 2D and 1D convolutions in the (2+1)D convolutional layers, according to the author, helped the ResNet (2+1)D model outperform ResNet3D, and the reduction of trainable parameters while maintaining the same number of layers helped the ResNet Mixed Convolution model succeed. The spatial relationship between the three dimensions is not preserved within the network like a fully 3D network as ResNet3D—which is a limitation of this architecture and may have some unanticipated negative effects. Despite the fact that the spatiospatial models performed better, it is important to note that they do not adequately maintain the 3D nature of the data. The author proposed that this relationship was indirectly maintained through the network's channels, and that the network might pick up on the general representation in order to make the proper classifications. The trials have additionally demonstrated that, for the presented brain tumour classification challenge, spatiospatial models outperform completely 3D models. Nevertheless, these models need to be further tested for various tasks before a general agreement is reached regarding this finding.

In this research, the "specially-treated" spatial dimension of the spatiospatial models—which can alternatively be thought of as the pseudo-temporal dimension of the spatiotemporal models—was assumed to be the slice dimension in the axial direction. It has still to be proven whether it is also possible to utilize the benefits of such models by similarly orienting the data in sagittal or coronal orientation, according to the author premise. Additionally, it was found that the pre-trained models took first place in each of the three classes. Pre-training had a different impact on each of the three models, though. Pre-training enhanced the performance of the models for both spatiospatial models, although to varying degrees: 2.24 percent for ResNet (2+1)D and 8.57 percent for ResNet Mixed Convolution (based on macro F1 scores). Pre-training, however, adversely affected the 3D ResNet18 model (for two out of three classes), resulting in a 1.87 percent reduction in the macro F1 score. The pre-training led to an overall improvement of 2.88 percent across all models, as seen by the average macro F1 scores for all the models with and without pre-training (0.9169 with pre-training, 0.8912 without pre-training). The fact that the pre-trained networks were first trained on RGB videos is important. The performance of the models might be enhanced by pre-training them using MRI volumes or MR films (dynamic MRIs). Regarding comparisons to other research that have been published, it's interesting to note that those earlier papers simply categorized distinct grades of brain tumours (LGG and HGG), whereas this paper additionally added a class for healthy brains. As more classes make the task harder, the outcomes are therefore not entirely comparable. Even so, the results of the winning model are superior to all previously published techniques, with the exception of one that showed outcomes that were on par with ResNet Mixed Convolution (that paper reported 0.12% better accuracy, and 0.41%less specificity).However, this paper used four different contrasts and an additional dataset apart from BraTS, making them have a larger dataset for training.

VII. CONCLUSION

This research confirms how ResNet(2+1)D and ResNet Mixed Convolution, acting as spatiospatial models, might enhance the classification of brain tumour grades (i.e. low-grade and high-grade glioma), as well as classifying brain pictures with and without tumours, while lowering the computing costs. The performance of the spatiospatial models was compared to a pure 3D convolution model using a 3D ResNet18 model. To examine the efficacy of pre-training in this configuration, each of the three models was trained from scratch as well as using weights from pre-trained models that were trained on an action recognition dataset. Three fold cross-validation was used to produce the final results. Despite having fewer trainable parameters, it was found that the spatiospatial models outperformed a pure 3D convolutional ResNet18 model in terms of performance. Further observation reveals that pre-training enhanced the models' functionality. Overall, the pre-trained ResNet Mixed Convolution model was shown to be the best model in terms of F1-score, attaining 0.8949 and 0.9123 F1-scores for low-grade glioma and high-grade glioma, respectively, and a macro F1-score of 0.9345 and a mean test accuracy of 96.98 percent. This research demonstrates the potential of spatiospatial models to outperform a fully 3D convolutional model. This was demonstrated here, however, only for the specific job of classifying brain tumours using the dataset BraTS. These models should be compared for other tasks in the future to build a common consensus regarding the spatiospatial models. One limitation of this study is that it only used T1 contrast-enhanced images for classifying the tumours, which already resulted in good accuracy. Incorporating all four available types of images (T1, T1ce, T2, T2-Flair) or any combination of them might improve the performance of the model even further.

REFERENCES

[1]. Fritz, A. et al. International Classification of Diseases for Oncology Vol. 3 (World Health Organization, Geneva, 2001).

[2]. Goodenberger, M. L. et al. Genetics of adult glioma. Cancer Genet. 205, 613–621 (2012).

[3]. Claus, E. B. et al. Survival and low-grade glioma: The emergence of genetic information. Neurosurg. Focus 38, E6 (2015).

[4]. Raza, S. M. et al. Necrosis and glioblastoma: A friend or a foe? A review and a hypothesis. Neurosurgery 51, 2–13 (2002).

[5]. Engelhorn, T. et al. Cellular characterization of the peritumoral edema zone in malignant brain tumors. Cancer Sci. 100, 1856–1862 (2009).

[6]. Menze, B. H. et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. 34, 1993–2024 (2014).

[7]. Rajpurkar, P. et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. PLoS Med. 15, e1002686 (2018).

[8]. Mzoughi, H. et al. Deep multi-scale 3d convolutional neural network (cnn) for mri gliomas brain tumor classification. J. Digit. Imaging 33, 903–915 (2020).

[9]. Pei, L. et al. Brain tumor classification using 3d convolutional neural network. In International MICCAI Brain lesion Workshop, 335–342 (2019).

[10]. Ge, C. et al. Deep learning and multi-sensor fusion for glioma classification using multistream 2d convolutional networks. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 5894–5897 (2018).

[11]. Ouerghi, H. et al. Glioma classification via mr images radiomics analysis. Vis. Comput. 2021, 1–15 (2021).

[12]. He, K. et al. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778 (2016).

[13]. Tran, D. et al. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 6450–6459 (2018).

[14]. Torrey, L. et al. Transfer learning. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, 242–264 (2010).

[15]. Zhuang, F. et al. A comprehensive survey on transfer learning. Proc. IEEE 109, 43–76 (2020).

[16]. Sarasaen, C. et al. Fine-tuning deep learning model parameters for improved super-resolution of dynamic mri with prior-knowledge. Artif. Intell. Med. 121, 102196 (2021).

[17]. Pallud, J. et al. Quantitative morphological magnetic resonance imaging follow-up of low-grade glioma: A plea for systematic

measurement of growth rates. Neurosurgery 71, 729–740 (2012).

[18]. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8026–8037 (2019).

[19]. Torchvision models. https://pytorch.org/vision/stable/models.html #video-classification. Accessed on 15th December 2021.

[20]. Kinetics-400 dataset. https://deepmind.com/research/open-source/kinetics. Accessed on 15th December 2021.

[21]. Micikevicius, P. et al. Mixed precision training. arXiv preprint arXiv:1710.03740 (2017).

[22]. Nvidia apex. https://github.com/NVIDIA/apex. Accessed on 15th December 2021.

[23]. Pérez-García, F. et al. Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Programs Biomed. 2021, 106236 (2021).

[24]. Bakas, S. et al. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Sci. data 4, 1–13 (2017).

[25]. Bakas, S. et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018).

[26]. Ixi dataset. https://brain-development.org/ixi-dataset. Accessed on 15th December 2021.

[27]. Yang, Y. et al. Glioma grading on conventional mr images: A deep learning study with transfer learning. Front. Neurosci. 12, 804 (2018).

[28]. Smith, S. M. et al. Advances in functional and structural mr image analysis and implementation as fsl. Neuroimage 23, S208–S219 (2004).

[29]. Jenkinson, M. et al. Smith sm. FSL Neuroimage 62, 782–90 (2012).

[30]. Isensee, F. et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018).

[31]. Shahzadi, I. et al. Cnn-lstm: Cascaded framework for brain tumour classification. In 2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 633–637 (IEEE, 2018).

[32]. Zhuge, Y. et al. Automated glioma grading on conventional mri images using deep convolutional neural networks. Med. Phys. 47, 3044–3053 (2020).

[33]. Johnson, J. M. et al. Survey on deep learning with class imbalance. J. Big Data 6, 1–54 (2019).

ETHICS DECLRATION

Competing interests. The author declare no competing interests.