Brain Tumor Detection and Classification from Multi-Channel MRIs using Deep Learning and Transfer Learning

Subhashis Banerjee, *Student Member, IEEE*
Supervisor: Francesco Masulli, *Senior Member, IEEE* and Sushmita Mitra, *Fellow, IEEE*

Abstract—**Glioblastoma Multiforme constitutes 80% of malignant primary brain tumors in adults, and is usually classified as High Grade Glioma (HGG) and Low Grade Glioma (LGG). LGG tumors are less aggressive, with slower growth rate as compared to HGG, and are responsive to therapy. Tumor biopsy being challenging for brain tumor patients, noninvasive imaging techniques like Magnetic Resonance Imaging (MRI) have been extensively employed in diagnosing brain tumors. Therefore, development of automated systems for the detection and prediction of the grade of tumors based on MRI data become necessary. In this paper, we investigate Deep Convolutional Neural Networks (ConvNets) for classification of brain tumors using multisequence MR images. We propose three ConvNets, which are trained from scratch, on MRI patches, slices, and multi-planar volumetric slices. The suitability of transfer learning for the task is also studied by applying two existing ConvNets models (VGGNet and ResNet) trained on ImageNet dataset, through fine-tuning of the last few layers. Leave-one-patient-out (LOPO) testing scheme is used to evaluate the performance of the ConvNets. Results demonstrate that ConvNet achieves better accuracy in all cases where the model is trained on the multi-planar volumetric dataset. It obtains a testing accuracy of 97% without any additional effort towards extraction and selection of features, as required in conventional models. We also compare our results with state-of-the-art methods that require manual feature engineering for the task. It shows a maximum improvement of 12% on grading performance of ConvNets. We also study the properties of self-learned kernels/filters in different layers through visualization of the intermediate layers outputs.**

I. Introduction

Magnetic Resonance Imaging (MRI) has become the standard non-invasive technique for brain tumor diagnosis over the last few decades, due to its improved soft tissue contrast [1], [2]. Gliomas constitute 80% of all malignant brain tumors originating from the glial cells in the central nervous system. Based on the aggressiveness and infiltrative nature of the gliomas the World Health Organization (WHO) broadly classified them into two categories Low-grade gliomas (LGG), consisting of low-grade and intermediate-grade gliomas (WHO grades II and III), and high-grade gliomas (HGG) or glioblastoma multiforme (GBM) (WHO grade IV) [3]. Although most of the LGG tumors have slower growth rate compared to HGG and are responsive to treatment, there is a subgroup of LGG tumors which if not diagnosed earlier and left untreated could lead to GBM. In both cases a correct treatment planning (including surgery, radiotherapy, and chemotherapy separately or in combination) becomes necessary, considering that an early and proper detection of the tumor grade can lead to a good prognosis [4].

Histological grading, based on a stereotactic biopsy test, is the gold standard for detecting the grade of a brain tumor. The biopsy procedure requires the neurosurgeon to drill a small hole into the skull (exact location of the tumor in the brain guided by MRI), from which the tissue is collected using specialized equipments [5]. There are many risk factors involving the biopsy test, including bleeding from the tumor and brain due to the biopsy needle, which can cause a severe migraine, stroke, coma and even death. Other risks involve infection or seizures [6], [7]. But the main concern with the stereotactic biopsy is that it is not 100% accurate. When it misleads the histological grading of the tumor, there may result in a serious diagnostic error followed by a wrong clinical management of the disease [8].

In this context multi-sequence MRI plays a major role in the detection, diagnosis, and management of brain cancers in a non-invasive manner. Studies in the recent literature report that that, automatic computerized detection and diagnosis of the disease, based on medical image analysis, could be a good alternative. Decoding of tumor phenotype using noninvasive methods is a recent field of research, known as *Radiomics* [9]–[11], involving the extraction of a large number of quantitative imaging features that may not be visible to the human eye from medical images. An integral part of the procedure involves manual or automated delineation of the 2D region of interest (ROI) or 3D volume of interest (VOI) [12]–[15], to focus attention on the malignant growth. This is typically followed by the extraction of suitable sets of hand-crafted quantitative imaging features from the ROI or VOI, to be subsequently analyzed through machine learning towards decision-making. Feature selection enables the elimination of redundant and/or less important subset(s) of features, for improvement in speed and accuracy of performance. This is particularly relevant for high-dimensional radiomic features, extracted from image data.

Subhashis Banerjee (e-mail: mail.sb88@gmail.com) and Sushmita Mitra (e-mail: sushmita@isical.ac.in) are with Machine Intelligence Unit (MIU), Indian Statistical Institute, Kolkata, INDIA. Francesco Masulli (e-mail: francesco.masulli@unige.it) is with Department of Informatics, Bioengineering, Robotics and Systems Engineering (DIBRIS), University of Genova, Genoa, ITALY.

This work was done while the author visited Department of Informatics, Bioengineering, Robotics and Systems Engineering (DIBRIS), University of Genova, Genoa, ITALY.

Quantitative imaging features, extracted from MR images, have been investigated in literature for the assessment of brain tumors [11], [16]. In Ref. [17] authors proposed an adaptive neuro-fuzzy classifier based on linguistic hedges (ANFC-LH), for predicting the brain tumor grade using 56 3D quantitative MRI features extracted from the corresponding segmented tumor volumes. Quantitative imaging features, extracted from pre-operative gadolinium-enhanced T1-weighted MRI were investigated for diagnosis of meningioma (type of brain tumor) grades [18]. A study of MR imaging features was made [19] to determine those which can differentiate among grades of soft-tissue sarcoma (STS). The features investigated include signal intensity, heterogeneity, margin, descriptive statistics, and perilesional characteristics on images, obtained from each MR sequence. Brain tumor classification and grading study based on 2D quantitative imaging features like texture and shape, involving gray-level co-occurrence, run-length, and morphological features were also reported [20].

Although the techniques demonstrate good disease classification, their dependence on hand-crafted features requires extensive domain knowledge, involves human bias, and is problem specific. Manual designing of features typically requires greater insight into the exact characteristics of normal and abnormal tissues, and may fail to accurately capture some important representative features; thereby hampering classifier performance. The generalization capability of such classifiers may also suffer due to the discriminative nature of the methods, with the hand-crafted features being usually designed over fixed training sets. Subsequently manual or semi-automatic localization and segmentation of the ROI or VOI is also needed to extract the quantitative imaging features [12], [13].

Convolutional Neural Networks (ConvNets) offer state-of-the-art framework for image recognition or classification [21]–[23]. ConvNet architecture is designed to loosely mimic the fundamental working of the mammalian visual cortex system. It has been shown that the visual cortex has multiple layers of abstractions which look for specific patterns in the input vision. A ConvNet is built upon a similar idea of stacking multiple layers to allow it to learn multiple different abstractions of the input data. These networks automatically learn mid-level and high-level representations or abstractions from the input training data, in the form of convolution filters that are updated during the training process. They work directly on raw input (image) data, and learn the underlying representative features of the input which are hierarchically complex, thereby ruling out the need for specialized hand-crafted image features. Moreover, ConvNets require no prior domain knowledge and can automatically learn to perform any task just by working through the training data.

However, training a ConvNet from scratch is generally difficult because it essentially requires large training data, along with the significant expertise to select an appropriate model architecture for proper convergence. In medical applications data is typically scarce, and expert annotation is expensive. Training a deep CNN requires huge computational and memory resources, thereby making it extremely time-consuming. Repetitive adjustments in architecture and/or learning parameters, while avoiding overfitting, make deep learning from scratch a tedious, time-consuming, and exhaustive procedure. Transfer learning offers a promising alternative, in case of inadequate data, to fine tune a ConvNet already pre-trained on a large set of available labeled images from some other category [24]. This helps in speeding up convergence, while lowering computational complexity during training [25], [26].

In this paper we investigate the performance of ConvNets, with and without transfer learning, for non-invasive brain tumor detection and grade prediction from multi-sequence MRI. Tumors are typically heterogeneous, depending on cancer subtypes, and contain a mixture of structural and patch-level variability. Since performance and complexity of ConvNets depend on the input data representation, we experimented with three types of datasets – i) Patch-based, ii) Slice-based, and iii) Volume-based, prepared from the original MRI dataset. In each case, a ConvNet model is developed and trained from scratch. We have also tested two popular convolutional neural network architectures VGGNet [27], and ResNet [21], with parameters, pre-trained on ImageNet images using transfer learning (via fine-tuning) for the problem.

The main contributions of this research are listed as follows:

- Adaptation of deep learning to radiomics, for non-invasive prediction of brain tumors grades from multi-channel MR images.
- Prediction of the grade of brain tumor without manual segmentation of tumor volume, or manual extraction and selection of features.
- Development of novel ConvNet architectures viz. PatchNet, SliceNet, and VolumeNet for tumor grade prediction based on the MRI patches, MRI slices, and multi-planar volumetric MR images, respectively.
- New framework for applying existing pre-trained deep ConvNets models on multi-channel MRI data using transfer learning, which can be extended to other tasks based on MRI data.

The rest of the paper is organized as follows. Section II provides details about the data, its preparation in patch, slice and volumetric modes, and some preliminaries of ConvNets and transfer learning. In Section III we present the proposed ConvNet architectures. Section IV describes the experimental results, demonstrating the effectiveness (both qualitatively and quantitatively) with respect to existing related methods. Finally conclusions are provided in Section V.

II. Materials and Methods

In this section we provide a brief description of the data preparation at three levels of resolution, followed by an introduction to convolutional neural networks and transfer learning.

A. Brain tumor data

All the experiments were performed on the BraTS 2017 dataset [28], [29], which includes data from BraTS 2012, 2013, 2014 and 2015 challenges along with data from the Cancer Imaging Archive (TCIA). The dataset consisted of 210 HGG and 75 LGG glioma cases. Each patient MRI scan set has four MRI sequences or channels, encompassing native (T1) and post-contrast enhanced T1-weighted (T1C), T2-weighted (T2), and T2 Fluid-Attenuated Inversion Recovery (FLAIR) volumes having 155 2D slices of 240×240 resolution. The data is already aligned to the same anatomical template, skull-stripped, and interpolated to $1mm^3$ voxel resolution. In the ground truth images, each voxel is labeled with zeros and nonzeros, corresponding to the normal pixel and parts of tumor cells, respectively. Sample image of the two grades is shown in Fig. 1. It can be observed from the figure that it is very hard to discriminate between these two categories based on the phenotypes visible to the human eye. Hence, abstract features learned by the deep layers of a ConvNet might be helpful in differentiating the grades noninvasively. Besides, the use of large public domain datasets would allow for more clinical impact as compared to controlled and dedicated prospective image acquisitions.

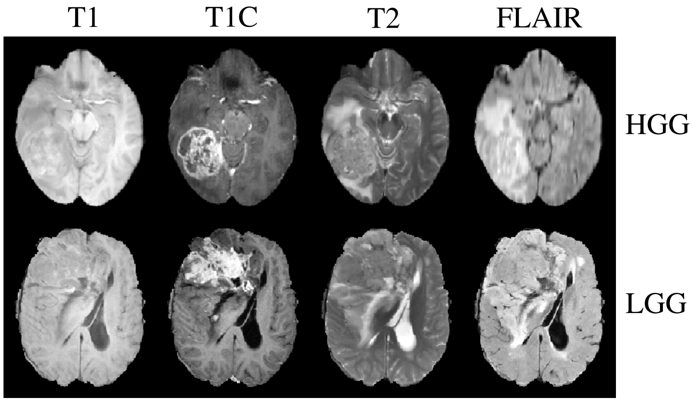


Fig. 1. MR images, of the two categories (HGG, LGG), from TCIA database [30]. Four sample sequences of a) HG, and b) LG patients.

B. Dataset preparation

Although the BraTS 2017 dataset consists MRI volumes, we cannot propose a 3D ConvNet model for the classification problem, mainly because the dataset has only 210 HGG and 75 LGG patients data, which is considered as inadequate to train a 3D ConvNet with a huge number of trainable parameters. Another problem with the dataset is its imbalanced class distribution i.e. about 35.72% of the data comes from the LGG class. Therefore formulate 2D ConvNet models based on the MRI patches (encompassing the tumor region) and slices, followed by a multi-planar slice-based ConvNet model that incorporates the volumetric information as well.

The tumor can be lying anywhere in the image and can be of any size (scale) or shape. Classifying the tumor grade from tumor patches is easier, than classifying the whole MRI slice, because here the ConvNet learns to localize only within the extent of the tumor in the image. Thereby the ConvNet needs to learn only the relevant details without getting distracted by irrelevant details. However, it may lack spatial and neighborhood details of the tumor, which may influence the grade prediction. Although classification based on the 2D slices and patches often achieves good accuracy, the incorporation of volumetric information from the dataset can enable the ConvNet to perform better.

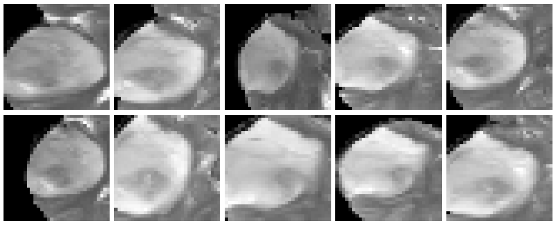


Fig. 2. Ten T2-MR patches extracted from contiguous slices from an LGG patient.

Along these lines, we propose schemes to prepare three different sets viz. (i) patch-based, (ii) slice-based, and (iii) multi-planar volumetric dataset, from the BraTs 2017 dataset.

1) Patch-based dataset: The slice with the largest tumor region is first identified. Keeping this slice in the middle a set of slices before and after that one considered for extracting 2D patches containing the tumor regions using bounding-box. Corresponding to each slice the bounding-box is marked based on the ground truth image, followed by the extraction of the image region enclosing within.

We use a set of 20 slices for extracting the patches. In case of MRI volumes from HGG (LGG) patients, four (ten) 2D patches [with a skip over 5 (2) slices] patches are extracted for each of the MR sequences. Therefore a total of $210 \times 4 = 840$ HGG and $75 \times 10 = 750$ LGG patches, with four channels each, constitute this dataset. Although the classes are still not perfectly balanced, this ratio is found to be good enough in the enhanced dataset.

In spite of significant dissimilarity visible between contiguous MRI slices at a global level, there may be little difference at the patch level. Therefore patches extracted from contiguous MRI slices look similar, particularly for LGG cases. This can lead to overfitting in the ConvNets. To overcome this problem we introduced a concept of static augmentation by randomly changing the perfect bounding-box coordinates by a small amount ($\in \{-5, 5\}$ pixels) before extracting the patch. This resulted in improved learning and convergence of the network. Fig. 2 depicts a set of 10 patches extracted from contiguous MR slices of an LGG patient.

2) Slice-based dataset: Complete 2D slices, with visible tumor region, are extracted from the MRI volume. The slice with the largest tumor region, with a set of 20 slices before and after it, are extracted from the MRI volume in a sequence similar to the patch-based approach.While for HGG patients 4 slices (with a skip over 5 slices) are used in the case of LGG patients 10 (with a skip of 2) slices are used.

3) Multi-planar volumetric dataset: Here 2D MRI slices are extracted along all three anatomical planes, viz. axial (X-Z axes), coronal (Y-X axes), and sagittal (Y-Z axes), in a manner similar to that described above.

C. Convolutional neural networks

Convolutional Neural Networks (ConvNets) can automatically learn low-level, mid-level and high-level abstractions from input training data in the form of convolution filter weights, that gets updated during the training process by back-propagation. The inputs percolating through the network are the responses of convoluting the images with various filters. These filters act as detectors of simple patterns like lines, edges, corners, from spatially contiguous regions in an image. When arranged in many layers, the filters can automatically detect prevalent patterns while blocking irrelevant regions. Parameter sharing and sparsity of connection are the two main concepts that make ConvNets easier to train with a small number of weights as compared to dense fully connected layers. This reduces the chance of overfitting, and enables learning translation invariant features. Some of the important concepts, in the context of ConvNets are next discussed.

1) Layers: The fundamental layers of a ConvNet consist of the input layer, convolution layer, activation layer, pooling layer and fully-connected layer. Some additional layers include the dropout layer, and batch-normalization layer.

- *Input layer:* This serves as the entry point of the ConvNet, taking the raw pixel value of the input image. Here input is a 4-channel brain MRI patch/slice denoted by $I \in \mathbb{R}^{4 \times w \times h}$, where w and h represent the resolution of the image.
- *Convolution layer:* It is the core building block of a ConvNet. Each convolution layer is composed of a filter bank (set of convolutional filters/kernels of same width and height). The number and size of filters in a bank are specified by the user for each convolutional layer. The depth of the filters in each filter bank is determined by the depth (channel) of its input volume. A convolutional layer takes an image or feature maps as input and performs the convolution operation between the input and each of these filters by sliding (also called stride) the filter over the image to generate a set of (same as the number of filters) activation maps or the feature map. The output feature map dimension, from a convolution layer, is calculated as
$$w_{out}/h_{out} = \frac{(w_{in}/h_{in} - F + 2P)}{Stride} + 1, \tag{1}$$
where w_{in} and h_{in} are the width and height of the input image, w_{out} and h_{out} are the width and height of the effective output. Here P denotes the input padding which if set to zero known as "valid" convolution involving nil zero-padding. The displacement $Stride = 1$, with F being the receptive field (kernel size) of the neurons in a particular layer.
- *Activation layer:* Output responses of the convolution and fully connected layers pass through some nonlinear activation function such as a rectified linear unit (ReLU) [31] for transforming the data. ReLU, defined as $f(a) = max(0, a)$, is a popular activation function for deep neural networks due to its computational efficiency and reduced likelihood of vanishing gradient.
- *Pooling layer:* Pooling layer follows each convolution layer to typically reduce computational complexity by downsampling of the convoluted response maps. It combines spatially close, possibly redundant features in the feature maps; thereby, making the representation more compact and invariant to small changes in an image like the insignificant details. Max pooling enables selection of the maximum feature response in local neighborhoods, while discarding its exact location, and thereby enhances translation invariance.
- *Fully-connected layer:* The features learned through a series of convolutional and pooling layers are eventually fed to a fully-connected layer, typically a Multilayer Perceptron. The term "fully-connected" implies that every neuron in a layer is connected to every neuron of the following layer. The purpose of the fully-connected layer is to use these features for categorizing the input image into different classes, based on the training dataset.

Additional layers like Batch-Normalization [32] reduces initial covariate shift. Dropout [33] is used as regularizer to randomly disable nodes of the network during training; thereby forcing all nodes in the fully connected layers to learn a better representation of the data, while preventing them from co-adapting to each other.

2) Loss: The cost function for all the proposed and fine-tuned ConvNets is chosen as binary cross-entropy (for a two-class problem) as
$$L_C = -\frac{1}{n}\sum_{i=1}^{n}\left\{y_i \log(f_i) + (1 - y_i)\log(1 - f_i)\right\}, \tag{2}$$
where n is the number of samples, y_i is the true label of a sample and f_i is its predicted label.

D. Transfer Learning

Typically the early layers of a ConvNet learn low-level image features, which are applicable to most vision tasks. The later layers, on the other hand, learn high-level features which are more application-specific. Therefore, shallow fine-tuning of the last few layers is usually sufficient for transfer learning. A common practice is to replace the last fully-connected layer of the pre-trained ConvNet with a new fully-connected layer, having as many neurons as the number of classes in the new target application. The rest of the weights, in the remaining layers, of the pre-trained network are retained. This corresponds to training a linear classifier with the features generated in the preceding layer. However, when the distance between the source and target applications is significant than one may need to induce deeper fine-tuning. This is equivalent to training a shallow neural network with one or more hidden layers. An effective strategy [34] is to initiate fine-tuning from the last layer, and then incrementally include deeper layers in the tuning process until the desired performance is achieved.

III. Three level ConvNets for Brain tumor Grading

A. Architectures

We propose three ConvNet architectures named as PatchNet, SliceNet, and VolumeNet, which are trained from scratch on the three datasets prepared as detailed in Section II-B. This is followed by transfer learning and fine-tuning of these networks. The ConvNet architectures are illustrated in Fig. 3. As the names suggested, PatchNet is trained on the patch-based dataset and provides the probability of a patch belong to HGG or LGG. SliceNet gets trained on the slice based-dataset and predicts the probability of a slice being from HGG or LGG. Finally, VolumeNet is trained on the multi-planar volumetric dataset and predicts the grade of the tumor from its 3D representation using the multi-planar 3D MRI data.

As reported in the literature, smaller size convolutional filters produce better regularization due to the smaller number of trainable weights; thereby allowing construction of deeper networks without losing too much information in the layers. We use filters of size (3×3) for our ConvNet architectures. A greater number of filters, involving deeper convolution layers, allows for more feature maps to be generated. Thus compensates for the decrease in the size of each feature map caused by "valid" convolution and pooling layers. Due to the complexity of the problem and bigger size of the input image, the SliceNet and VolumeNet architectures are deeper as compared to the PatchNet.

B. *Fine-tuning*

Pre-trained VGGNet (16 layers), and ResNet (50 layers) architectures trained on the ImageNet dataset are employed for transfer learning. Even though ResNet is much deeper than VGGNet, the model size of ResNet is actually substantially smaller due to the usage of global average pooling rather than fully-fully-connected layers. Transfering from the non-medical to the medical image domain was achieved through fine-tuning of the last convolutional block of each model alongside the fully-connected layers (top-level classifier) of each model. Fine-tuning of a trained network is achieved by re-training it on the new dataset with very small weight updates. In our case we did it in the following four steps:

- Instantiate the convolutional base of the model and load its weights.
- Replace the last fully-connected layer of the pre-trained ConvNet with a new fully-connected layer, having single neuron with sigmoid activation.
- Freeze the layers of the model up to the last convolutional block.
- Finally retrain the last convolution block and the fully-connected layers with a very slow learning rate with the SGD optimizer.

Since the models were trained on the RGB images, and accept single input with three channels, we train and tested them on the slice-based dataset with the three MR sequences ($T1C$, $T2$, $FLAIR$). We fine-tuned the models using $T1$ instead $T1C$ along with the other two sequences and found that $T1C$ gives much more accuracy than $T1$. Although running any of the two models from the scratch is very expensive, especially if you're working on CPU, here we just train the last few layers which could be easily done on a CPU. Results for both, ConvNets trained from scratch and using transfer learning are presented in the next section.

IV. Experimental Results

A. *Implementation*

The ConvNets were developed using TensorFlow, with Keras in Python. The experiments were performed on a desktop machine with Intel i7 CPU (clock speeds $3.40GHz$), having 4 cores, 32GB RAM, and NVIDIA GeForce GTX 1080 GPU with 8GB VRAM. The operating system was Ubuntu 16.04.

B. *Quantitative Evaluation*

Due to the small size (only 285 patients), and uneven class distributions (210 HGG and 75 LGG patients), we propose leave-one-patient-out (LOPO) test scheme for quantitative evaluation. So in each iteration, one patient is used for testing and remaining patients are used for training the ConvNets, this iterates for each patient. Although LOPO test scheme is computationally expensive, using this we can have more training data which is required for ConvNets training. LOPO testing is robust and most applicable to our application, where we get test result for each individual patient. So, if classifier misclassifies a patient then we can further investigate it separately.

The three dataset preparation schemes discussed in Section II-B are used to create the three separate training and testing data sets. Proposed ConvNetmodels – PatchNet, SliceNet, VolumeNet are trained on the corresponding datasets using the Stochastic Gradient Descent (SGD) optimization algorithm with learning rate $= 0.001$, and momentum $= 0.9$ using mini-batches of size 32 samples generated from the corresponding training dataset. During the training small part of the training set (20%) used as the validation set for validating the ConvNet model after each epoch for parameter selection and to inspect overfitting.

Since deep ConvNets entail a large number of free trainable parameters, the effective number of training samples were artificially enhanced using real-time data augmentation through some linear transformation such as random rotation ($0° - 10°$), horizontal and vertical shifts, horizontal and vertical flips. This type of augmentation works on the CPU parallel to the training process running on GPU, thereby saving computing time and improving resource usage when the CPU is idle during training. After each epoch, the model was validated on the corresponding validation dataset. Training and validation performance of the three ConvNets measured using the following two metrics.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$F_1 Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

$Accuracy$ is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. $F_1 Score$ is the weighted average of $Precision$ and $Recall$, which are defined as $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$. TP, TN, FP, and FN indicate numbers of true positives, true negatives, false positive and false negative detections. When we have an unbalanced dataset $F_1 Score$ favored over accuracy because it takes both false positives and false negatives into account.

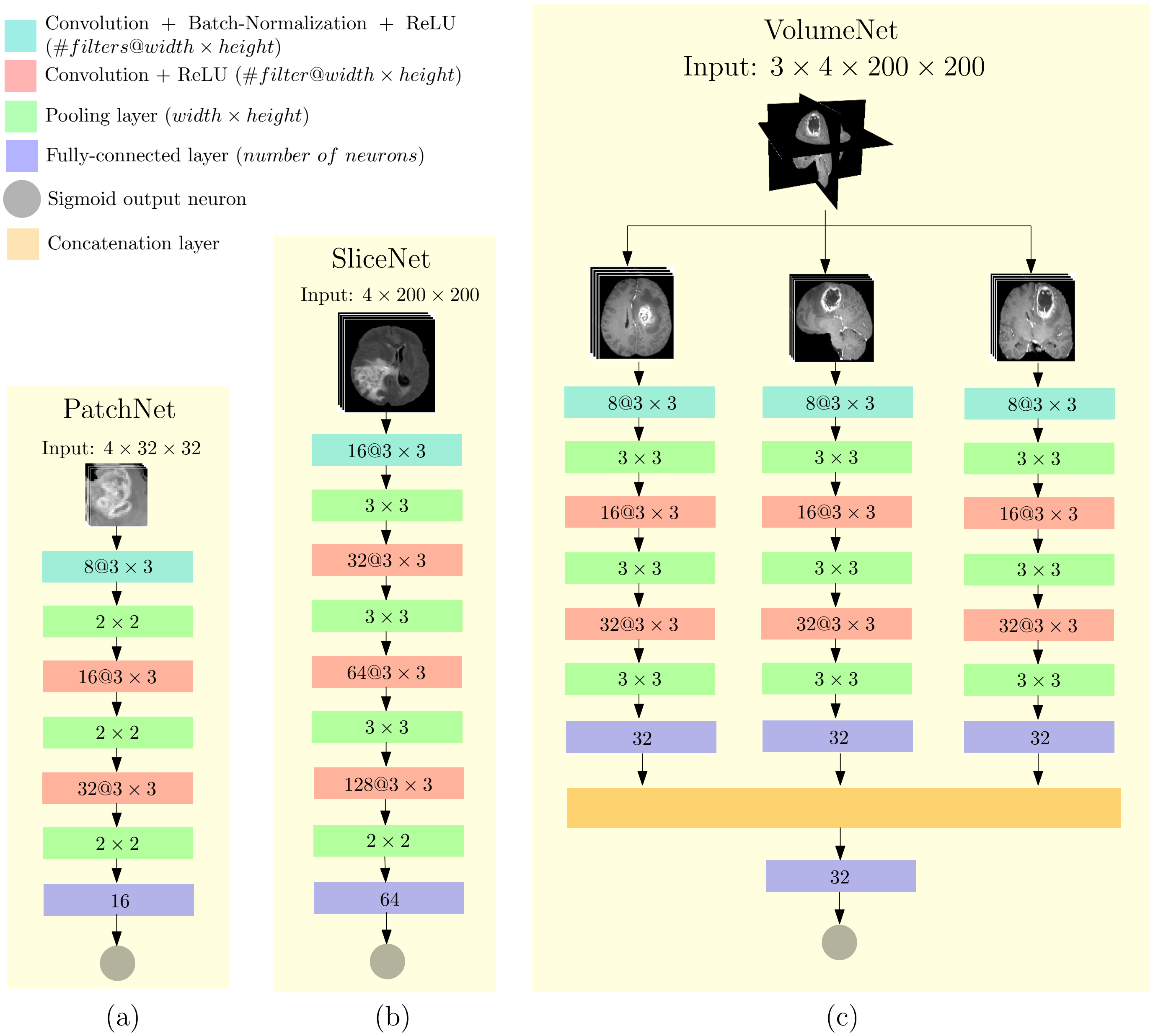


Fig. 3. Three level ConvNet architectures, (a) PatchNet, (b) SliceNet, and (c) VolumeNet.

Training and validation accuracy and loss, F1-score on the validation dataset for a sample iteration of the three proposed ConvNets (PatchNet, SliceNet, and VolumeNet), trained from scratch and the two pre-trained ConvNets (VGGNet, and ResNet) fine-tuned on the Brats2017 dataset are given in Fig. 4. The plots demonstrate that VolumeNet gives the highest classification performance during training, it reaches the maximum accuracy on the training set (100%) and the validation set (98%) just within 20 epochs. The performance of PatchNet and SliceNet are quite similar on the validation set (PatchNet - 90%, SliceNet - 92%) although on the training set SliceNet achieves better accuracy (95%), which is due to some overfitting after 50 epochs. The performance of two the pre-trained models (VGGNet and ResNet) show similar results, and both achieve around 85% accuracy on the validation set. All the networks plateau after the 50th epoch.

After the model was trained, it was evaluated on the hold-out test set using majority voting scheme. So, each individual patch or slice is classified as HGG or LGG from the test dataset which is from a single test patient. Then the class with maximum slices or patches classified, selected as the grade of the tumor. In case of an equal vote in each class, the patient is marked as ambiguous. LOPO testing scores are shown in the Table. I. VolumeNet achieves the best LOPO test accuracy (97.19%), with zero ambiguous compared to other four networks. SliceNet also achieves good LOPO test accuracy (90.18%). The pre-trained models show similar LOPO test accuracy as PatchNet, which is very interesting because with a little fine-tuning we can achieve test accuracy similar to a ConvNet trained from scratch on the specific

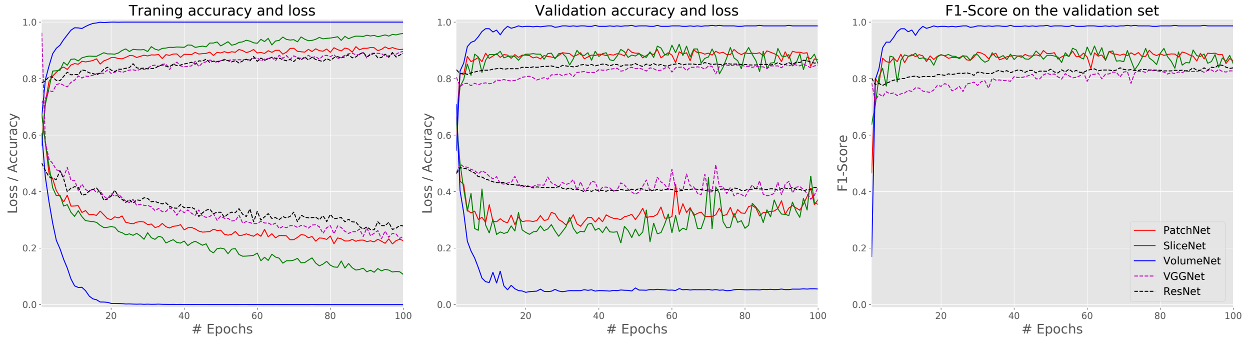


Fig. 4. Training and validation accuracy and loss, F1-Score on the validation dataset for the five ConvNets.

TABLE I
LOPO TEST PERFORMANCE OF THE FIVE CONVNETS

ConvNets	Classified	Misclassied	Ambiguous	Accuracy
PatchNet	242	39	4	84.91 %
SliceNet	257	26	2	90.18 %
VolumeNet	**277**	**8**	**0**	**97.19 %**
VGGNet	239	40	6	83.86 %
ResNet	242	42	1	84.91 %

TABLE II
TRAINING TIME

ConvNet	Time $(Mean \pm SD)$	Training type
PatchNet	10.75 ± 0.05 min	from scratch
SliceNet	65.95 ± 0.02 min	from scratch
VolumeNet	132.48 ± 0.05 min	from scratch
VGGNet	8.56 ± 0.03 min	fine-tuning
ResNet	12.14 ± 0.03 min	fine-tuning

dataset. So, if we fine-tune some more intermediate layers then there is a chance of getting very high scores with a little amount of training. The total time required for training each network for 100 epochs are mentioned in Table. II, mean over several runs.

In Table. III we compared the proposed ConvNets with other existing shallow learning models used for the same application from literature, which requires additional effort to extract and select features from the manually segmented ROI / VOI, in terms of classification accuracy. Ref. [17] reports the accuracy achieved by seven standard classifiers, viz. i) Adaptive Neuro-Fuzzy Classifier (ANFC), ii) Naive Bayes (NB), iii) Logistic Regression (LR), iv) Multilayer Perceptron (MLP), v) Support Vector Machine (SVM), vi) Classification and Regression Tree (CART), and vii) k-nearest neighbors (k-NN). The accuracy reported in Ref. [17] are on the BraTS 2015 dataset (a subset of BraTS 2017 dataset) which consists 200 HGG and 54 LGG cases. 56 three-dimensional quantitative MRI features extracted manually from each patient MRI and used for the classification. Where in our case, we leverage the learning capability of deep convolutional neural networks for automatically learning the features from the data.

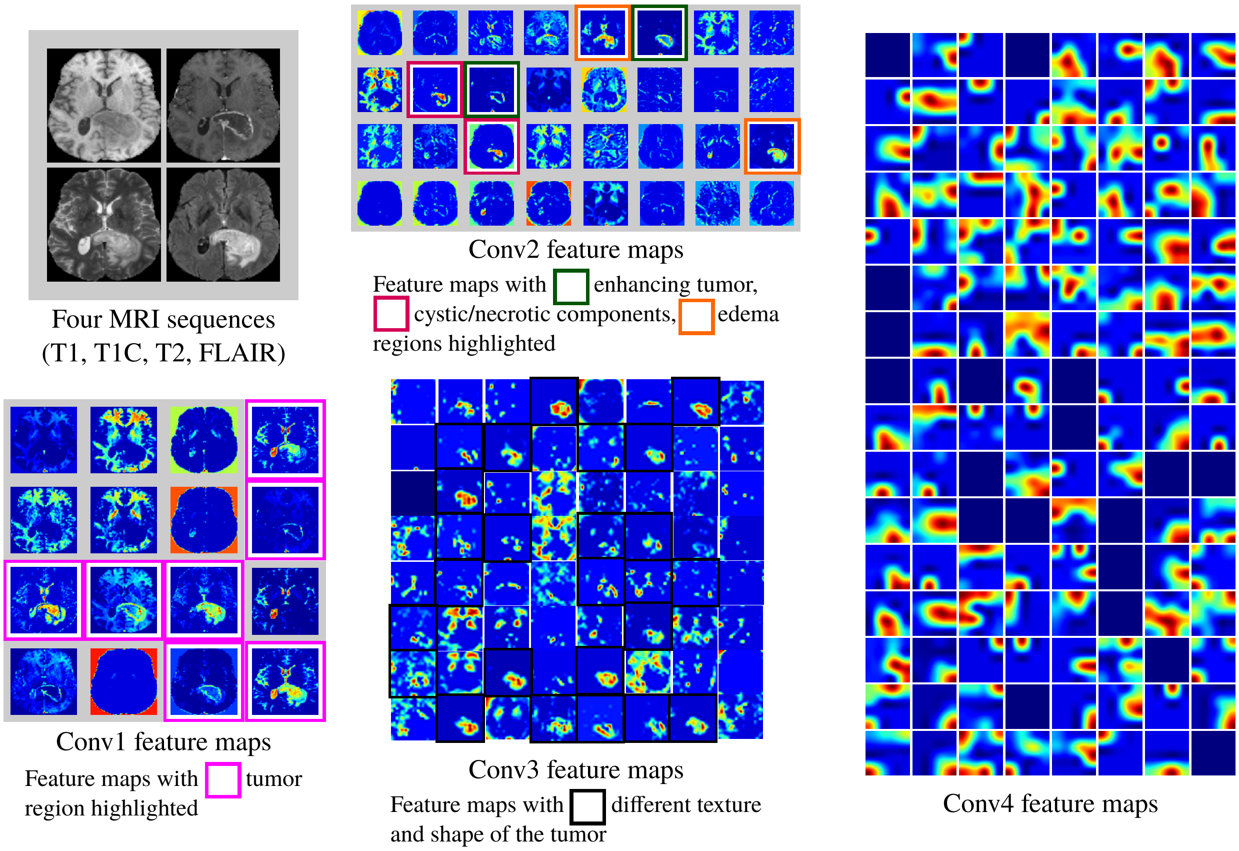
TABLE III
THE CLASSIFICATION ACCURACY OF DIFFERENT DEEP AND SHALLOW LEARNING MODELS.

Classifier	Accuracy (%)	Details
PatchNet	84.91	Trained and tested on 2D MRI patches of size 32 × 32.
SliceNet	90.18	Trained and tested on MRI slices of size 200 × 200.
VolumeNet	**97.19**	Trained and tested on multi planar MRI slices of size 200 × 200.
VGGNet	83.86	Trained on ImaeNet dataset, fine-tuned and tested on MRI slices of size 200 × 200.
ResNet	84.91	Trained on ImaeNet dataset, and fine tuned and tested on MRI slices of size 200 × 200.
ANFC-LH	85.83	Trained on manually extracted quantitative MRI features, based on 10 fuzzy rules.
NB	69.48	Trained on manually extracted quantitative MRI features.
LR	72.07	Trained on manually extracted quantitative MRI features based on multinomial logistic regression model with a ridge estimator.
MLP	78.57	Trained on manually extracted quantitative MRI features using single hidden layer with 23 neurons, learning rate = 0.1, momentum = 0.8.
SVM	64.94	Trained on manually extracted quantitative MRI features, LibSVM with RBF kernel, cost = 1,gamma = 0.
CART	70.78	Trained on manually extracted quantitative MRI features using minimal cost-complexity pruning.
k-NN	73.81	Trained on manually extracted quantitative MRI features, accuracy averaged over scores for $k = 3, 5, 7$.

C. Qualitative Evaluation

We further investigate the ConvNets through visual analysis of the intermediate layers outputs. The performance of a ConvNet fully depends on the convolution kernels which are the feature extractors, learned from the unsupervised learning process. By visualizing the outputs of any convolution layer, description of the kernels learned can be determined. Fig. 5, illustrates the intermediate convolution layer outputs (after the ReLU activation) of the proposed SliceNet architecture on a sample MRI slices from an HGG patient.

The visualization of the first convolution layer activations or



Four MRI sequences
(T1, T1C, T2, FLAIR)

Conv2 feature maps

Feature maps with 🟩 enhancing tumor, 🟥 cystic/necrotic components, 🟧 edema regions highlighted

Conv1 feature maps

Feature maps with 🟪 tumor region highlighted

Conv3 feature maps

Feature maps with ⬛ different texture and shape of the tumor

Conv4 feature maps

Fig. 5. Intermediate layers outputs/feature maps generated by SliceNet, on an HGG MRI slice.

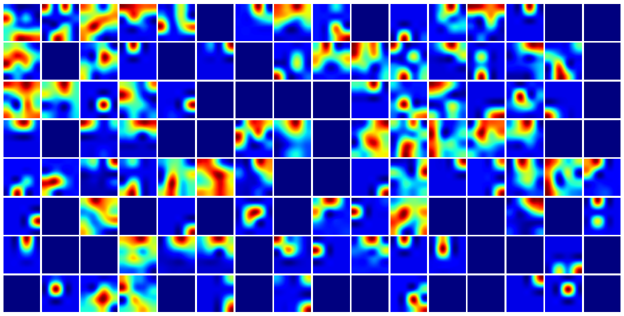


Fig. 6. Feature maps generated from the last convolution layer by SliceNet, on an LGG MRI slice.

feature maps indicates that the ConvNet has learned a variety of filters that can detect edges and distinguish different brain tissues such as white matter (WM), gray matter (GM), cerebrospinal fluid (CSF), skull and background. Most importantly some filters isolate the region of interest or the tumor on the basis of which we want to classify the whole MRI slice. Most of the feature maps generated by the second convolution layer highlight mainly the tumor region and its subregions like enhancing tumor structures, surrounding cystic/necrotic components and the edema region of the tumor. So, the filters in the second convolution layer have learned to extract deeper features from the tumor by concentrating particularly to the ROI or the tumor. The texture and shape of the tumor get enhanced in the feature maps generated from the third convolution layer, like small-sized, distributed and enhanced tumor cells which is one of the most important tumor grading criteria called “CE-Heterogeneity”, irregular, nodule or flower shape are formed. Such that, next layer will be able to extract more detailed information about more discriminating features by combining these to produce a clear distinction in the images of different types of tumors. By visualizing the final feature maps generated from the last convolution layer a clear discrimination between two grades can be noticed in Figs. 5-6.

V. Conclusion

In this paper, we have presented three novel ConvNet architectures for grading brain tumors non-invasively, into HGG and LGG, from the MR images of tumors and explore transfer learning for the same task, by fine-tuning two existing ConvNet models. An improvement about 12% in terms of classification accuracy on the test dataset was observed from deep ConvNets compared to shallow learning models. Visualizations of the intermediate layer outputs/feature maps show that kernels/filters in the convolution layers automatically learned to detect different tumor features that are closely resembled different tumor grading criteria. We also noticed that existing ConvNets trained on natural images performed adequately by only fine-tuning their final convolution layer on the MRI dataset. In our experiments, we proposed a scheme for incorporating volumetric tumor information using multi-planar MRI slices, that achieved the best testing accuracy 97.19%. So, we conclude that deep ConvNets could be a feasible alternative to surgical biopsy for brain tumors.

Acknowledgment

This research is funded by the IEEE Computational Intelligence Society Graduate Student Research Grant 2017.

The author would like to thank Professor Francesco Masulli, Professor Sushmita Mitra and Professor Stefano Rovetta for their guidance in the completion of this research.

References

[1] L. M. DeAngelis, "Brain tumors," *New England Journal of Medicine*, vol. 344, no. 2, pp. 114–123, 2001.

[2] S. Cha, "Update on brain tumor imaging: From anatomy to physiology," *American Journal of Neuroradiology*, vol. 27, no. 3, pp. 475–487, 2006.

[3] D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvet, B. W. Scheithauer, and P. Kleihues, "The 2007 WHO classification of tumours of the Central Nervous System," *Acta neuropathologica*, vol. 114, no. 2, pp. 97–109, 2007.

[4] M. J. van den Bent, A. A. Brandes, M. J. Taphoorn, J. M. Kros, M. C. Kouwenhoven, J.-Y. Delattre, H. J. Bernsen, M. Frenay, C. C. Tijssen, W. Grisold *et al.*, "Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: Long-term follow-up of EORTC brain tumor group study 26951," *Journal of clinical oncology*, vol. 31, no. 3, pp. 344–350, 2012.

[5] J. F. Hahn, W. J. Levy, and M. J. Weinstein, "Needle biopsy of intracranial lesions guided by computerized tomography," *Neurosurgery*, vol. 5, no. 1, pp. 11–15, 1979.

[6] M. Field, T. F. Witham, J. C. Flickinger, D. Kondziolka, and L. D. Lunsford, "Comprehensive assessment of hemorrhage risks and outcomes after stereotactic brain biopsy," *Journal of neurosurgery*, vol. 94, no. 4, pp. 545–551, 2001.

[7] M. J. McGirt, G. F. Woodworth, A. L. Coon, J. M. Frazier, E. Amundson, I. Garonzik, A. Olivi, and J. D. Weingart, "Independent predictors of morbidity after image-guided stereotactic brain biopsy: a risk assessment of 270 cases," *Journal of neurosurgery*, vol. 102, no. 5, pp. 897–901, 2005.

[8] P. T. Chandrasoma, M. M. Smith, and M. L. J. Apuzzo, "Stereotactic biopsy in the diagnosis of brain masses: Comparison of results of biopsy and resected surgical specimen," *Neurosurgery*, vol. 24, no. 2, pp. 160–165, 1989.

[9] S. Mitra and B. Uma Shankar, "Medical image analysis for cancer management in natural computing framework," *Information Sciences*, vol. 306, pp. 111–131, 2015.

[10] ——, "Integrating radio imaging with gene expressions toward a personalized management of cancer," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 5, pp. 664–677, 2014.

[11] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, pp. 563–577, 2015.

[12] S. Banerjee, S. Mitra, and B. Uma Shankar, "Single seed delineation of brain tumor using multi-thresholding," *Information Sciences*, vol. 330, pp. 88–103, 2016.

[13] S. Banerjee, S. Mitra, B. Uma Shankar, and Y. Hayashi, "A novel GBM saliency detection model using multi-channel MRI," *PLOS ONE*, vol. 11, no. 1, p. e0146388, 2016.

[14] S. Banerjee, S. Mitra, and B. Uma Shankar, "Automated 3D segmentation of brain tumor using visual saliency," *Information Sciences*, vol. 424, pp. 337–353, 2018.

[15] S. Mitra, S. Banerjee, and Y. Hayashi, "Volumetric brain tumour detection from mri using visual saliency," *PLOS ONE*, vol. 12, pp. 1–14, 2017.

[16] M. Zhou, J. Scott, B. Chaudhury, L. Hall, D. Goldgof, K. Yeom, M. Iv, Y. Ou, J. Kalpathy-Cramer, S. Napel, R. Gillies, O. Gevaert, and R. Gatenby, "Radiomics in brain tumor: Image assessment, quantitative feature descriptors, and machine-learning approaches," *American Journal of Neuroradiology*, 2017.

[17] S. Banerjee, S. Mitra, and B. U. Shankar, "Synergetic neuro-fuzzy feature selection and classification of brain tumors," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017, pp. 1–6.

[18] T. Coroller, W. Bi, M. Abedalthagafi, A. Aizer, W. Wu, N. Greenwald, R. Beroukhim, O. Al-Mefty, S. Santagata, I. Dunn *et al.*, "Early grade classification in meningioma patients combining radiomics and semantics data," *Medical Physics*, vol. 43, pp. 3348–3349, 2016.

[19] F. Zhao, S. Ahlawat, S. J. Farahani, K. L. Weber, E. A. Montgomery, J. A. Carrino, and L. M. Fayad, "Can MR imaging be used to predict tumor grade in soft-tissue sarcoma?" *Radiology*, vol. 272, pp. 192–201, 2014.

[20] E. I. Zacharaki, S. Wang, S. Chawla, D. Soo Y., R. Wolf, E. R. Melhem, and C. Davatzikos, "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme," *Magnetic Resonance in Medicine*, vol. 62, pp. 1609–1618, 2009.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[24] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1717–1724.

[25] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[26] H. T. H. Phan, A. Kumar, J. Kim, and D. Feng, "Transfer learning of a convolutional neural network for hep-2 cell image classification," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 1208–1211.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1993–2024, 2015.

[29] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, p. sdata2017117, 2017.

[30] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, pp. 1045–1057, 2013.

[31] 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, pp. 315–323.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[33] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[34] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1299–1312, 2016.