frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 17 May 2021
doi: 10.3389/frai.2021.635766

A Transfer Learning–Based Active Learning Framework for Brain Tumor Classification

Ruqian Hao[1,2,3], *Khashayar Namdar*[4,3], *Lin Liu*[1] *and Farzad Khalvati*[2,4,3,5]*

[1]School of Optoelectronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, [2]Institute of Medical Science, University of Toronto, Toronto, ON, Canada, [3]Department of Diagnostic Imaging, and Neurosciences and Mental Health, The Hospital for Sick Children (SickKids), Toronto, ON, Canada, [4]Department of Medical Imaging, University of Toronto, Toronto, ON, Canada, [5]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada

Brain tumor is one of the leading causes of cancer-related death globally among children and adults. Precise classification of brain tumor grade (low-grade and high-grade glioma) at an early stage plays a key role in successful prognosis and treatment planning. With recent advances in deep learning, artificial intelligence–enabled brain tumor grading systems can assist radiologists in the interpretation of medical images within seconds. The performance of deep learning techniques is, however, highly depended on the size of the annotated dataset. It is extremely challenging to label a large quantity of medical images, given the complexity and volume of medical data. In this work, we propose a novel transfer learning–based active learning framework to reduce the annotation cost while maintaining stability and robustness of the model performance for brain tumor classification. In this retrospective research, we employed a 2D slice–based approach to train and fine-tune our model on the magnetic resonance imaging (MRI) training dataset of 203 patients and a validation dataset of 66 patients which was used as the baseline. With our proposed method, the model achieved area under receiver operating characteristic (ROC) curve (AUC) of 82.89% on a separate test dataset of 66 patients, which was 2.92% higher than the baseline AUC while saving at least 40% of labeling cost. In order to further examine the robustness of our method, we created a balanced dataset, which underwent the same procedure. The model achieved AUC of 82% compared with AUC of 78.48% for the baseline, which reassures the robustness and stability of our proposed transfer learning augmented with active learning framework while significantly reducing the size of training data.

Keywords: brain tumor, transfer learning, active learning, MRI, classification

OPEN ACCESS

Edited by:
Tuan D. Pham,
Prince Mohammad bin Fahd University, Saudi Arabia

Reviewed by:
Akram Mohammed,
University of Tennessee Health Science Center, United States
Chirag Kamal Ahuja,
Post Graduate Institute of Medical Education and Research, India

***Correspondence:**
Farzad Khalvati
farzad.khalvati@utoronto.ca

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *30 November 2020*
Accepted: *26 April 2021*
Published: *17 May 2021*

Citation:
Hao R, Namdar K, Liu L and Khalvati F (2021) A Transfer Learning–Based Active Learning Framework for Brain Tumor Classification. Front. Artif. Intell. 4:635766. doi: 10.3389/frai.2021.635766

INTRODUCTION

Brain tumor is one of the leading causes of cancer-related death globally among children and adults (Siegel et al., 2019). According to the World Health Organization (WHO) classification 2016 (Louis et al., 2016), brain tumors are divided into different grades (grades I, II, III, or IV) based on histology and molecular characteristics. The higher the grade of the tumor is, the more malignant it becomes. Patients with low-grade glioma (LGG, grade I/II) usually have better survival than those diagnosed with high-grade glioma

(HGG, grade III/IV), which is incurable and universally fatal. LGG has high possibility of eventually progressing to HGG if it is not diagnosed and the treatment is delayed (Claus et al., 2016).

Precise classification of brain tumor grade at the early stage plays a key role in successful prognosis (Delattre et al., 2014). Magnetic resonance imaging (MRI) is the favored imaging technique in glioma diagnostics due to good contrast enhancement and noninvasive features (Essig et al., 2012). The conventional method for tumor detection is followed by radiologists who observe and diagnose tumors, which is extremely laborious and time-consuming. Recent advances in artificial intelligence (AI) and deep learning have made great strides in computer-aided medical diagnosis (CAMD), which can assist doctors in the interpretation of medical images within seconds (Hosny et al., 2018).

The performance of deep learning technique is highly dependent on the quality and size of the dataset. Deep learning techniques require a large number of images with high-quality annotations. However, labeling large quantities of medical images is quite challenging as annotation can be expensive in terms of both time and expertise (Razzak et al., 2018). Insufficient amount of imaging data and scarcity of human expert annotations for images are the two major barriers in success of deep learning for medical imaging (Razzak et al., 2018).

To address and resolve the abovementioned challenges, numerous efforts have been made. For instance, transfer learning is a promising strategy in case of limited domain training samples. It fine-tunes a network which is already pretrained on a large labeled dataset, typically from another domain. By transferring learned knowledge to the target dataset, the speed of network convergence becomes faster while maintaining low computational complexity level at the training stage (Tajbakhsh et al., 2016).

Active learning algorithms have also been investigated to train a competitive classifier with minimal annotation cost. The underlying idea behind active learning is that different training examples have different effects on the performance of the current model. Instead of labeling the complete dataset, an active learning method selects a subset of informative samples to annotate and then train the classification model without compromising its performance. There are two important metrics to describe the informativeness of an unlabeled sample: uncertainty, which is the inverse of the confidence of predicted results by the model; and representativeness, which measures the degree of similarity in distribution and structure between selected samples and target dataset (Du et al., 2017). Based on different query schemes of informative unlabeled samples, conventional active learning algorithms can be listed as follows: uncertainty sampling, query by committee, expected model change, expected error reduction, variance reduction, and density-weighted methods (Settles 2011).

In this work, we propose an active learning method which integrates traditional uncertainty sampling technique and query-by-committee method, and transfer learning to reduce the amount of required training samples while maintaining stability and robustness of convolutional neural network (CNN) performance for brain tumor classification.

MATERIALS AND METHODS

Related Work

Brain Tumor Classification Using Deep Learning

Pereira et al. (2018) proposed a novel CNN with deeper architectures and small kernels for automatic LGG and HGG brain tumor grading prediction on both whole brain and only tumor region MRI images, and the accuracies were 89.5% and 92.98%, respectively. The datasets they used were BRATS 2013 and BRATS 2015. Suganthe et al. (2020) employed recurrent neural network (RNN) architecture for detection of tumors on a 600 MRI brain image dataset and achieved an accuracy of 90%. On a brain tumor dataset consisting of 3,064 MRI images from 233 patients, there has been multiple experiments (Afshar et al., 2018; Das et al., 2019; Badža and Barjaktarović 2020). Each patient in the dataset has one of the three types of brain tumor (glioma, meningioma, and pituitary). Badža and Barjaktarović (2020) presented a new CNN architecture for the three types of brain tumor classification, and the best accuracy was 96.56%. Das et al. (2019) also explored a CNN model for the classification of the three types of brain tumor MRI images, and an accuracy of 94.39% was achieved. Afshar et al. (2018) proposed a modified CapsNet architecture (Ballal and Joseph 2004) combined with tumor boundaries information for brain tumor classification and achieved 90.89% accuracy.

Transfer Learning and Active Learning for Medical Imaging

Yang et al. (2018) compared the classification performance of fine-tuned pretrained CNNs and CNNs trained from scratch on a private glioma MRI dataset containing 113 LGG and HGG patients. The experiments showed that transfer learning and fine-tuning improved performance for classifying HGG and LGG. They achieved their best test accuracy of 90%, using GoogLeNet. Banerjee et al. (2019) proposed three CNN models (PatchNet, SliceNet, and VolumeNet), trained from scratch, and compared with the two pretrained ConvNets (VGGNet (Simonyan and Zisserman 2015) and ResNet (Li et al., 2019)) fine-tuned on the BRATS 2017 dataset for HGG and LGG classification problem. Results demonstrate that the proposed VolumeNet achieved best testing accuracy of 95%. Swati et al. (2019) used a block-wise fine-tuning algorithm based on transfer learning to fine-tune pretrained CNN on an MRI brain tumor dataset and obtained average accuracy of 94.82% under five-fold cross validation. Rehman et al. (2019) employed three pretrained CNNs (AlexNet (Krizhevsky et al., 2017), GoogLeNet (Zeng et al., 2016), and VGGNet (Simonyan and Zisserman 2015)) to classify brain tumor MRI images with two different transfer learning techniques (fine-tune and freeze), and the fine-tuned VGG16 architecture showed the highest accuracy of 98.69%.

Smailagic et al. (2018) sampled the instances which had the longest distance from other training samples in a learned feature space. The proposed strategy reduced the annotated examples by 32% and 40%, respectively, compared to the conventional uncertainty and random sampling methods on the task of

diabetic retinopathy detection. Dai et al. (2020) proposed a gradient-guided suggestive annotation framework which computes gradient of training loss and then selects informative examples which have the shortest Euclidean distance to the gradient-integrated samples projected onto the data manifold learned by a variational auto-encoder (VAE). Through employing this framework, they selected 19% of the MRI images from BRATS 2019 dataset to train a CNN for brain tumor segmentation task and achieved competitive results (a Dice score of 0.853) compared with when the whole labeled dataset was used. Zhou et al. (2017) augmented each sample by data augmentation technique, and then computed entropy and relative entropy for original and augmented samples. Next, they continuously selected the most uncertain samples to label and added them to the training dataset to fine-tune AlexNet at each iteration. They managed to cut the needed annotated training data by half in three different biomedical imaging applications. Li et al. (2019) proposed an active learning strategy for breast cancer classification on pathological image dataset. Instead of selecting the most informative samples, the algorithm removed 4,440 misleading samples from the training dataset which contained 68,640 samples. They obtained patch-level average classification accuracy of 97.63%, compared to 85.69% which was resulted by training on the whole dataset.

Methods

Transfer learning is a widely used approach in which a network is trained on a large labeled source dataset, and the resulting pretrained network is fine-tuned on the small target dataset, transferring the learned knowledge from the source to target dataset. Active learning, on the other hand, is a promising strategy which has been investigated to train a competitive classifier with minimal annotation cost. In this retrospective work, transfer learning and active learning are the components of our proposed uncertainty sampling method for achieving stable test results using a smaller subset of training cohort. We chose the MICCAI BRATS 2019 dataset (Menze et al., 2015; Bakas et al., 2017; Bakas et al., 2018) as the target dataset, which is a new, well-annotated, well-preprocessed, and skull-striped dataset with interpolation and registration.

Dataset

All the experiments in this work were performed on the BRATS 2019 dataset which consists of 335 patients diagnosed with brain tumors (259 patients with HGG and 76 patients with LGG). According to the available age information of 240 patients, the mean age is 60.31 years. Each patient MRI scan set has four MRI sequences, which are T1-weighted, post–contrast-enhanced T1-weighted (T1C), T2-weighted (T2), and T2 fluid-attenuated inversion recovery (FLAIR) volumes. The dataset was preprocessed with skull-striping, interpolated to a uniform isotropic resolution of 1 mm^3, and registered to SRI24 space with a dimension of 240 × 240 × 155. The annotations of the dataset include four labels: background, gadolinium-enhancing tumor, the peri-tumoral edema, and the necrotic and non-enhancing tumor core. The area identified by the last three of the four labels represents the complete tumor region.

TABLE 1 | Detailed architecture of AlexNet.

Layer	Kernel size	Stride	Padding	Output size
Conv1	11 × 11	4	2	64 × 55 × 55
Maxpool1	3 × 3	2	0	64 × 27 × 27
Conv2	5 × 5	2	2	192 × 27 × 27
Maxpool2	3 × 3	2	0	192 × 14 × 14
Conv3	3 × 3	1	1	384 × 13 × 13
Conv4	3 × 3	1	1	256 × 13 × 13
Conv5	3 × 3	1	1	256 × 13 × 13
Maxpool3	3 × 3	2	0	256 × 6 × 6
FC1	–	–	–	4096 × 1
FC2	–	–	–	4096 × 1
FC3	–	–	–	2 × 1

To implement the proposed method in this work, we randomly extracted 20 slices with the tumor region from each patient MRI scan in axial plane, and kept T1, T1C, and T2 channels for each slice. The pretrained AlexNet requires three channel input, and we chose T1, T1C, and T2 channels from total four channels based on the results of the initial experiments. The obtained 6,700 2D 3-channel slice dataset was further split into training set (203 patients), validation set (66 patients), and test set (66 patients). All the three cohorts have the same ratio of HGG patient number and LGG patient number as the full dataset. Every slice with LGG tumor was annotated as label 0, and HGG tumor slices were labeled as 1. The images were resized from 240 × 240 pixels to 224 × 224 pixels in order to fit the pretrained CNN.

Transfer Learning

Training a CNN from scratch (with random initialization) requires massive amount of annotated training samples and relatively more time and computational resources than employing a CNN pretrained on a very large dataset. In general, there are two main scenarios of transfer learning: fine-tuning and freezing. In fine-tuning, instead of random initialization, weights and biases of a pretrained CNN are adopted, and then a conventional training process on the target dataset is performed. In the freezing scenario, we consider the pretrained CNN layers as a fixed feature extractor. In this context, we freeze the weights and biases of our desired convolutional layers, and let the fully connected layers be fine-tuned over the target dataset. The frozen layers do not have to be limited to the convolutional layers. Frozen layers can be chosen to be any subset of convolutional or fully connected layers; however, a common practice is to freeze the shallower convolutional layers. In our research, the CNNs are pretrained on ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) dataset (Russakovsky et al., 2015) which includes natural images. Due to the large difference between our target medical image domain and the ImageNet dataset, we chose fine-tuning to be our strategy of transfer learning.

Based on the purpose of reducing the annotation cost, we opted the pretrained AlexNet and fine-tuned it on the BRATS 19 dataset. AlexNet is composed of five convolutional layers, three max-pooling layers, and three fully connected layers. The

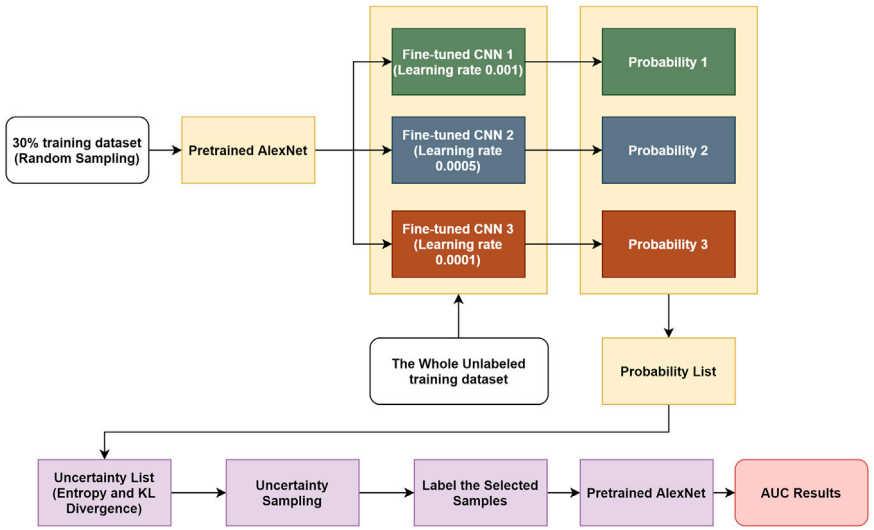

FIGURE 1 | Workflow of proposed transfer learning–based active learning framework.

detailed architecture used in this work is shown in **Table 1**. AlexNet depth is capable for brain tumor classification, and it is considerably shallower than other benchmark CNNs (e.g., ResNet (He et al., 2016) and VGG (Simonyan and Zisserman 2015)), which leads to faster convergence and less required computational resources.

Uncertainty Score Calculation

We use entropy and relative entropy as measures to estimate the informativeness of each training example. Given a discrete random variable X, with possible outcomes x_1, $x_2, \ldots,$ x_n which occurs with probabilities $\mathrm{P}(x_1)$, $\mathrm{P}(x_2)$, $\ldots$, $\mathrm{P}(x_n)$, the entropy formula of X is given by **Equation (1)**.

$$\mathrm{H}(\mathrm{X}) = -\sum_{i=1}^{n} P(x_i)\log P(x_i) \tag{1}$$

.

Another useful measure for estimating the amount of mutual information between two possibility distributions on a random variable is relative entropy, also known as the symmetric Kullback–Leibler (KL) divergence. Formally, given two probability distributions $\mathrm{P}(x)$ and $\mathrm{Q}(x)$ over a discrete random variable X which has n possible outcomes, the relative entropy given by $\mathrm{D}(\mathrm{p}||\mathrm{q})$ is given by **Equation (2)**.

$$\mathrm{D}(\mathrm{p}||\mathrm{q}) = \sum_{i=1}^{n} P(x_i)\log\frac{P(x_i)}{Q(x_i)} \tag{2}$$

.

In this scenario, the probability distributions are the outputs of the pretrained CNNs.

Workflow

In this work, we present a novel transfer learning–based active learning framework to reduce the annotation cost while maintaining stability and robustness of CNN performance for brain tumor classification. Our active learning workflow is described in **Figure 1**.

We assume the training dataset consists of labeled and unlabeled subsets. The goal is to find the best informative samples in the entire training set, which may or may not overlap with the labeled training subset. The workflow is divided into four steps: 1) For the labeled training subset, we randomly selected 30% training samples and assumed the remaining 70% samples were unlabeled. We then used the 30% labeled training subset to fine-tune the pretrained AlexNet, and the learning rate α was set to different values (i.e., 0.001, 0.0005, and 0.0001). By performing this step, we obtained three fine-tuned CNNs. 2) We used these fine-tuned CNNs to compute the classification probabilities of each sample in the entire training dataset. In this step, the CNNs only perform forward propagation to calculate outputs; therefore, no labels are required. 3) Once each training sample produced three predicted possibilities in step 2, we computed the individual entropy (**Equation 1**) and pairwise KL divergence (**Equation 2**). The uncertainty score is the sum of the entropy and KL divergence of each sample. Through this approach, an uncertainty score list of the entire training dataset was obtained. 4) We sorted the uncertainty score list in descending order, and we sampled 30% of the training cohort, which consisted of the best informative samples. This selected subset required labeling and was consequently used to fine-tune a pretrained AlexNet.

If there was no overlap between the original labeled training subset (30%) and the discovered best informative subset (30%), then the maximum training size needed is 30 + 30 = 60% (40% reduction in training size) of the entire training cohort. If the discovered best informative samples happen to be exactly the same as the original labeled training subset (30%), then the maximum training size needed is only 30% (70% reduction in training size) of the entire training cohort. In other words, between 40 and 70% of annotation cost (average of 55%) can

TABLE 2 | AUC results of AlexNet trained from scratch and fine-tuned from the pretrained model.

AUC (95% CI)	Pretrained AlexNet	AlexNet trained from scratch
Validation dataset	87.46% (87.11, 87.81)	86.14% (85.60, 86.68)
Test dataset	79.91% (78.95, 80.87)	71.93% (70.76, 73.10)

be saved by our proposed transfer learning–based active learning framework.

RESULTS

All the experiments were conducted on a NVIDIA GeForce RTX 2070 platform, using Python 3.8 and PyTorch 1.5.1. In order to prove the stability and reproducibility, all the AUC results below are averages of 10 runs of a single experiment and presented as mean along with the 95% confidence interval (CI).

In *Results of Using Transfer Learning*, we will show transfer learning is an effective approach and improves our baseline models. In *AUC Results of Selecting a Different Range of Uncertainty Distribution*, we will demonstrate the top 10% certain and uncertain examples are not informative, and thus, omitting them helps the models to better generalize. In *AUC Results of the Uncertainty Sampling Method*, we will experimentally show our uncertainty sampling approach improves the baseline with sample size fixed at 30%. Finally, in *AUC Results of the Uncertainty Sampling Method on Balanced Dataset*, we will demonstrate the following: 1) Regardless, if the dataset is balanced or imbalanced, our sampling method is effective. 2) The fact that our sampling approach improves the baseline is not arbitrary or as a result of filtering noisy examples through chance. It in fact always outperforms random sampling. 3) Although 30% is the optimum sample size, our sampling method works at other sample sizes as well.

Results of Using Transfer Learning

Training AlexNet from scratch requires massive data with high-quality annotation. Employing transfer learning technique improves performance of the model when sufficient data are not available. The baseline AUC was computed by fine-tuning the pretrained AlexNet on the entire training dataset. The maximum number of epochs was 30, the learning rate was set to 0.001, the batch size was set to 16, momentum in stochastic gradient descent (SGD) optimizer was 0.8, and L2 regularization penalty was set to 0.0001 based on a grid search strategy. We also explored training AlexNet from scratch, with the same hyperparameter settings, except that epoch number was increased to 80 because it needed more iterations to converge.

Table 2 lists AUC results with and without transfer learning strategy on both validation dataset and test dataset. As it can be seen, the validation AUC and test AUC improved by 1.51% and 7.98%, respectively, when employing the transfer learning method.

AUC Results of Selecting a Different Range of Uncertainty Distribution

As described in *Workflow*, we fine-tuned the pretrained AlexNet on 30% of the training dataset, which was labeled, and obtained three fine-tuned CNNs with learning rate α set to 0.001, 0.0005, and 0.0001, respectively. The uncertainty score list of the entire training samples was computed based on the output of these CNNs. **Figure 2** visualizes uncertainty distribution of the training dataset, where uncertainty score list is unsorted in **Figure 2A**, and uncertainty scores are ranked in the descending order in **Figure 2B**.

While keeping the number of samples constant (i.e., 30% of the training dataset), we fine-tuned the pretrained AlexNet on different ranges of uncertainty distribution. This was done to assess the effect of sampling from diverse uncertainty ranges on the performance of the CNN. In all experiments, we stopped our training or fine-tuning procedure at the highest validation AUC.

As reflected in **Figure 3**, AUC results for validation and test sets were calculated on samples from different uncertainty ranges

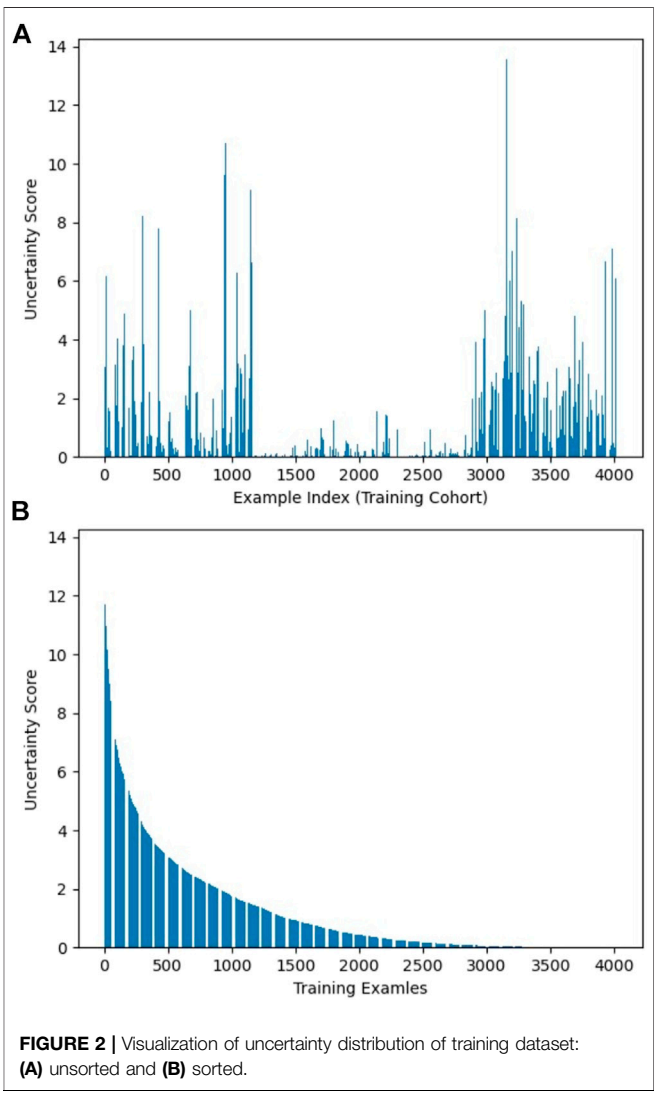

FIGURE 2 | Visualization of uncertainty distribution of training dataset: **(A)** unsorted and **(B)** sorted.

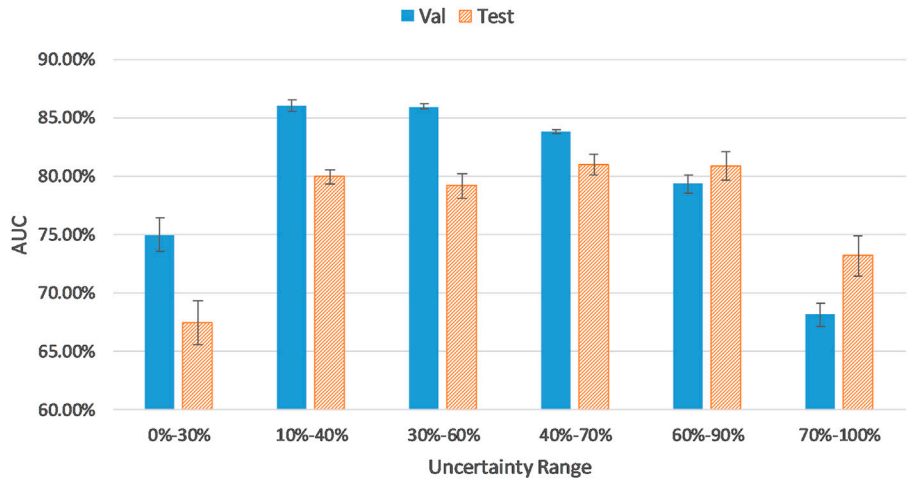

FIGURE 3 | CNN performance on samples from different uncertainty ranges.

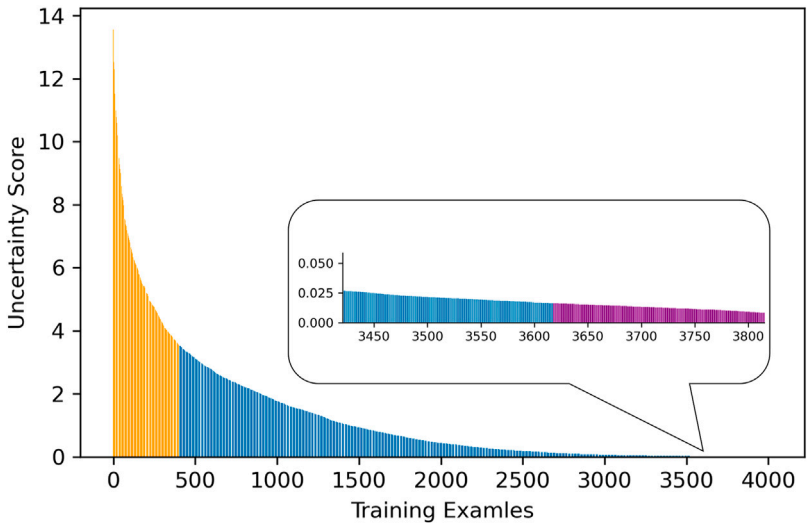

FIGURE 4 | The distribution of 10% examples with the highest and lowest uncertainty scores.

according to the sorted uncertainty list. As it can be seen, the biggest jumps of validation AUC occur when the first and last 10% of the sorted list (the top 10% certain and uncertain examples) are excluded. As shown in **Figure 3**, using the top 30% certain examples or the top 30% uncertain examples results in a decrease of AUC results for the validation (and test) cohort. Thus, we removed the top 10% (highest uncertainty scores) and the bottom 10% (lowest uncertainty scores) samples to eliminate outliers with least training values. As it can be seen in **Figure 3**, the uncertainty range of 10–40% improves AUC results by 12.51% compared to the range of 0–30%. Similarly, the uncertainty range of 60–90% elevates AUC by 7.72% in comparison to the range of 70–100%.

TABLE 3 | AUC results of the proposed method and baseline AUC.

AUC (95% CI)	Proposed method	Baseline
Validation dataset	86.86% (86.48, 87.24)	87.46% (87.11, 87.81)
Test dataset	82.89% (81.87, 83.91)	79.91% (78.95, 80.87)

The distribution and proportion of the hardest 10% samples and the easiest 10% samples in the entire uncertainty distribution are visualized in orange color and purple color, respectively, in **Figure 4**. We hypothesize the top 10% uncertain examples are outliers, and the bottom 10% do not provide training value for the model, which will result in a poor model generalization.

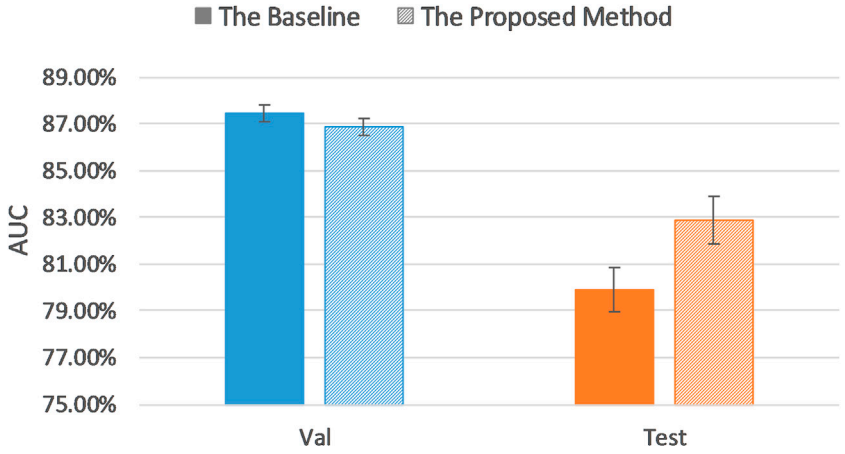

FIGURE 5 | Comparison of AUC results of the proposed method and baseline.

TABLE 4 | AUC results of the proposed method and baseline AUC on the balanced dataset.

AUC (95% CI)	Proposed method	Baseline
Validation dataset	85.20% (84.88, 85.52)	87.17% (86.87, 87.47)
Test dataset	82.00% (81.18, 82.82)	78.48% (77.60, 79.36)

AUC Results of the Uncertainty Sampling Method

In our uncertainty sampling algorithm, in order to train a model with better generalization, we discarded the top 10% and the bottom 10% training examples to eliminate outliers and least informative samples, respectively. Next, we randomly sampled 30% of the entire training cohort from the remaining dataset. We hypothesized that because this sample set did not include the top and bottom most uncertain and certain cases, it was the best informative and representative part of the dataset, and hence, we used it to fine-tune a pretrained AlexNet, in order to achieve competitive model performance compared with using the whole training dataset.

Table 3 lists model classification performance based on the proposed uncertainty sampling method and compares it with the baseline in which we fine-tuned the pretrained AlexNet on the entire training dataset. **Figure 5** illustrates contents of the **Table 3**.

It can be seen that our proposed uncertainty sampling method achieved similar classification performance on the validation dataset, and the AUC on the test set was 2.92% higher than the baseline AUC. Overall, the proposed method could save 40–70% of labeling cost while maintaining high classification performance of the model.

AUC Results of the Uncertainty Sampling Method on Balanced Dataset

For the purpose of verifying the robustness of our proposed method, we further created a balanced dataset and applied uncertainty sampling method. In order to better control the variables, we did not change the way the training, validation, and test sets were divided. Rather, we changed the number of slices extracted from each patient's MRI scan. Because the ratio of the number of HGG patients (259 patients) and LGG patients (76 patients) is close to 3:1, the ratio of the number of HGG and LGG slices can be changed to 1:3 to form a balanced dataset. Therefore, 30 slices were extracted from MRI scan instead of 20 slices for each LGG patient, and the number of MRI slices for every HGG patient reduced from 20 slices to 10 slices. This yielded a dataset of 4,870 2D 3-channel slices.

The baseline AUC was computed when the pretrained AlexNet was fine-tuned on the entire balanced training set, and the uncertainty sampling method was the same as described previously. As **Table 4** and **Figure 6** indicate, even on a balanced dataset, our proposed method achieved better classification performance than the baseline test AUC with significantly less annotations, which demonstrates robustness of our uncertainty sampling method.

Comparison of AUC Results of the Uncertainty Sampling Method and the Random Sampling Method

In the previous sections, our sample size was fixed at 30% of the training dataset, excluding the top and bottom most certain and uncertain samples. In this section, we investigate the effect of the sample size. In order to compare the efficacy of the uncertainty sampling method and the random sampling method, we fine-tuned the pretrained AlexNet on the fixed number of examples which were created using these two sampling methods.

In our random sampling method, we started with random sampling of 10% of the training cohort (N samples), and then increased the number of samples by 10% of the training dataset (N) until it accounted for 80% of the total training set samples (8xN) (top and bottom 10% already removed). Thus, 8 sampled datasets with a sample size of 10–80% (N to 8xN) of the total training set were obtained, with interval of 10% (N). For the

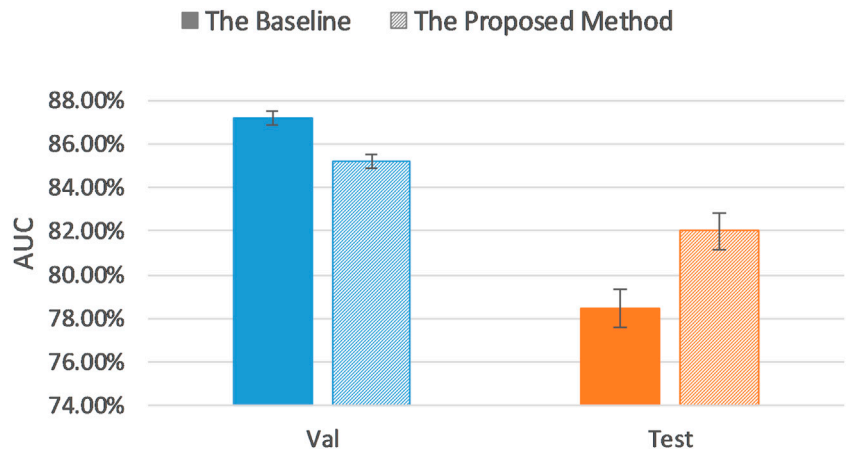

FIGURE 6 | Comparison of AUC results of the proposed method and baseline on the balanced dataset.

TABLE 5 | Correspondence between the proportion of sample size and the number of examples on the imbalanced dataset and the balanced dataset.

Proportion of sample size		**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**
Number of examples	Imbalanced dataset	406	812	1218	1624	2030	2436	2842	3248
	Balanced dataset	487	974	1461	1948	2435	2922	3409	3896

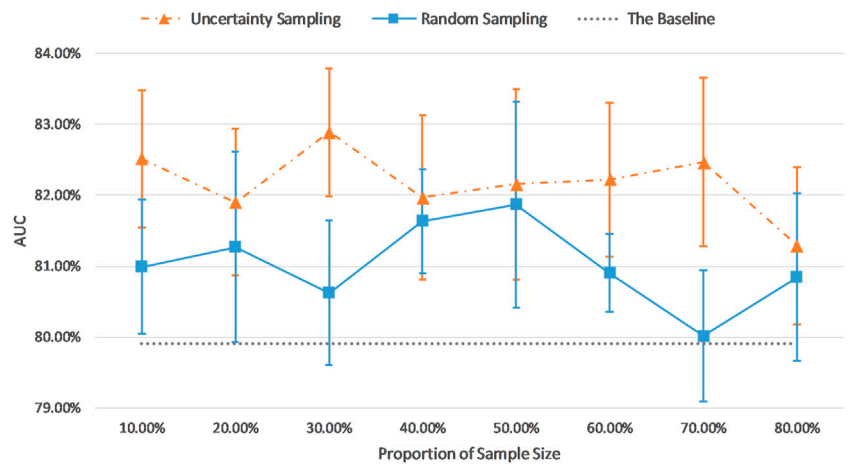

FIGURE 7 | Comparison of test AUC results of the uncertainty sampling method and random sampling method on the imbalanced dataset.

uncertainty sampling method, we removed the top 10% and bottom 10% samples according to the sorted uncertainty list, and randomly selected a subset whose sample size is 10% of the total training cohort (N) from the remaining part of the dataset. Similar to the previous sampling process, we created eight different datasets and conducted our experiments on them. **Table 5** describes the details of correspondence between the proportion of sample size and the number of examples on imbalanced and balanced datasets.

Figure 7 and **Figure 8** show the visualizations of test AUC results using the uncertainty sampling method and the random sampling method on the imbalanced as well as the balanced datasets. In each figure, the solid and dash dotted lines indicate the AUC values obtained on the samples corresponding to the parameters of the horizontal axis, and the dotted lines represent the baseline AUC which were computed when the pretrained CNN was trained on the entire labeled training datasets. The two colors orange and blue in each figure represent AUC results calculated by the uncertainty sampling method and the random sampling method, respectively.

As shown in **Figures 7**, **8**, for both imbalanced and balanced datasets, our proposed method performs better than the random sampling method, and the AUC results are higher than the

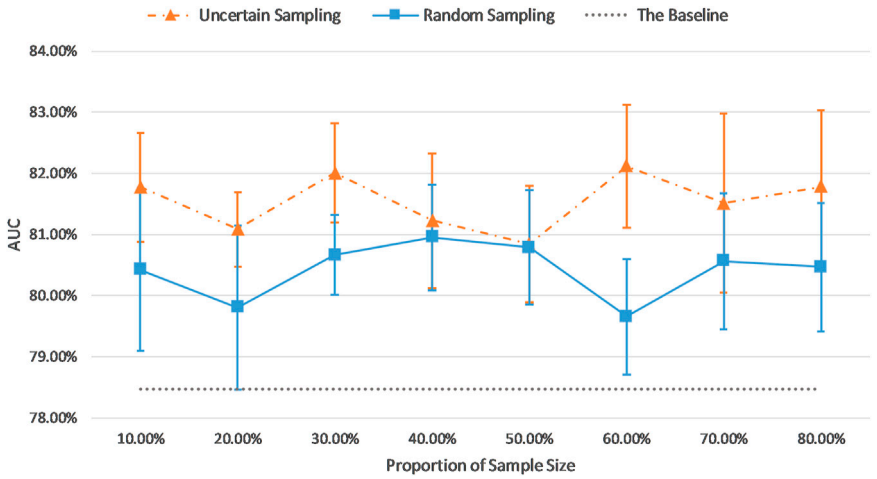

FIGURE 8 | Comparison of test AUC results of the uncertainty sampling method and random sampling method on the balanced dataset.

baseline on every proportion of sample size, which demonstrates the stability and robustness of our proposed uncertainty sampling strategy.

DISCUSSION

Deep learning algorithms for detection of tumors in medical images require large annotated datasets for training. The annotation is usually done manually by subspecialty radiologists. The associated cost (time and expertize) is prohibitively high, which hinders success of AI in medical imaging. Transfer learning is a widely used approach which can transfer the knowledge that the model has learned on large datasets to the new recognition and classification tasks. Active learning algorithms have been investigated to train a competitive classifier with minimal annotation cost. In this work, we combine transfer learning and active learning to propose a novel uncertainty sampling method which can reduce the amount of required training samples while maintaining stability and robustness of CNN performance for brain tumor classification.

There are two important metrics to describe the informativeness of an unlabeled sample: uncertainty, which is the inverse of the confidence of predicted results by the model; and representativeness, which measures the degree of similarity in distribution and structure between selected samples and target dataset (Du et al., 2017). Most studies consider only one of these two metrics. For example, Smailagic et al. (2018) selected samples which had the longest distance from other training samples in a learned feature space as the informative subset. Although this is an efficient approach in terms of uncertainty metric, it is hard to detect outlier samples because it is not guaranteed that the selected samples are all representative of the whole training cohort. Our proposed active learning method integrates the traditional uncertainty sampling technique and the query-by-committee method (Settles 2011), which selects the subset of informative samples in terms of both uncertainty and representativeness.

All BraTS multimodal scans were acquired with different clinical protocols and various scanners from multiple (n = 19) institutions. The details about the patients' demographics, region, racial diversity, clinical setting, and data extraction techniques are not provided, and it is highly possible that they are not exactly the same in these 19 institutions. The patients in the test set were randomly sampled from the dataset with the same ratio of HGG and LGG cases as those in the full dataset. Therefore, the selected diverse test set has a good representation of the population cohort, which could provide a valid and comprehensive evaluation on the model performance.

Our proposed sampling method selects samples with representativeness and informativeness by discarding subsets of training samples with the highest and lowest uncertainty scores. We set the proportion of discarded samples as 10% because the top 10% examples with highest uncertainty and the bottom 10% samples with the lowest uncertainty resulted in a poor model generalization as shown in **Figure 3**. We then had multiple options for using the remaining 80% of the training dataset. Our experiments revealed that a sample as big as 30% of the dataset is the optimum choice (**Figures 7**, **8**). By using 30% of the training dataset conditioned on excluding top and bottom 10% of our uncertainty list, the uncertainty sampling method achieved AUC of 82.89% and 82.00% on the imbalanced and balanced datasets, respectively, which was comparable or better than the baseline AUC. Although the best sampling size for the balanced dataset would be 60%, given the slight difference between AUC results at 30 and 60% (82.00 vs. 82.11%), we chose 30% to save a considerable amount of labeling costs and to be consistent with the imbalanced scenario. The proposed method can save 40–70% of the labeling cost. We also compared our uncertainty method with random sampling and demonstrated that our proposed method outperforms random sampling. It should be noted that random sampling is inherently unstable compared to the proposed systematic sampling

approach, and the results for random sampling are not reliable as they may not be repeatable.

To apply the proposed method in a prospective setting and generalize to other cohorts, the same hyperparameter setting proposed in this research can be used to fine-tune the pretrained CNNs and obtain the list of uncertainty scores for the entire training dataset. According to the distribution of the obtained uncertainty scores, we could set the threshold proportion to discard the samples with extreme scores. Then the proportion of samples selected from the remaining datasets can be set to 30%, similar to this research or based on further analysis of the new training data.

Although there is no mathematical proof or guarantee that our results will generalize to other medical imaging datasets, our research introduces an annotation reduction method for AI applied to medical imaging projects, which was proved to effectively reduce annotation cost in brain tumor classification task.

CONCLUSION

A transfer learning–based active learning framework can significantly reduce the size of required labeled training data while maintaining high accuracy of the classification of tumors in brain MRI.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://www.med.upenn.edu/cbica/brats2019/data.html.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

RH, KN, and FK contributed to the design of the concept and the study, and the design and implementation of machine learning modules. RH, KN, and FK contributed to the writing and reviewing of the first draft of the manuscript. RH, KN, LL, and FK contributed to the manuscript revision and approved the final manuscript.

FUNDING

This research received funding support in part from China Scholarship Council, and Chair in Medical Imaging and Artificial Intelligence, a joint Hospital-University Chair between the University of Toronto, the Hospital for Sick Children, and the SickKids Foundation.

ACKNOWLEDGMENTS

This manuscript has been released as a preprint at arXiv (Hao et al., 2020).

REFERENCES

Afshar, P., Mohammadi, A., and Plataniotis, K. N. (2018). Brain Tumor Type Classification via Capsule Networks. Proceedings–International Conference on Image Processing, ICIP, Athens, Greece, October 7–October 10, 2018, 3129–3133. doi:10.1109/ICIP.2018.8451379

Badža, M. M., and Barjaktarović, M. C. (2020). Classification of Brain Tumors from Mri Images Using a Convolutional Neural Network. *Appl. Sci. (Switzerland)* 10 (6). doi:10.3390/app10061999

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017). Advancing The Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features. *Scientific Data* 4 (September), 170117. doi:10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). "*Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge*." arXiv [Preprint]. Available at: https://arxiv.org/abs/1811.02629

Ballal, D. R., and Zelina, J. (2004). Progress in Aeroengine Technology (1939–2003). *J. Aircraft* 41 (1), 43–50. doi:10.2514/1.562

Banerjee, S., Mitra, S., Masulli, F., and Rovetta, S. (2019). "Brain Tumor Detection and Classification from Multi-Sequence MRI: Study Using ConvNets," in *In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Editors C Alessandro, B Spyridon, K Hugo, K Farahani, R Mauricio, and W Theo van, 170–179. (Cham: Springer International Publishing).

Claus, E. B., Walsh, K. M., Wiencke, J., Annette, M., Wiemels, J. L., Joellen, M., et al. (2016). Survival and Low-Grade Glioma: The Emergence of Genetic Information, *Neurosurg Focus* 38 (1), 1–19. doi:10.3171/2014.10.FOCUS12367

Dai, C., Wang, S., Mo, Y., Zhou, K., Angelini, E., Guo, Y., et al. (2020). "Suggestive Annotation of Brain Tumour Images with Gradient-Guided Sampling," .In A.L. Martel (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020, Lima, Peru, October 4-8, 2020: Springer, Cham, 1–10.

Das, S., Riaz Rahman Aranya, O. F. M., and Nishat, N. L. (2019). "Brain Tumor Classification Using Convolutional Neural Network." In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, May 3–5, 2019 1–5. doi:10.1109/icasert.2019.8934603

Delattre, J-Y., Bernsen, H. J. J. A., Frenay, M., Tijssen, C. C., and Grisold, W. (2014). Adjuvant Procarbazine , Lomustine , and Vincristine Chemotherapy in Newly Diagnosed Anaplastic Oligodendroglioma : Long-Term Follow-Up of EORTC Brain Tumor Group Study 26951. *J. Clin. Oncol.* 31 (3), 344–350. doi:10.1200/JCO.2012.43.2229

Du, B., Wang, Z., Zhang, L., Zhang, L., Liu, W., Shen, J., et al. (2017). Exploring Representativeness and Informativeness for Active Learning. *IEEE Trans. Cybern.* 47 (1), 14–26. doi:10.1109/tcyb.2015.2496974

Essig, M., Anzalone, N., Combs, S. E., Dörfler, A., Lee, S-K., Picozzi, P., et al. (2012). MR Imaging of Neoplastic Central Nervous System Lesions: Review and Recommendations for Current Practice. *AJNR Am. J. Neuroradiol* 33 (5), 803–817. doi:10.3174/ajnr.a2640

Hao, R., Namdar, K., Liu, L., and Khalvati, F. (2020). *A Transfer Learning Based Active Learning Framework for Brain Tumor Classification*. arXiv [Preprint]. Available at: https://arxiv.org/abs/2011.09265

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 27–30, 2016.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. W. L. (2018). Artificial Intelligence in Radiology. *Nat. Rev. Cancer* 18 (8), 500–510. doi:10.1038/s41568-018-0016-5

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet Classification with deep convolutional neural networks. *Commun. ACM* 60 (6), 84–90. doi:10.1145/3065386

Li, Y., Xie, X., Shen, L., and Liu, S. (2019). Reverse Active Learning Based Atrous DenseNet for Pathological Image Classification. *BMC Bioinformatics* 20 (1), 1–15. doi:10.1186/s12859-019-2979-y

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A Summary. *Acta Neuropathol.* 131 (6), 803–820. doi:10.1007/s00401-016-1545-1

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* 34 (10), 1993–2024. doi:10.1109/tmi.2014.2377694

Pereira, S., Meier, R., Alves, V., Reyes, M., and Silva, C. A. (2018). Automatic Brain Tumor Grading from MRI Data Using Convolutional Neural Networks And Quality Assessment."*Lecture Notes In Computer Science (Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics) 11038 LNCS*, 106–114. doi:10.1007/978-3-030-02628-8_12

Razzak, M. I., Naz, S., Zaib, A., and Ahmad, Z. (2018). Deep Learning for Medical Image Processing: Overview, Challenges and the Future. *Lecture Notes Comput. Vis. Biomech.* 26, 323–350. doi:10.1007/978-3-319-65981-7_12

Rehman, A., Naz, S., Imran, M., Akram, F., and Imran, M. (2019). A Deep Learning-Based Framework for Automatic Brain Tumors Classification Using Transfer Learning. *Circuits, Systems, Signal. Process.* 39, 757–775. doi:10.1007/s00034-019-01246-3

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). "ImageNet Large Scale Visual Recognition Challenge." *Int. J. Comput. Vis.* 115 (3), 211–252. doi:10.1007/s11263-015-0816-y

Settles, B. (2011). "Active Learning Literature Survay." *Mater. Lett.* 65 (5), 854–856. doi:10.1016/j.matlet.2010.11.072

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer Statistics, 2019: *CA A. Cancer J. Clin.* 69 (1), 7–34. doi:10.3322/caac.21551

Simonyan, K., and Zisserman, A. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition." In International Conference on Learning Representations, San Diego, United States, May 7–9, 2015. (ICLR, 1–14). doi:10.1016/j.infsof.2008.09.005

Smailagic, A., Noh, H. Y., Costa, P., Walawalkar, D., Khandelwal, K., Mirshekari, M., et al. (2018). "MedAL: Deep Active Learning Sampling Method for Medical Image Analysis." In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, December 17–20, 2018, doi:10.1109/icmla.2018.00078

Suganthe, R. C., Revathi, G., Monisha, S., and Pavithran, R. (2020). "Deep Learning Based Brain Tumor Classification Using Magnetic Resonance Imaging." *J. Crit. Rev.* 7 (9), 347–350. doi:10.31838/jcr.07.09.74

Swati, Z. N. K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S., et al. (2019). "Brain Tumor Classification for MR Images Using Transfer Learning and Fine-Tuning." *Comput. Med. Imaging Graphics* 75, 34–46. doi:10.1016/j.compmedimag.2019.05.001

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., et al. (2016). "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" *IEEE Trans. Med. Imaging* 35 (5), 1299–1312. doi:10.1109/tmi.2016.2535302

Yang, Y., Yan, L. F., Zhang, X., Han, Y., Nan, H. Y., Yu, C. H., et al. (2018). "Glioma Grading on Conventional MR Images: A Deep Learning Study with Transfer Learning." *Front. Neurosci.* 12, 804. doi:10.3389/fnins.2018.00804

Zeng, G., He, Y., Yu, Z., Yang, X., Yang, R., and Zhang, L. (2016). "Preparation of Novel High Copper Ions Removal Membranes by Embedding Organosilane-Functionalized Multi-Walled Carbon Nanotube." *J. Chem. Technol. Biotechnol.* 91 (8), 2322–2330. doi:10.1002/jctb.4820

Zhou, Z., Jae, S., Zhang, L., Gurudu, S., Gotway, M., and Liang, J. (2017). "Fine-Tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally." Proceedings–30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, HI, USA, July 21–26, 2017, 4761–4772. doi:10.1109/cvpr.2017.506

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.