Diagnosis of Brain Tumors from MRI using Deep Transfer Learning

Armita Kazemi

Century High School, Rochester, MN

Acknowledgement

I would like to thank and acknowledge my mentor, Oscar O’Rahilly for teaching me the skills necessary to complete this project, finding the dataset, and helping troubleshoot the code.

Abstract

Each year, more than 100,000 people in the United States are diagnosed with a brain tumor. An early and accurate diagnosis is crucial in getting patients the necessary treatment and increasing survival rates. In recent years, machine learning algorithms have become increasingly popular in the medical field due to their ability to recognize complex patterns and reduce human errors. However, accurate diagnosis using deep learning algorithms requires a large amount of training data, which is not always available. Additionally, training a model from scratch can take a long time and requires vast amounts of computational power. As a solution, this study aims to utilize a transfer learning method in which the prior knowledge of a pretrained model is used to aid in a new classification problem. In this study, a dataset of MRI images consisting of four classes (no tumor, pituitary tumor, meningioma, and glioma) were used. The performance of seven pretrained models (ResNet18, ResNet50, VGG16, DenseNet, GoogLeNet, ShuffleNet, and MobileNet) were evaluated in order to see which would achieve the highest classification accuracy. Additionally, this study examined two different methods for the implementation of transfer learning. In the first method, all layers of the pretrained model were frozen and in the second method, all layers of the pretrained model were trained. The best performing models proved to be ResNet18 and ShuffleNet with all layers trained, achieving an accuracy of 97.86%. The results also showed that the unfrozen models outperformed their frozen counterparts.

Table of Contents

Introduction

Each year, hundreds of thousands of people are diagnosed with brain tumors across the world. In 2021, 83,570 individuals were diagnosed with brain tumors in the United States, with 24,530 being malignant and 59,040 being benign. Of those, 18,600 patients died of the disease.

An efficient and accurate brain tumor diagnosis is critical in determining the best treatment and increasing the chances of recovery. Brain tumors are commonly misdiagnosed, affecting the health of thousands of patients and costing millions of dollars. In recent years, machine learning—a type of artificial intelligence (AI)—has emerged as a powerful tool for improving medical diagnosis. Machine learning programs train on large amounts of data and then identify structures and patterns in that data. They then use those patterns to predict answers to problems. Not only can machine learning greatly improve the accuracy of brain tumor diagnosis, but it can also aid countries with underdeveloped healthcare systems. In places like sub-Saharan Africa with an average of 0.2 doctors per 1,000 people, machine learning can provide a way to diagnose people that would be impossible otherwise.

The problem that arises with previous medical image classification methods is the lack of large, high-quality datasets. Machine learning models rely on large quantities of training data to achieve high classification accuracies. However, datasets of this scale are not available for medical image classification. The solution to this problem lies in transfer learning, a new deep learning technique that has been dominating the studies on image classification.

In this paper, we compare the performance of multiple pretrained Convolutional Neural Network (CNN) models in distinguishing between four classes of brain tumors (no tumor, pituitary tumor, meningioma, and glioma). Furthermore, we compare the implementation of transfer learning with two different methods (freezing the convolutional base and training the convolutional base) to discover which yields the best results. Following testing, a complete evaluation of each model is conducted. The top transfer learning model achieved the best classification performance compared to previous methods. Transfer learning also allowed the models to achieve high accuracies with a small number of training samples. The major contributions of this paper are listed below.

- A transfer learning approach is applied to a four class brain tumor classification task
- Award-winning CNN models are evaluated on the classification of brain tumor types
- For the first time, the significance of trainable layers in transfer learning is discovered in relation to brain tumor classification
- The proposed method achieves high accuracies compared to other methods while using a relatively small dataset

Materials and Methods

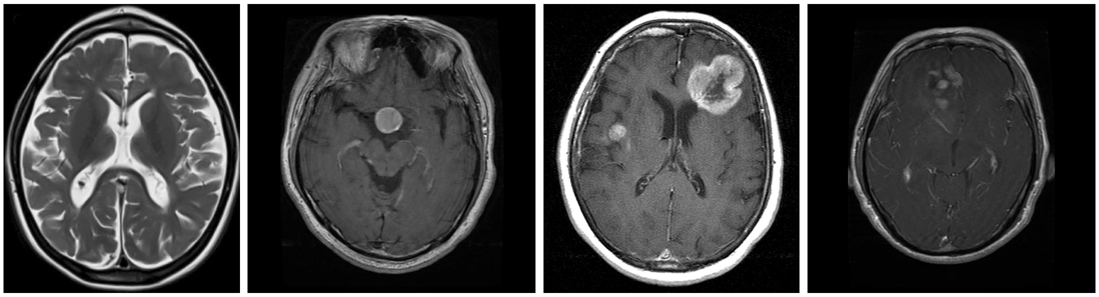
Dataset

The dataset used in this study was publicly available on Kaggle (Bhuvaji et al., 2020). It contains a total of 3264 brain MRI images. The images had a class distribution of 500 with no tumor, 901 with pituitary tumors, 937 with meningiomas, and the remaining 926 with gliomas. The images are available as .jpg files and the size of each image is 512×512. The tumor type labels were changed to numerical values by assigning 0 to no tumor, 1 to pituitary tumor, 2 to meningioma, and 3 to glioma.

In the data preprocessing stage, the images were resized to 224×224 and normalized to scale the intensity values between 0 and 1. The dataset was divided into 80% and 20% for training and testing, respectively. Then the dataset was converted to a PyTorch Tensor. The dataset was passed through the PyTorch DataLoader with a batch size of 32 and shuffling for the training set.

Figure 1

Sample Brain MRI Images of the Four Classes of Brain Tumors in This Study



No tumor Pituitary tumor Meningioma Glioma

Methodology

In this work, we applied different pre-trained machine learning models to the classification task via transfer learning.

Transfer learning is a machine learning technique in which the model uses the knowledge gained during training on a large dataset to solve a different but related problem (Figure 2). With transfer learning, there is no need to completely train a model from scratch, which eliminates the need for large datasets and great computational power.

The transfer learning models for this experiment were imported from PyTorch, with all the models pretrained on the ImageNet database. Each model consists of a convolutional base and a final classifier layer (Figure 3). The convolutional base generally consists of multiple convolution layers, pooling layers, and activation functions. The classifier layer consists of a few fully connected layers and pooling layers. The last fully connected layer of each original model was removed and instead replaced with a new fully connected layer with an output size of four to fit the specifications of our classification task.

During the training and testing stages, Adam was chosen as the optimizer with a learning rate of 0.001. The loss was calculated using the Cross-Entropy Loss function. Each model was trained and tested over 100 epochs.

In this paper, we used two methods of implementing the pretrained models in our classification task. The methods are described below (Figure 3).

i) In method 1, the convolutional base of the pretrained model was frozen, and only the classification layer was trained. This is because after each model is pretrained on a dataset, it associates specific weights with each layer of the model. By freezing the convolutional base, the weights assigned from training on the ImageNet dataset are preserved. This method results in a decrease in training and testing time as only the top classification layer is being trained.

ii) In method 2, both the convolutional base and the classification layer were trained. This method deletes all the old weights associated with ImageNet and assigns new weights that correspond to the new dataset. This method can achieve higher accuracy rates as it is adapting the pretrained features to the new data. Method 2 is preferred in scenarios where the source domain and target domain are different from each other. The drawbacks of this method is the increase in training and testing time for the model.

Figure 2

How Traditional Machine Learning Works Compared to Transfer Learning

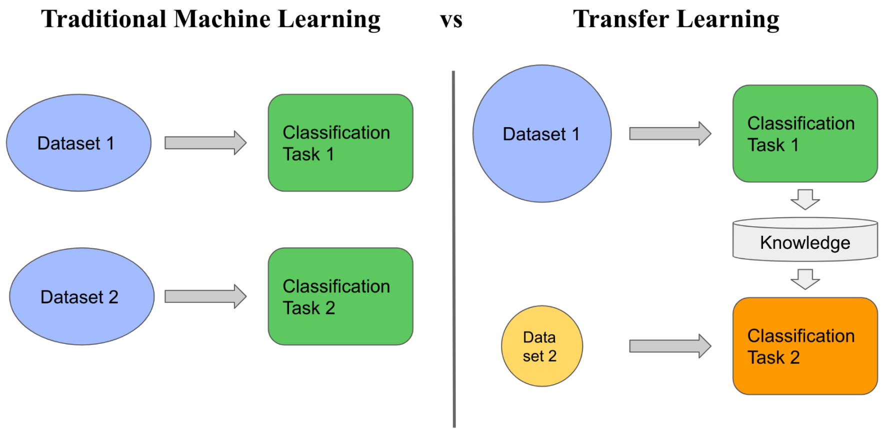
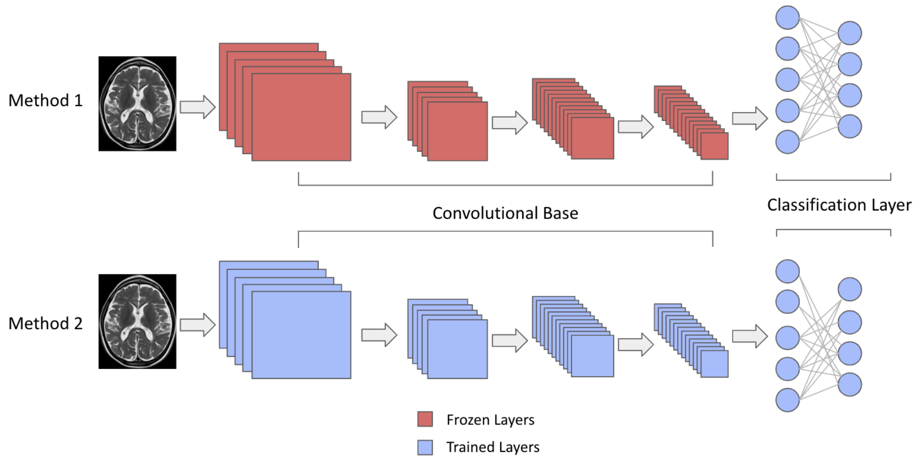


Figure 3

Different Implementation Methods of Transfer Learning



Models

In this study, we used the pretrained models ResNet18, ResNet50, VGG16, DenseNet, GoogLeNet, ShuffleNet, and MobileNet.

ResNet, short for Residual Network, is a type of neural network that was introduced in 2015 by He, Zhang, Ren, and Kan Sun in their paper “Deep Residual Learning for Image Recognition” (He et al., 2016). The ResNet models were extremely successful, as can be ascertained from the fact that they won 1st place in the ILSVRC 2015 classification competition with an error of only 3.57%. VGG 16 was proposed by Simoyan and Zisserman in 2014 in their paper “Very Deep Convolutional Networks for Large-Scale Image Recognition” (Simonyan et al., 2014). This model won 1st and 2nd place in the 2014 ILSVRC challenge for object detection and classification. GoogLeNet was proposed by researchers at Google in 2014 in the research paper “Going Deeper with Convolutions” (Szegedy et al., 2015). This architecture won the LSVRC 2014 image classification challenge. The MobileNetV2 model was introduced by Google in 2018 with the ability to run deep networks on personal mobile devices. The DenseNet model was introduced in the “Densely Connected Convolutional Network” paper in 2017 (Huang et al., 2017). ShuffleNet was introduced by Zhang, Zhou, Lin, and Sun in their paper "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices" (Zhang et al., 2018).

Results and Discussion

Metrics

The classification of brain tumors from MRIs into four classes was evaluated on five performance metrics. The most common metric is the classification accuracy, which is the percentage of correctly classified samples from the total number of samples. However, final classification accuracy is most reliable when the dataset contains an equal number of samples from each class. The unbalanced nature of our dataset requires further evaluation with more performance metrics. For further evaluation of each model, a confusion matrix was generated using Scikit Learn. The confusion matrix shows the distribution of true positives, true negatives, false positives, and false negatives. From the confusion matrix, the precision (1), recall (2), and F1 score (3) can be determined. The training and testing times per epoch were also recorded.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3}$$

Results

ResNet

As can be seen in Table 1, freezing the convolutional base (method 1) resulted in a classification accuracy of 89.13% by ResNet18 and 88.97% by ResNet50. ResNet18 had a shorter run time with an average of 6.59 seconds per epoch as opposed to ResNet50 with an average of 19.03 seconds per epoch. When the convolutional base was trained (method 2), ResNet18 achieved a classification accuracy of 97.86% and ResNet50 achieved 95.71%. The accuracies with method 2 were significantly higher than with method 1. However, the models ran longer with method 2 when compared to method 1. With method 2, ResNet18 had an average 11.59 seconds per epoch, while ResNet50 had 37.25 seconds per epoch.

Figures 4 and 5 show the confusion matrices for ResNet18, and ResNet50. ResNet18 had the highest accuracy for no tumor at 100% with method 2 and the lowest for meningioma at 80.11% with method 1. ResNet50 had the highest accuracy for pituitary tumors at 98.01% with method 2 and the lowest for glioma at 87.28% with method 1. For further evaluation, the precision, recall and F1 scores can be found in Tables 2 and 3.

Table 1

Results of Transfer Learning with Both Methods

Model	Accuracy	Best Accuracy	Average Time Per Epoch
ResNet18, method 1	89.13	89.43	6.59
ResNet18, method 2	**97.86**	**98.62**	11.59
ResNet50, method 1	88.97	91.11	19.03
ResNet50, method 2	95.71	97.24	37.25
DenseNet, method 1	91.58	92.19	42.00
DenseNet, method 2	94.95	97.55	81.43
GoogLeNet, method 1	86.22	86.37	7.88
GoogLeNet, method 2	88.97	**98.62**	14.97
ShuffleNet, method 1	86.83	86.83	**4.04**
ShuffleNet, method 2	**97.86**	98.32	6.58
MobileNet, method 1	88.97	89.74	7.29
MobileNet, method 2	97.55	98.47	13.99
VGG16, method 1	88.97	90.51	28.67
VGG16, method 2	86.52	86.83	53.30

Figure 4

Confusion Matrices for ResNet18 with Method 1 and Method 2, Respectively

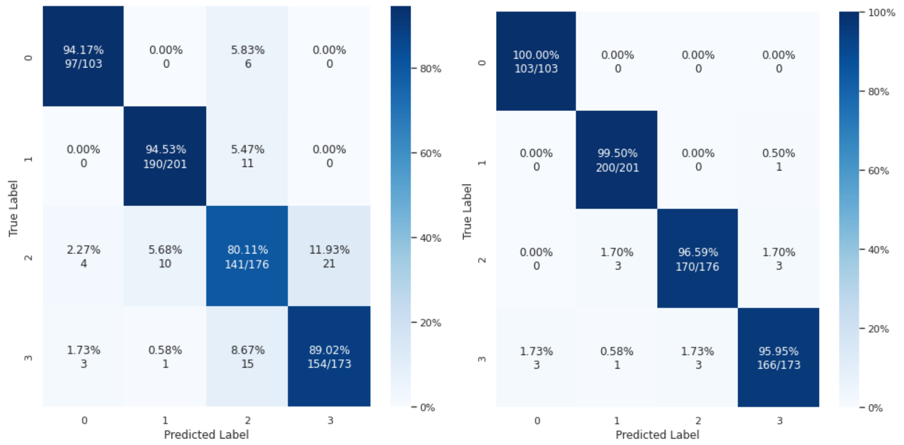


Figure 5

Confusion Matrices for ResNet50 with Method 1 and Method 2, Respectively

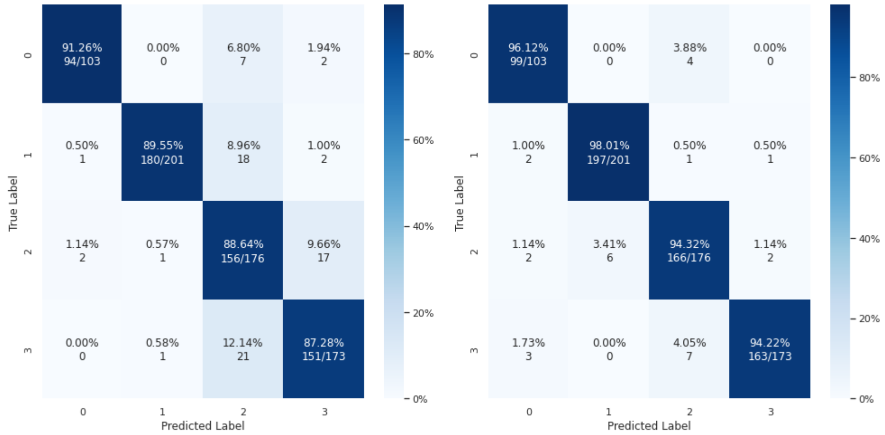


Table 2

Class-specific Evaluation of ResNet18

Method 1				Method 2			
Tumor Type	Precision	Recall	F1 Score	Tumor Type	Precision	Recall	F1 Score
No Tumor	93	94	94	No Tumor	97	100	99
Pituitary	95	95	95	Pituitary	98	100	99
Meningioma	82	80	81	Meningioma	98	97	97
Glioma	88	89	89	Glioma	98	96	97

Table 3

Class-specific Evaluation of ResNet50

Method 1				Method 2			
Tumor Type	Precision	Recall	F1 Score	Tumor Type	Precision	Recall	F1 Score
No Tumor	97	91	94	No Tumor	93	96	95
Pituitary	99	90	94	Pituitary	97	98	98
Meningioma	77	89	83	Meningioma	93	94	94
Glioma	88	87	88	Glioma	98	84	86

DenseNet

As can be seen in Table 1, DenseNet achieved the best classification accuracy out of all the models run using method 1 at 91.58%. With method 2, DenseNet achieved an accuracy of 94.95%. DenseNet took the longest time to run with both methods. With method 1, DenseNet averaged 42.00 seconds per epoch and with method 2 it averaged 81.43 seconds per epoch. These times are significantly longer than all the other models. The longer run time of DenseNet can be attributed to the enormous output that results from all the layers being connected in its architecture.

Figure 6 shows the confusion matrices for DenseNet. With method 1, the model had the highest accuracy for pituitary tumors at 99.41% and the lowest for meningioma at 84.07%. With method 2, the model had the highest accuracy for pituitary tumors at 100% and the lowest for glioma at 91.98%. For further evaluation, the precision, recall and F1 scores can be found in Table 4.

Figure 6

Confusion Matrices for DenseNet with Method 1 and Method 2, Respectively

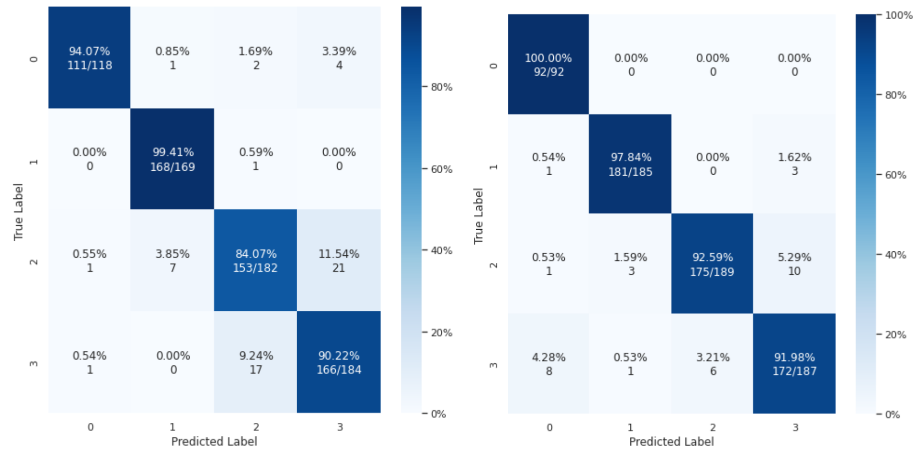


Table 4

Class-specific Evaluation of DenseNet

Method 1				Method 2			
Tumor Type	Precision	Recall	F1 Score	Tumor Type	Precision	Recall	F1 Score
No Tumor	98	94	96	No Tumor	90	100	95
Pituitary	95	99	97	Pituitary	98	98	98
Meningioma	88	84	86	Meningioma	97	93	95
Glioma	87	90	89	Glioma	93	92	92

GoogLeNet

As can be seen in Table 1, GoogLeNet had the lowest accuracy of all the models run using method 1 at 86.22%. With method 2, the accuracy increased to 88.97%, although this increase was not as significant as with other models. GoogLeNet had an average runtime of 7.88 seconds per epoch with method 1 and 14.97 seconds per epoch with method 2.

Figure 7 shows the confusion matrices for GoogLeNet. With method 1, the model had the highest accuracy for no tumors at 96.61% and the lowest for meningioma at 76.92%. With method 2, GoogLeNet had the highest accuracy for pituitary tumors at 99.46% and the lowest for glioma at 85.56%. For further evaluation, the precision, recall and F1 scores can be found in Table 5.

Figure 7

Confusion Matrices for GoogLeNet with Method 1 and Method 2, Respectively

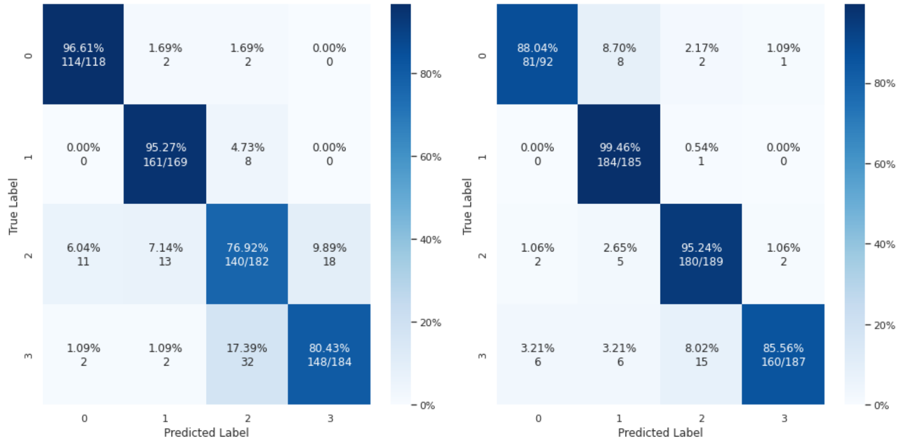


Table 5

Class-specific Evaluation of GoogLeNet

Method 1			
Tumor Type	Precision	Recall	F1 Score
No Tumor	90	97	93
Pituitary	90	95	93
Meningioma	77	77	77
Glioma	89	80	85

Method 2			
Tumor Type	Precision	Recall	F1 Score
No Tumor	91	88	90
Pituitary	91	99	95
Meningioma	91	95	93
Glioma	98	86	91

ShuffleNet

As can be seen in Table 1, ShuffleNet achieved a classification accuracy of 86.83% with method 1. The model improved significantly with method 2, achieving the highest classification accuracy of all the models with 97.86%. ShuffleNet proved to be the fastest model regardless of the method used. It ran at 4.04 seconds per epoch with method 1 and 6.58 seconds per epoch with method 2.

Figure 8 shows the confusion matrices for ShuffleNet. With method 1, the model had the highest accuracy for pituitary tumors at 92.31% and the lowest for glioma at 81.52%. With method 2, the model had the highest accuracy for pituitary tumors at 99.41% and the lowest for glioma at 95.65%. For further evaluation, the precision, recall and F1 scores can be found in Table 6.

Figure 8

Confusion Matrices for ShuffleNet with Method 1 and Method 2, Respectively

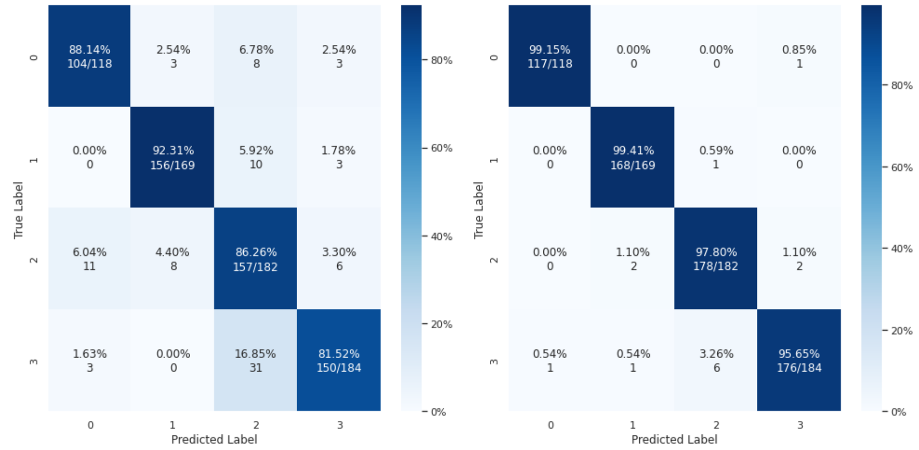


Table 6

Class-specific Evaluation of ShuffleNet

Method 1				Method 2			
Tumor Type	Precision	Recall	F1 Score	Tumor Type	Precision	Recall	F1 Score
No Tumor	88	88	88	No Tumor	99	99	99
Pituitary	93	92	93	Pituitary	98	99	99
Meningioma	76	86	81	Meningioma	96	98	97
Glioma	93	82	87	Glioma	98	96	97

MobileNet

As can be seen in Table 1, MobileNet achieved a classification accuracy of 88.97% when run with method 1. However, the model improved greatly when run with method 2, achieving the second highest accuracy of 97.55%. MobileNet was one of the fastest models, averaging 7.29 seconds per epoch with method 1 and 13.99 seconds per epoch with method 2.

Figure 9 shows the confusion matrices for MobileNet. With method 1, the model had the highest accuracy for pituitary tumors at 98.82% and the lowest for meningioma at 73.08%. With method 2, the model had the highest accuracy for glioma at 98.37% and the lowest for meningioma at 96.15%. For further evaluation, the precision, recall and F1 scores can be found in Table 7.

Figure 9

Confusion Matrices for MobileNet with Method 1 and Method 2, Respectively

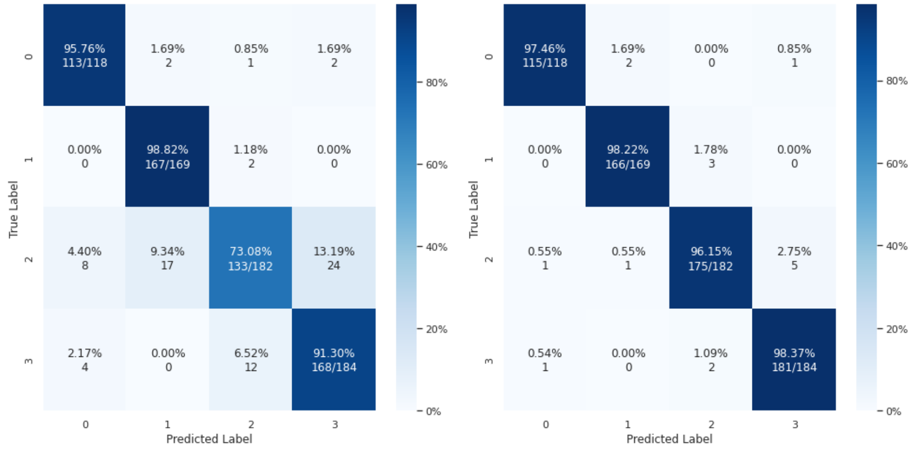


Table 7

Class-specific Evaluation of MobileNet

Method 1			
Tumor Type	Precision	Recall	F1 Score
No Tumor	90	96	93
Pituitary	90	99	94
Meningioma	90	73	81
Glioma	87	91	89

Method 2			
Tumor Type	Precision	Recall	F1 Score
No Tumor	98	97	98
Pituitary	98	98	98
Meningioma	97	96	97
Glioma	97	98	98

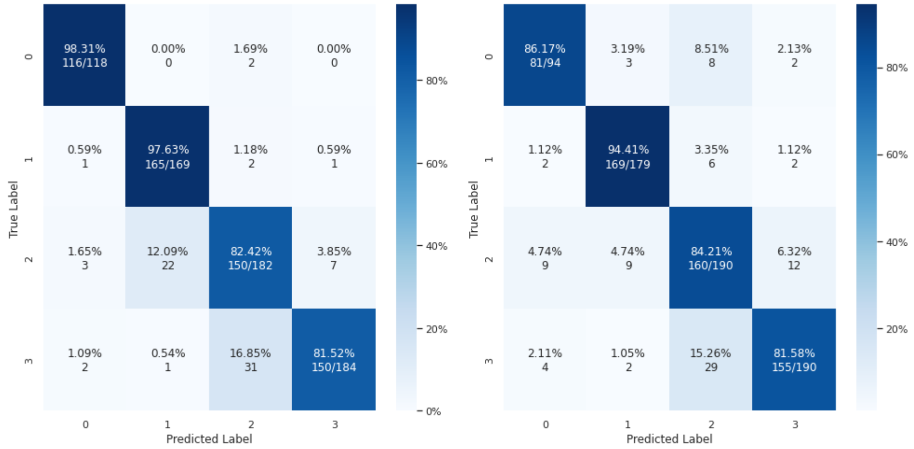
VGG16

As seen in Table 1, VGG16 performed with an accuracy of 88.97% when run with method 1, but the accuracy dropped to 86.53% with method 2. This is the opposite of what was seen with all the other models. VGG16 was also the second slowest model, with 28.67 seconds per epoch with method 1 and 53.30 seconds per epoch with method 2.

Figure 10 shows the confusion matrices for VGG16. With method 1 the model had the highest accuracy for no tumors at 98.31% and the lowest for glioma at 81.52%. With method 2 the model had the highest accuracy for pituitary tumors at 94.41% and the lowest for glioma at 81.58%.

Figure 10

Confusion Matrices for VGG16 with Method 1 and Method 2, Respectively



Discussion

The highest classification accuracy was achieved by ResNet18 and ShuffleNet at 97.86%. Both of these accuracies were achieved when the models were run using method 2. However, ShuffleNet ran almost twice as fast as ResNet18. The precision scores for ResNet18 and ShuffleNet were the same for pituitary tumors and glioma, and slightly different for no tumor and meningioma. ShuffleNet had a higher precision score for no tumors at 99% but ResNet had a higher recall score at 100%. ResNet18 had a higher precision score for meningioma, but it had a slightly lower recall than ShuffleNet. The F1 scores for both models were identical. High precision and recall scores are especially important when evaluating models on a medical classification problem. A high precision score indicates that most of the positives detected were true positives. A high precision score is critical in diagnosing brain tumors because it ensures that patients won't undergo unnecessary treatment due to false positives. On the other hand, a high recall score indicates that the model was able to correctly detect most occurrences of that brain tumor. A high recall score is especially important in diagnosing possibly malignant brain tumors because it ensures that there are very few false negatives. If there was a high false negative rate, it would mean that patients who actually had a brain tumor were falsely told they didn't, which could lead to delayed treatment and increased progression of the tumor.

On average, the models overwhelmingly had the highest classification accuracy for pituitary tumors and the lowest accuracy for meningioma or glioma. One possible cause for the lower accuracies of meningioma and glioma could be a difference in image quality compared to pituitary tumors. Additionally, meningioma and glioma were often misclassified as each other. The factor most likely contributing to this misclassification is the similar appearance of meningioma and glioma in the MRI images.

With both methods, the fastest models were ShuffleNet, Resnet18, and MobileNet. The slowest model for both methods was DenseNet. The long run time of DenseNet can be attributed to the model architecture. The DenseNet architecture interconnects all the layer in the model. While this method resulted in high accuracies, the architecture is not practical for deep neural nets as it increases the run time and consequently the computational costs. Additionally, while method 2 resulted in higher accuracies, it also doubled the run time for each model.

Future extensions of this work should explore implementing machine learning to classify subtypes of each brain tumor type. Additionally, future works could expand on the machine learning methods used in this study by implementing Support Vector Machines (SVM) or K-Nearest Neighbors (KNN) to improve the model accuracies.

Conclusion

In this study, the pretrained models ResNet18, ReNet50, VGG 16, DenseNet, GoogLeNet, ShuffleNet, and MobileNet were used to classify MRI images into the four classes of no tumor, pituitary tumors, meningioma, and glioma. These models were tested on a dataset using 3264 images. For each model, two methods of transfer learning were applied. In the first method, the convolutional base was frozen and only the classifier layer was trained for each model. In the second method, both the convolutional base and the classifier layer were trained. With method 1, DenseNet achieved the best accuracy at 91.58% followed by ResNet18 at 89.13%. The accuracy was improved by using method 2, with ResNet18 and ShuffleNet both achieving 97.86% accuracies. The results of this paper show that the transfer learning method can be used to accurately classify brain tumors into four classes using a small dataset of MRI images.

References

Bhuvaji, S., Kadam, A., Bhumkar, P., Dedge, S., & Kanchan, S. (2020, May 24). *Brain tumor classification (MRI)*. Kaggle. Retrieved February 10, 2023, from https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri?select=Test

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.90

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected Convolutional Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2017.243

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv*. https://doi.org/arXiv:1409.1556

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2015.7298594

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2018.00716