Evaluating Different Classifiers for Detecting Brain Tumors from MRI Scans Using Image Recognition

Kousthubh Veturi

Abstract

The diagnosis and treatment of brain tumors is a very difficult and critical process that requires timely action and the most accurate information.The goal is to test and select the best performing classifier to detect Brain Tumors and Types of Brain Tumors(Glioma, Pituitary, and Meningioma) from an MRI scan using Machine Learning Techniques. A well functioning model would intend to replace manual work in a physical medical environment, where trained medical professionals are not just expensive, but prone to human error, fatigue, and other unpredictable issues which could affect accuracy. By leveraging the power of machine learning, this study intends to advance the medical field by proposing a high-accuracy automated and reliable system of detecting brain tumors that performs better than the current manual diagnosis options available. Results from this project will be useful in the development of advanced technologies and applications which will integrate machine learning into the medical diagnosis process and reduce time consumption while increasing accuracy and reliability, and by extension improving healthcare quality.

0.1 Keywords:

Brain Tumors, Classification, VGG16, MRI Scans

1 Introduction

Cancer diagnosis and treatment is a field in which errors can often be damaging and dangerous. Currently, when manual diagnoses are used, the possibility of discrepancies and errors is high, with a study showing that out of 357 examined cases, 68.9% had at least one discrepancy[1]. More alarming was that the largest cause of discrepancies was faulty reasoning - not identifying an abnormality - amounting to 34.3%[1]. The second largest cause was satisfaction of search at 10.5%[1]. Satisfaction of search is when radiologists feel satisfied with what they have found. A hypothetical example is only identifying one tumor and not examining it to see if other tumors or points of concern exist because the single diagnosis is perceived as enough. In comparision, "overread" reports, reports done by specialised radiologists who often work at seperate facilities, had average discrepancy rates ranging from 2-6%[1].

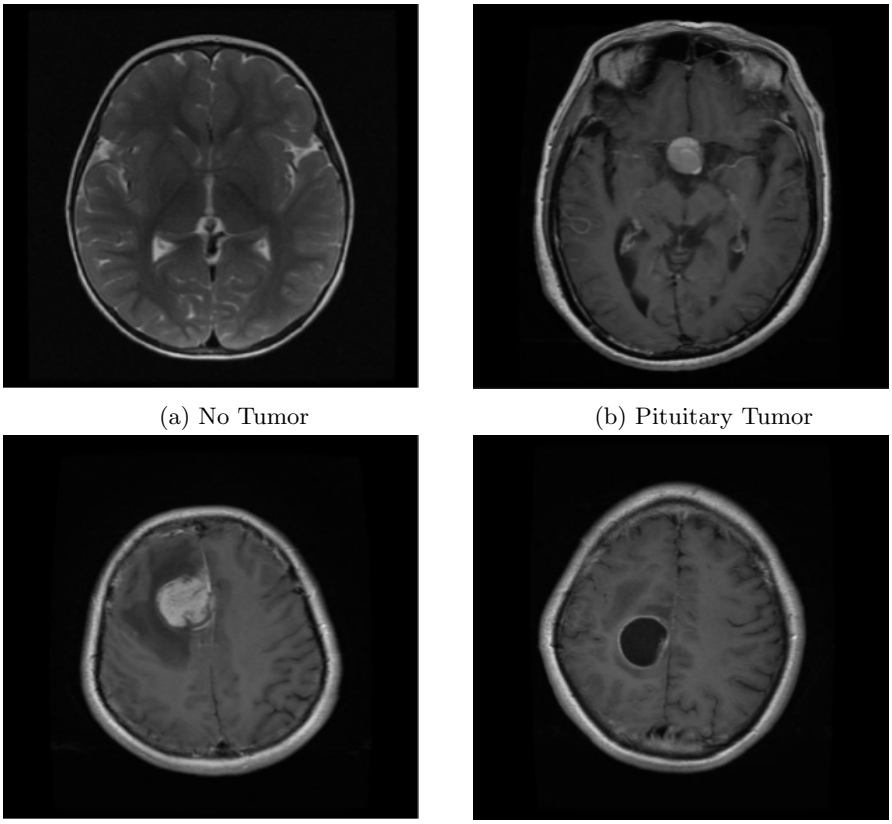
A rising alternative to manual diagnosis is diagnosis using Machine Learning. As it does not suffer from human problems such as lack of detail, the satisfaction of search, and fatigue, machine learning has the potential to play a primary role in the modernization and improvement of medical technology.

The goal of the investigation will be to assess the accuracy of machine learning-based classifiers at detecting whether brain MRI scans have tumors or not and to detect which type of tumor is present: meningioma, glioma, or pituitary tumor. This will be accomplished using processed feature vectors from a Convolutional Neural Network (CNN) which will then be used to train a classifier that will then be tested with other feature vectors that are grouped for testing. Various classifiers with different parameters and feature vectors from different stages of the CNN will be trained, and results will be analyzed to evaluate the best-performing classifier and whether machine learning-based methods of diagnosis outperform manual diagnosis systems. The findings of this study intend to provide more clarity and information required for the development of safer, more reliable, and more optimized medical technology.

2 Materials and Methods

2.1 Dataset

This project utilizes the "Brain Tumor MRI Dataset" by Masoud Nickparvar, sourced from Kaggle, to provide the images for classification[2]. This dataset contains a total of 7,022 images which are already split into training and testing sets of 5712 and 1311 images respectively. These images seem to be black-and-white scale at first glance, with there not being that much variation in color, resembling a grey-scale scan, which is what an MRI is intended to look like. The MRI scans are further split into four classes, "notumor" - no presence of a tumor, "pituitary" - a tumor detected in the pitutary gland of the brain, "meningioma"- tumor detected in the surrounding membrane meninges of the brain, and "glioma"- tumor in the glial cells of the brain[3][4].



(a) No Tumor (b) Pituitary Tumor

(c) Meningioma (d) Glioma

Figure 1: These are examples of each type of tumor diagnosis being tested.

2.2 VGG-16 Model

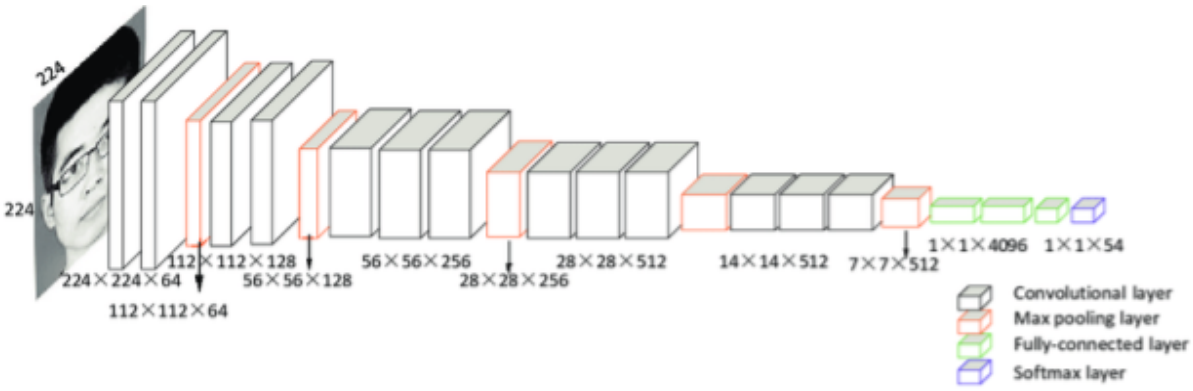


Figure 2: This is a diagram of the Convolutional VGG-16 Model and the respective dimensions of each layer[5]

This project utilized the VGG model as the Convolutional Neural Network for its relative simplicity as a Sequential model, meaning that the model continues in a fixed order every time without any

backtracks or loops[6]. This project used the VGG-16 layer model variant which was pre-trained in the Keras library. The usage of the VGG model was warranted in order to be able to generate the feature values that would then further be used to train and test classifiers of different types. The output feature array from the model was based on the layer that the features were extracted from.

2.3 Experimental Design

The method for classifying brain tumors from MRI scans involves 3 major parts: Generating Feature Vectors, Training the Classifiers, and Testing them.

For the project, feature vectors were extracted from the *block_5_conv3*, *fc2*, and *flatten* layers. None of these include the prediction layer, as the predictions for the 1000 classes of ImageNet are not helpful for this classification task because the goal is not to classify everyday objects. This task only needs feature vectors that can then be used for the classifiers. The reasons for selecting said layers to shave off the successive layers are as follows. Shaving off after the *block_5_conv3* layer gives features for after data has gone through only the convolution layers, *fc2* gives the vectors for data that has gone through everything except for the flattening and prediction layers, and *flatten* gives us the feature values that have not been converted into 1000 class predictions. In summary, these three layers are critical parts of the VGG-16 model, working as the last layers of each "block" of the VGG-16 model.

Classifiers were then trained using the feature vectors of images pre-dedicated to training. The classifiers were built using the sci-kit learn libarary For the K-NearestNeighbors classifiers, all classifiers starting from 1 all the way to 8 neighbors were trained for each category of feature vectors from different layers of the model. Logistic Regression classifiers were trained for all three layer types, and Decision Tree Classifiers from a depth value of 1 all the way to 8 were trained.

After this stage, each trained classifier was tested using the feature vectors dedicated to testing from the layer the classifier was trained on and the accuracy rates were noted.

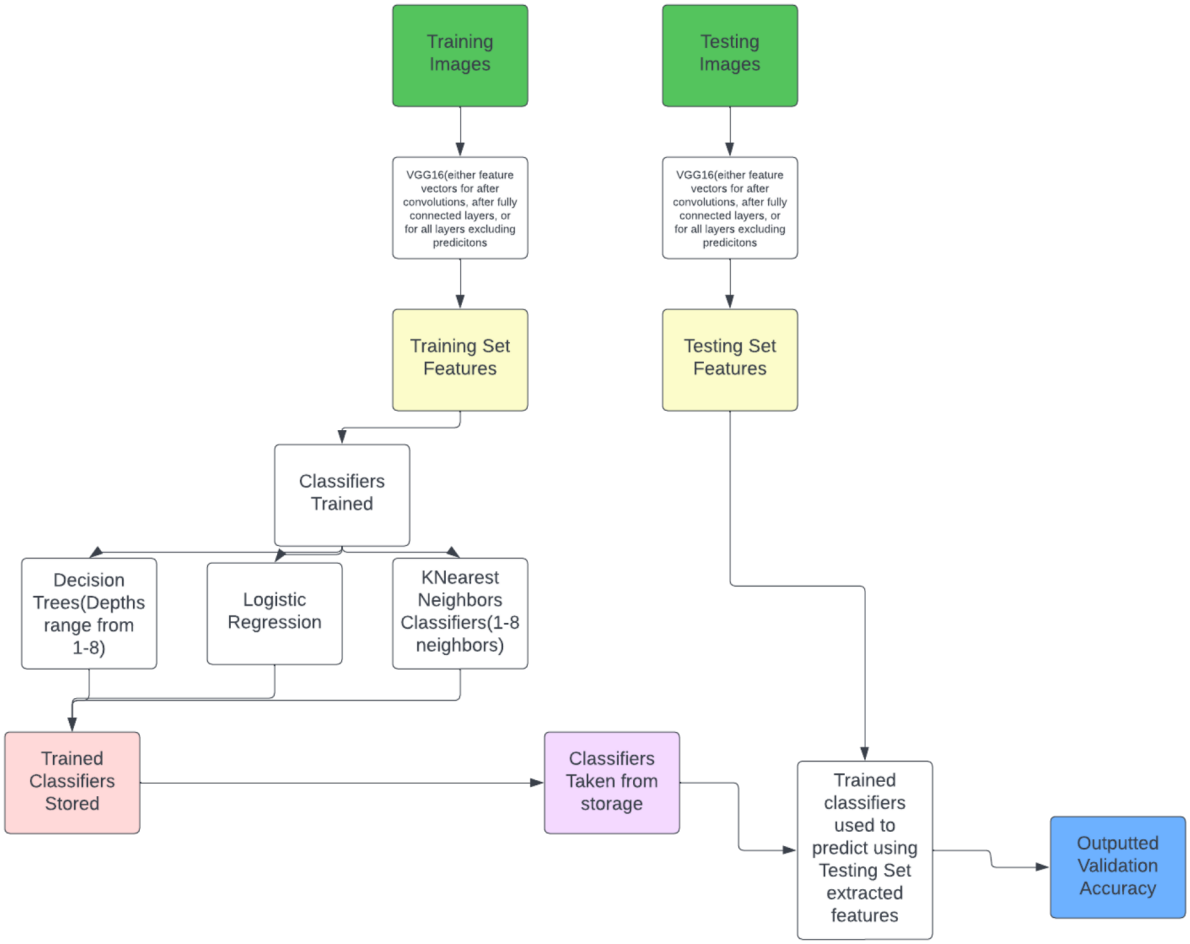
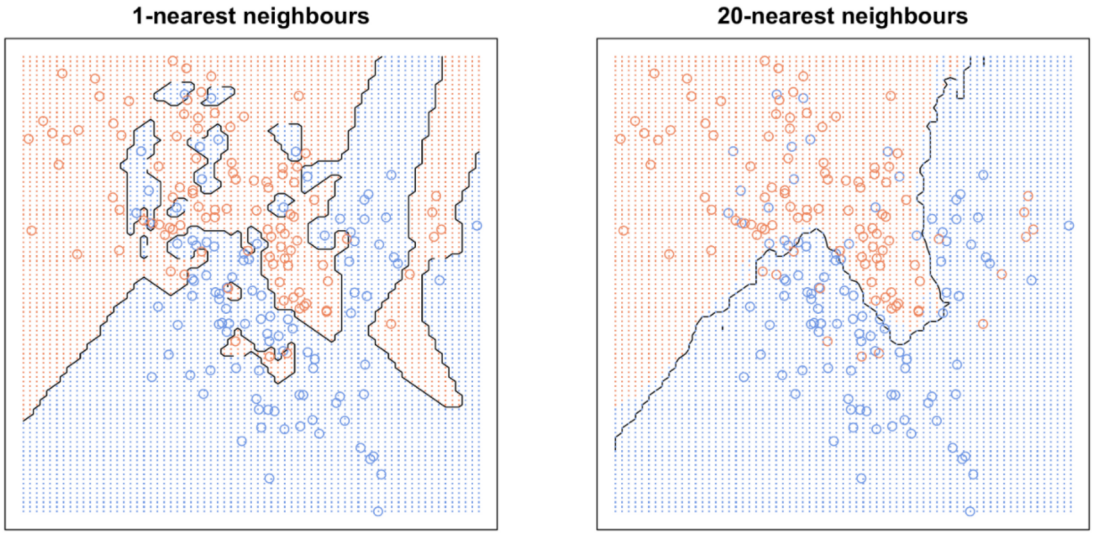


Figure 3: This is a pipeline diagram of the proposed design of the classification task

2.4 KNearest Neighbor Classifier

KNearestNeighbors is a supervised learning classifier which utilizes proximity and distance to make a prediction about a specific data point, hence the name "nearest neighbors". In the context of this task, there are four different classes, hence, there would be four different clusters on a plane. In this project, the number of neighbors as a value is set from 1 to 8 inclusive. The more number of neighbors there are, the more flexible the classifier is to outliers. Below is an diagram of how a KNearestNeighbors classifier works[7], and below the first image is a diagram displaying the effect of more neighbors[8].



Initially, values for distances from the input point to the other data points are calculated with Euclidean Distance, which returns a distance vector, calculated as follows[9]:

$$Distance_{euclidean}\,(x, y) = \sqrt{\sum_{i=1}^{n} (y_0 - x_0)^2} \tag{1}$$

In the equation, y_0 and x_0 are the input euclidean vectors that start from the origin. The variable n refers to the n-space, or the number of dimensions being analyzed and worked with. This euclidean distance is used to implement the idea of "plurality vote", that is, if the majority of close points have a specific label, then that point is given said label. If a point has a result of 35% class A, 25% class b, 20% class C, 15% class D, and 5% class E, the point is assigned class A because the plurality of close points are of class A.

Ultimately, advantages to this classifier include the relative simplicity of the construction of the classifier and the simplicity of the mathematics that make this classifier work.

2.5 Logistic Regression

Logistic Regression estimates the probability of a class based on the dataset given. Logistic Regression involves the application of a "logit transformation" on the probabilities[10]. Logistic Regression learns by calculating the sum of input features and calculates the logit as follows:

$$logit\,(Y) = naturallog\,(odds) = ln\left(\frac{\pi}{1-\pi}\right) \tag{2}$$

Since the project's task involves multiple classes, Multinomial Logistic Regression is utilized. Multinomial Logistic Regression uses softmax instead of sigmoid for the loss function. The softmax and sigmoid functions are as follows:

$$softmax\,(x_0) = \frac{e^{x_0}}{\sum_{i=1}^{n} e^{x_i}} \tag{3}$$

$$\widehat{multimodalprobability} = softmax\,(x_0) \tag{4}$$

For this project, the base Logistic Regression classifier in Scikit-learn is utilized for classification without any modification of parameters. The default parameters include defaulting to l2 penalties and multinomial as there are four classes being worked with.

2.6 Decision Trees Classifier

Decision Tree Classifiers are unsupervised classifiers that work based on a tree-like structure. There is a Primary node, from which internal nodes break off, which further break into leaf nodes. This classification task would involve four classes, so the leaf nodes would always be either no tumor detected or the three types of brain tumors. In order to find splits into various nodes, Decision Tree classifiers employ the usage of greedy searches, an algorithm that chooses the best option available at the moment. The decision for what to split into nodes is called an attribute and is calculated using the Gini Impurity. This is the probability of incorrectly classifying data points, calculated as:

$$Gini\,Impurity = 1 - \sum_{i=1}^{n} p_i^2 \tag{5}$$

, where p_i denotes the probability of each class, and n denotes the number of classes. Furthermore, an essential aspect of decision trees is the concept of depth. Each split into further nodes is one level of depth. For example, splitting a question into two answers is one level of depth. Depth is the distance of the longest path from a leaf to a primary node in a decision tree. The more depth there is, the longer the distance from the primary node to the farthest leaf node possible, highlighting the increasing complexity and, on paper, a higher classification accuracy. Below is a simple potential representation of a decision tree classifier in a non-technical scenario.

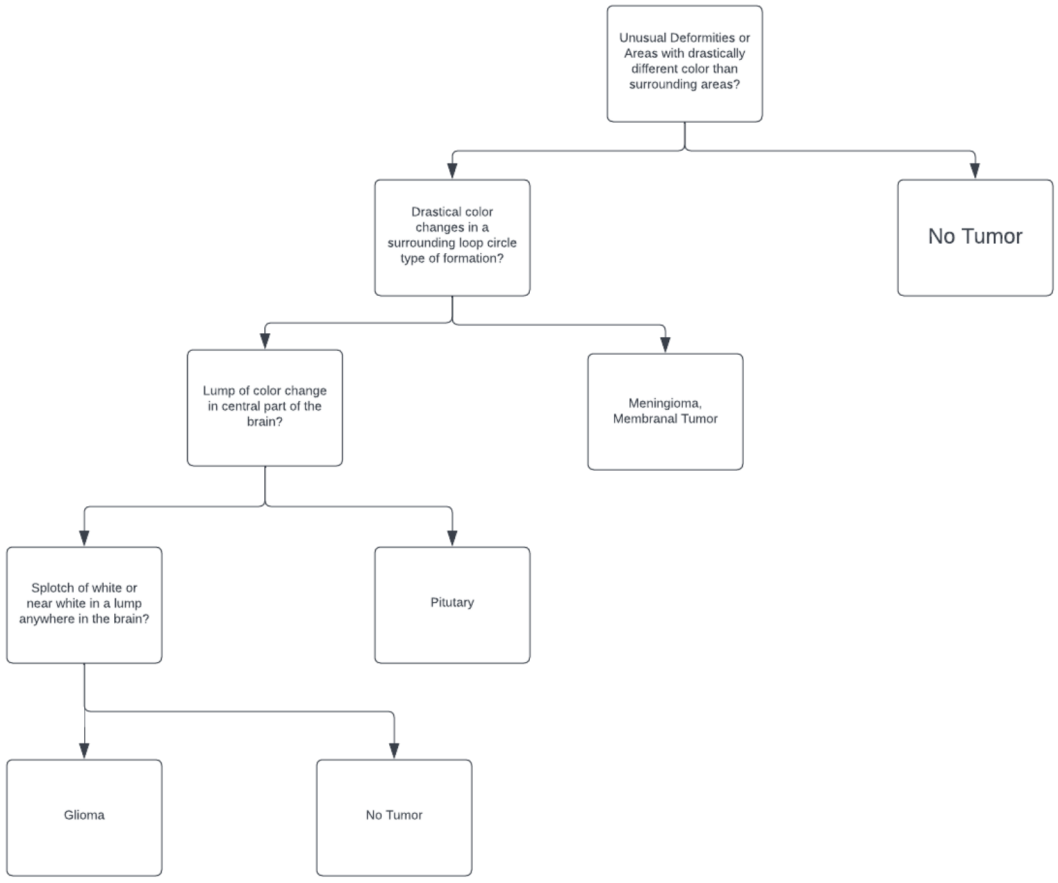


Figure 4: A hypothetical non technical scenario of a decision tree model

As visible, the more the depth, the more combinations of nodes that can result in selection and prediction of classes. The less the depth, the more limited the options are, and thus, the lower accuracy.

3 Results

3.1 KNearestNeighbor Classifier Results

The K-Nearest Neighbors classifiers often performed best with a lower parameter value for the number of neighbors. This is understandable, as earlier discussed; the increase in the number of neighbors results in the increased flexibility of including outliers, hence the lower accuracy.

Table 1: Accuracy Rates of Tested KNearestNeighbor Classifiers of Various Layers and Neighbor Parameter Values

Number of Neighbors	*fc2* Layer	*flatten* Layer	*block5_conv3* Layer
1	97.25%	98.39%	90.54%
2	92.98%	94.89%	87.71%
3	90.31%	93.44%	83.29%
4	89.70%	92.14%	81.99%
5	88.02%	90.77%	80.16%
6	87.26%	89.54%	79.17%
7	86.11%	88.86%	77.87%
8	86.95%	87.64%	77.95%

Observable is that the highest accuracies still do remain at the classifiers with only one neighbor, with the Highest Accuracy belonging to the KNearestNeighbor classifier that had one neighbor, trained with the features extracted from the flattened layer. Across all tested features, the highest accuracies are always in the one neighbor form of the classifier, which may be consistent with the more rigidness on the flexibility of inclusion of outliers. To further analyze the results confusion matrices were used. Below are confusion matrices for the three best, and the three worst classifiers:

Table 2: Matrices of best classifiers from each layer - Each number on the Y and X axis indicates the classes 0-3 in the order of: glioma, meningioma, notumor, pituitary

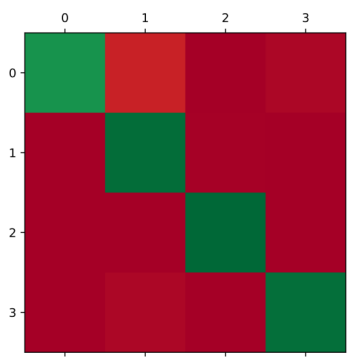
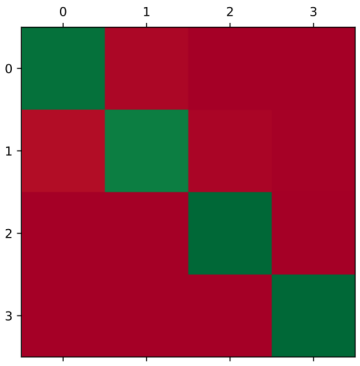
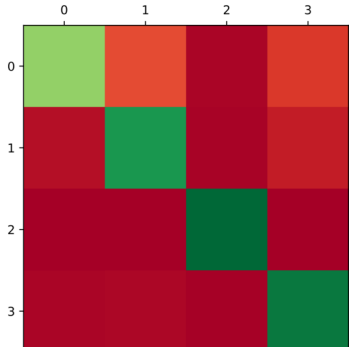
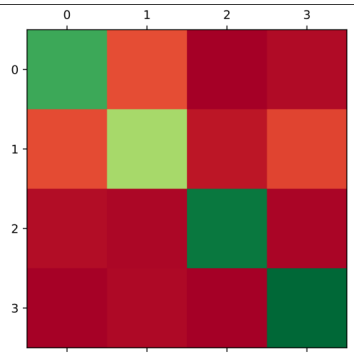
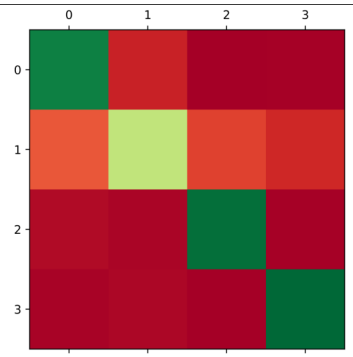
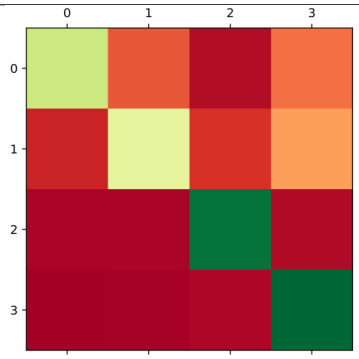
fc2 Layer	*flatten* Layer	*block5_conv3* Layer
		

Table 3: Matrices of lowest performing classifiers from each layer - Each number on the Y and X axis indicates the classes 0-3 in the order of: glioma, meningioma, notumor, pituitary

fc2 Layer	*flatten* Layer	*block5_conv3* Layer
		

3.2 Logistic Regression Results

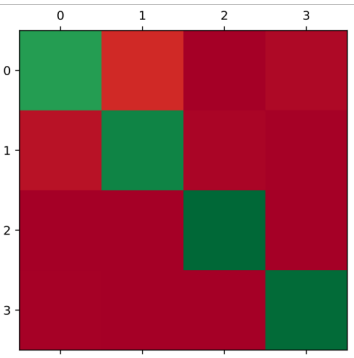
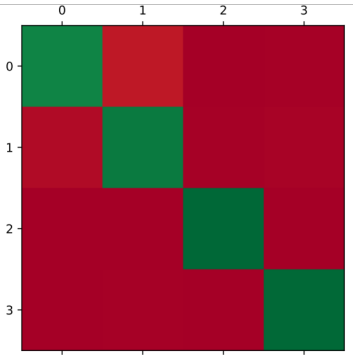
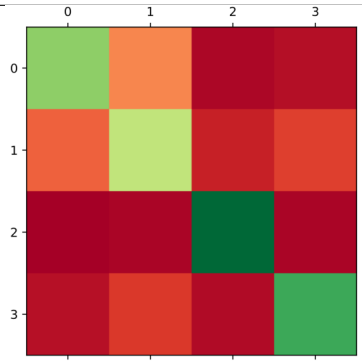
Since only the Multinomial Logistic Regression classifiers with L2 penalty were tested, there was only one classifier for each layer tested

Table 4: Performance of each Logistic Regression classifier

fc2 Layer	*flatten* Layer	*block5_conv3* Layer
95.80%	97.48%	78.94%

The classifier tested with the flatten layer's features performed the best, with the block5_conv3 layer's features testing the worst out of the tested features extracted. This may be because the earlier extractions do not have the features as developed as the flatten layer does because there are fewer trainable parameters in earlier layers(flatten layer is closest to the end of the model, and block5_conv3 is the farthest). These are the Confusion Matrices):

Table 5: Confusion Matrices for each Logistic Regression classifier from different layers

fc2 Layer	*flatten* Layer	*block5_conv3* Layer
		

3.3 Decision Tree Results

The Decision Tree Classifiers performed worse across all parameter depth values and chosen layers. The most plausible reason for this may be the reliance on the greedy algorithm which causes the implication that even one modification of a single node will require drastic changes in the classifier, hampering performance. This problem is exacerbated by the usage of in this case, multidimensional non-linear data in the form of features extracted from the VGG-16 model. In the case of using images, not all images are the same, thus different data points can sometimes result in less-than-optimal results.

Table 6: Accuracy Rates of Tested Decision Tree Classifiers of Various Layers and Depth Parameter Values

Depth	*fc2* Layer	*flatten* Layer	*block5_conv3* Layer
1	48.58%	42.02%	43.32%
2	58.81%	58.58%	50.72%
3	64.45%	66.06%	54.91%
4	67.96%	71.31%	58.35%
5	72.99%	72.92%	62.39%
6	76.04%	76.12%	62.40%
7	78.87%	78.71%	62.54%
8	79.17%	80.32%	65.67%

Observable is the overall increase in accuracy as depth increases. This is likely because a larger depth allows for the creation of more nodes, aiding in making the final classification decision. Below are confusion matrices for the three best ,and three worst classifiers:

Table 7: Matrices of best classifiers from each layer(all have depth of 8) - Each number on the Y and X axis indicates the classes 0-3 in the order of: glioma, meningioma, notumor, pituitary

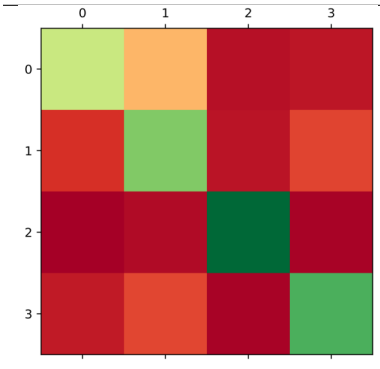
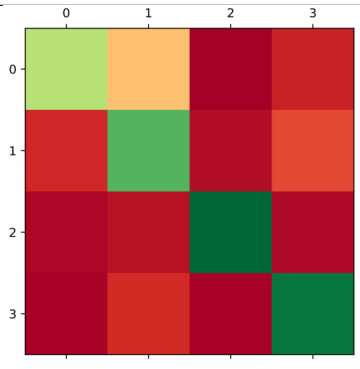
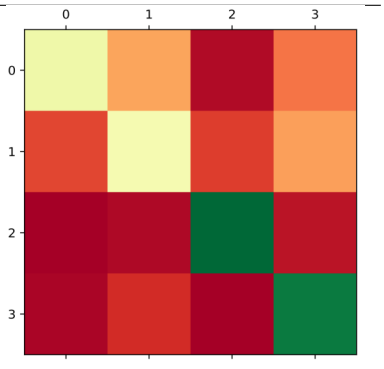
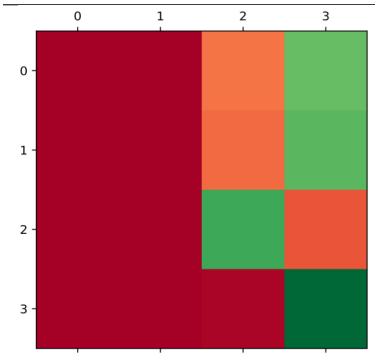
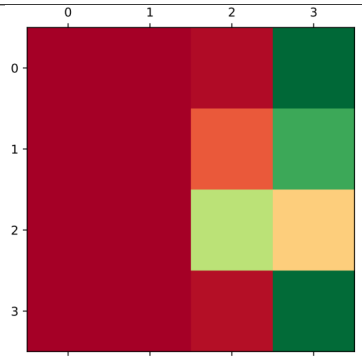
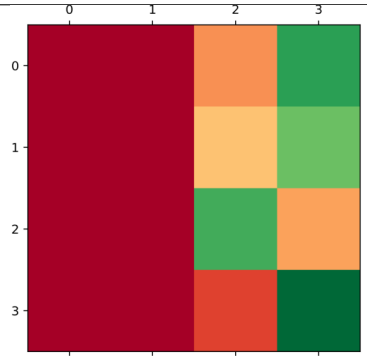
fc2 Layer	*flatten* Layer	*block5_conv3* Layer
		

Table 8: Matrices of lowest performing classifiers from each layer(all have depth of 1) - Each number on the Y and X axis indicates the classes 0-3 in the order of: glioma, meningioma, notumor, pituitary

fc2 Layer	*flatten* Layer	*block5_conv3* Layer
		

4 Discussion

4.1 Interpretation of Confusion Matrices

4.1.1 KNearestNeighbors

In both the best and worst performing confusion matrices, the classes of *notumor* and *pituitary tumors* have the highest classification accuracy. This can be explained by the characteristics of both classes. The presence of no tumor often signifies a normal brain, with no noticeable abnormalities visible. Pituitary Tumors occur in the pituitary gland of the brain, making the tumor more localised. In comparison, Meningioma occurs in the membrane surrounding the brain, resulting in a less localised tumor. Glioma originates in the glial cells, which surround neurons and protect them. Difficulties arise because glioma does not have clear defined borders or mass, making detection difficult[11]. Even with these in mind, the single-neighbor KNearestNeighbor classifiers perform exceptionally well in all categories, with accuracy rates ranging from 90.54% to 98.39%, with the best performing classifier being the single-neighbor classifier that uses the *flatten* layer.

4.1.2 LogisticRegression

The general trend in breakdown of accuracies of classes continues to be observed in the matrices for the Logistic Regression classifiers. The glioma and meningioma tumors continue to have slightly lower accuracies. The *block_5conv3* layer classifier performs the worse, likely because this layer is earlier in the VGG-16 model and has less parameters to be trained on. The best performing classifiers were the ones which used feature vectors from the later layers of the model *fc2* and *flatten*.

4.1.3 Decision Trees

On average Decision Tree classifiers performed the poorest when tested. Observable in the 3 best classifiers is the recurrent issue with the harder-to-distinguish difference in the classification menin-

gioma and glioma due to the fundamental lack of definite mass and localization in these types of cancers. The single depth classifiers, each one from each layer performed the worst. The reason may be that the depth is simply too low for the decision tree to be able to make a calculated decision to classify the samples. The number of nodes is much lower than in the top performing decision tree classifiers, resulting in a lower performance. Confusion Matrices for the single-depth decision tree classifiers show a complete red area on the left of each diagram, showing that the classifier is simply unable to classify the glioma and meningioma types at a high accuracy rate, while no tumor and pituitary tumors have marginally higher accuracy rates due to localised tumor locations which are easier to capture.

4.2 Conclusion

Overall, on average, the single-neighbor KNearestNeighbor classifiers performed best, with a maximum accuracy of 98.39%. A consistent trend in the KNearestNeighbor classifiers showed that a higher value for neighbors resulted in lower accuracy. This is likely because classifiers with more neighbors are prone to facing overfitting. Logistic Regression performed with accuracies above 90% for the fc2 and flatten layers located at the end of the model, likely because the number of usable parameters is higher at the end layers of a model. Additionally, the L2 regulariser likely prevented the overfitting of the model. Decision Tree Classifiers performed the worst due to the non-linearity of the data and the fundamental structure of Decision Trees not being flexible for a broader range of changes in data that are present in MRI scans. Thus, the best-performing classifier was the Single-Neighbor KNearestNeighbors classifier, which used feature vectors from the flatten layer of the VGG16 mode and had a 98.39% accuracy rate.

Compared to the rate of discrepancies found in the manual diagnosis of 68.9%, the model and classifier performed better with an accuracy rate of 98.39% and an error rate of 1.61%. This is likely because the model does not suffer from human error, fatigue, and other conditions present in humans. However, compared to overread MRI scans, the model and classifier perform only slightly better than the overread error rates of 2-6%. Thus, although only slightly better than overread scans of MRI images, the model and classifier perform better than the present manual types of diagnosis.

Future steps will be to research further and develop an application using AI for diagnosis that can be implemented in medical technologies and distributed to hospitals worldwide. This model is a step towards ensuring equal access to quality medical resources because it can be deployed in various regions, including regions with less access to medical resources and specialists, helping thousands of people worldwide. The benefit of this is that thousands of people worldwide will have their treatments done with more reliable information, improving patient experience and the odds of recovery from conditions. Additionally, this model and classifier design can be used to diagnose various conditions, not just limited to brain tumors, by simply re-training the model and generating feature vectors using different images and scans. In conclusion, the model has demonstrated a higher accuracy rate at 98.39%, performing better than manual diagnoses, shows potential to be utilized for various diagnoses, and can be deployed in various regions worldwide, closing the medical gap and modernizing the medical field.

Acknowledgments

The author would like to thank Dr.Mason McGill for serving as an advisor on this project.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publishing of this paper

Data Availability and Supplementary Materials

The data used for this research paper is accessible on Kaggle at this link:
kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset
This data contains all MRI scan images that were used for the purpose of this paper.

The folder with all the code can be accessed at:
https://drive.google.com/drive/folders/1ErnH_sZvve5DnPI752mRPlV04PFkdQfY?usp=sharing

References

[1] Danielle E. Kostrubiak et al. "Body MRI Subspecialty Reinterpretations at a Tertiary Care Center: Discrepancy Rates and Error Types". In: *American Journal of Roentgenology* 215.6 (2020). PMID: 33052740, pp. 1384–1388. DOI: 10.2214/AJR.20.22797. eprint: https://doi.org/10.2214/AJR.20.22797. URL: https://doi.org/10.2214/AJR.20.22797.

[2] Msoud Nickparvar. *Brain Tumor MRI Dataset*. 2021. DOI: 10.34740/KAGGLE/DSV/2645886. URL: https://www.kaggle.com/dsv/2645886.

[3] B M Arafah and M P Nasrallah. "Pituitary tumors: pathophysiology, clinical manifestations and management." In: *Endocrine-related cancer Endocr Relat Cancer Endocr. Relat. Cancer* 8.4 (2001), pp. 287–305. URL: https://erc.bioscientifica.com/view/journals/erc/8/4/11733226.xml.

[4] Christian Ogasawara, Brandon D. Philbrick, and D. Cory Adamson. "Meningioma: A Review of Epidemiology, Pathology, Diagnosis, Treatment, and Future Directions". In: *Biomedicines* 9.3 (2021). ISSN: 2227-9059. DOI: 10.3390/biomedicines9030319. URL: https://www.mdpi.com/2227-9059/9/3/319.

[5] Zhao Pei et al. "Face Recognition via Deep Learning Using Data Augmentation Based on Orthogonal Experiments". In: *Electronics* 8 (Sept. 2019), p. 1088. DOI: 10.3390/electronics8101088.

[6] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[7] *What is the k-nearest neighbors algorithm? — IBM — ibm.com*. https://www.ibm.com/topics/knn.

[8] Dhawan. *KNN(K-nearest neighbour) algorithm, maths behind it and how to find the best value for K*. Oct. 2019. URL: https://medium.com/@rdhawan201455/knn-k-nearest-neighbour-algorithm-maths-behind-it-and-how-to-find-the-best-value-for-k-6ff5b0955e3d.

[9] Pá draig Cunningham and Sarah Jane Delany. “k-Nearest Neighbour Classifiers - A Tutorial”. In: *ACM Computing Surveys* 54.6 (July 2021), pp. 1–25. DOI: 10.1145/3459665. URL: https://doi.org/10.1145%2F3459665.

[10] Joanne Peng, Kuk Lee, and Gary Ingersoll. “An Introduction to Logistic Regression Analysis and Reporting”. In: *Journal of Educational Research - J EDUC RES* 96 (Sept. 2002), pp. 3–14. DOI: 10.1080/00220670209598786.

[11] Makoto Hishii, Toshiharu Matsumoto, and Hajime Arai. “Diagnosis and treatment of early-stage glioblastoma”. In: *Asian Journal of Neurosurgery* 14.02 (2019), pp. 589–592. DOI: 10.4103/ajns.ajns_18_19.