Research Paper: Brain Tumor Classification and Segmentation with baseline models, CNNs, and Transfer Learning

Joanna Andrews

Abstract:

For this study, I utilized the Brain Tumor MRI dataset with the goal of developing a classifier capable of classifying MRI Images as no tumor, pituitary tumor, meningioma, or glioma with Machine Learning. The dataset included 7,022 images in total, and was split into 5,712 training images and 1,311 testing images. The training images included 1,321 glioma images, 1,339 meningioma images, 1,457 pituitary images, and 1,595 non-tumor images. The testing images included 300 glioma images, 306 meningioma images, 300 pituitary tumor images, and 405 non-tumor images. I first experimented with different baseline classifiers (including KNN, Decision Tree, Random Forest Classifier, Decision Tree, Support Vector Machine, Ridge Classifier, and Gradient Boosting Machine), then moved on to CNN models with different epoch levels and a batch sizes of 100. To evaluate the accuracy of each classifier, I calculated the accuracy value for each classifier. The classifier(s) which performed the best based on these metrics was the CNN classifier, which had an accuracy value of . With this classifier, I generated a saliency map to highlight the features most integral to the classifier's decisions. In doing this, I discovered that despite the relatively high performance of CNN models for the classification task, the primary features used by my CNN model to make its predictions were outside of the brain. This study elucidated the fact that even well-performing machine models could be seemingly misleading, highlighting both the importance of understanding how Machine Learning models make their decisions as well as a potential concern that must be considered while implementing Machine Learning models in the medical imaging space. This study highlights the importance of evaluating and interpreting even high-performing models before implementing them as reliable tools for medical care.

Introduction:

A brain tumor arises from irregular growth of cells within the brain. Brain tumors can be either malignant (cancerous) or benign (non-cancerous). In 2020, Primary cancerous brain and CNS tumors accounted for approximately 251,329 deaths worldwide. [1] For both malignant and benign tumors, WHO has developed a grading system (1-4) to classify brain tumors by their level of growth, with 1 being assigned to the tumors showing the least rapid growth and 4 being assigned showing the most rapid growth. Tumor grade assignments are informed by the type of tumor, natural history (progression of the cancer without treatment), or biological markers. [2] Brain tumors are also classified by stage (1-4), which is an indicator of how much the tumor has spread throughout the body. When brain tumors are diagnosed at an earlier stage, they are more treatable and patient survival outcomes are

better. However, brain tumor diagnosis can be costly, invasive, and time-consuming. Biopsies, considered one of the most reliable diagnostic methods for brain tumors, require the extraction of brain tissue so that the tissue can be examined for its histological features. These features provide valuable information about the tumors and can be used to determine tumor malignancy, grade, and stage. However, the biopsy process is both invasive and costly, and it can be time-consuming because pathologists must examine the extracted brain tissue. [2] The less invasive diagnostic methods are CT scan and MRI imaging, which can provide faster and safer insight into brain tumors. However, since these images must be interpreted by radiologists, there is room for potential error or variability in the conclusions obtained from these images, [3] and it can take time for radiologists to interpret the images.

Meningioma, among the most common types of brain tumor, originates within the meninges – a series of 3 layers coating the outside of the brain and spinal cord. Meningioma tumors are most commonly found in regions containing high concentrations of arachnoid villi, which are small protruding structures located in the 2nd layer of the menges. Pituitary tumors, on the other hand, impact the functions of the pituitary, which is a small gland in the brain which is responsible for releasing regulatory hormones. Therefore, these types of tumors can interfere with the production of these hormones, as well as the processes regulated by these hormones. Although these tumors develop slowly, they have the potential to cause significant damage to a patient. Gliomas are another type of brain tumor. Under normal conditions, glial cells protect and maintain the neurons, keeping them insulated and supplied with food so that they can continue to function well. Examples of glial cells include astrocytes, which help repair damaged nerves and provide nutrients to the nerves, and oligodendrocytes, which produce a protective coating for neurons known as a myelin sheath. Gliomas form when tumors develop from a type of glial cell. Different tumor types are characterized by different locations and histologies, and typically a trained medical professional determines tumor type using observations from tumor tissue extracted through a biopsy or through identifying tumor features from medical imaging. In this study, we attempt to develop a model capable of identifying the unique features of these different tumor types to successfully categorize the MRI brain images.

Materials and Methods:

For my experiment, I analyzed MRI images, which are used to visualize the location of tumors and other structures in the brain to aid treatment planning. MRI images are generated by creating a magnetic field within the brain using large magnets. The protons in the brain are drawn to the magnets and shifted out of place using radio waves. When the radio waves are cut off, the protons move back towards the magnets, sending out individual radio signals which can be built into an image displaying the brain's structures. [4] my dataset included 7,022 MRI images obtained from figshare (linked here) and from the Br35H study (linked here). The dataset included 5,712 training images and 1,311 testing images. Both the train and test images were divided into 4 classes: no tumor, glioma, meningioma, and

pituitary, indicating whether a tumor was present in the image and what type of tumor (if any) was present. The training images included 1,321 glioma images, 1,339 meningioma images, 1,457 pituitary images, and 1,595 non-tumor images. The testing images included 300 glioma images, 306 meningioma images, 300 pituitary tumor images, and 405 non-tumor images. Each image, such as the images displayed in figure 1, had dimensions of 256 by 256 pixels. As you can see from the images provided in Figure 1, the images featured a variety of angles/portions of the brain. This was done so that the model would be able to identify features based on a more diverse dataset representing multiple views of the brain.

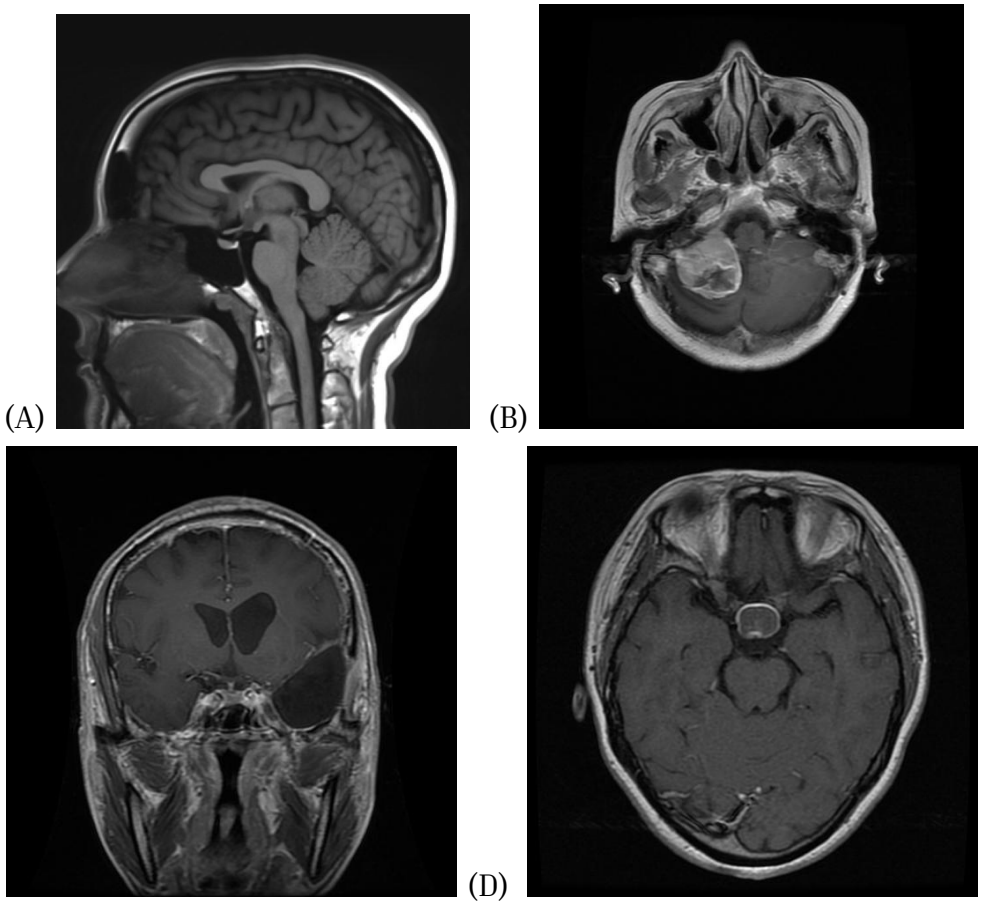


Figure 1. Example brain MRI images. (A) example image of no tumor. (B) Example image of meningioma (C) Example of glioma (D) Example of pituitary

Preprocessing

To normalize the images, I created a function to transform each pixel value by subtracting the minimum pixel value and dividing this difference by the range in pixel values. To load and process my input images and labels, I created image data generators for both the Training and Testing images. These generators loaded the images, normalized the images using my custom normalization function, and encoded the labels. I converted the

images to grayscale and converted my image labels from their one-hot-encoded format to a categorical format including the numerical category of each input image, where 1, 2, 3, and 4 corresponded with the glioma, meningioma, pituitary, and no tumor labels, respectively. After processing both the images and the labels, I then began building different models for image classification. For the image data generator, I trained the models with a batch size of 100.

Baseline Models

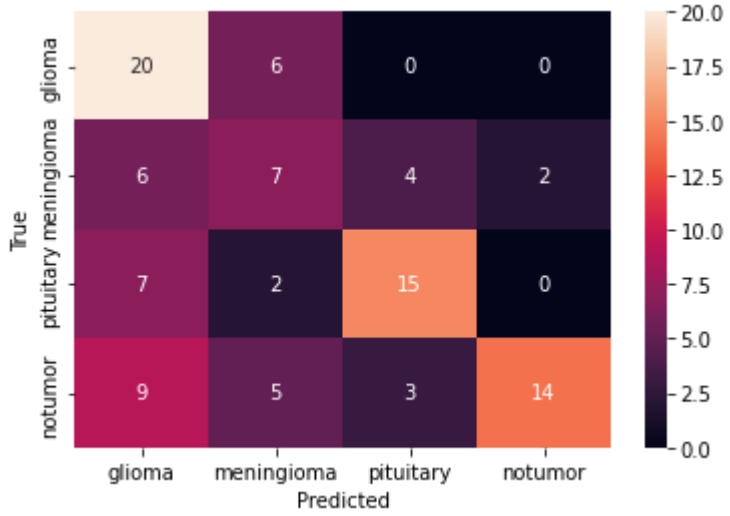
First, I trained some baseline models to assess their classification accuracies on the MRI brain image dataset. The baseline models I selected were: Logistic Regression, KNN, Decision Tree Regression, Random Forest Regression, Support Vector Machine, Gradient Boosting, and Ridge. To evaluate these models, I used the accuracy score and confusion matrices as metrics. The baseline model with the highest performance overall was the Random Forest Classifier with an accuracy value of .71. For each baseline classifier, I created a GridSearch function to test different parameters and identify the parameters which produced the highest model accuracy. To evaluate the accuracy of my models, I used the accuracy score, which is simply the fraction of total predictions which are correct. This value is obtained by taking the sum of the true positives and true negatives (the 'correct' predictions) over the total number of predictions, or the sum of the true and false positives and the true and false negatives. Here are the results of the hyperparameter tuning for these models and the highest performance overall for each model.

Model	Parameters Tested	Best Parameters	Accuracy with optimal parameters
KNN Classifier	n_neighbors: 1-10	n_neighbors = 1	.66
Logistic Regression Classifier	C: 1-10	C = 3	.74
Support Vector Machine Classifier	kernel: linear, polynomial, rbf C: 1-9	C = 1 Kernel = linear	.72
Gradient Boosting Classifier	random_state: 1,3,5 Random_state: 10,20	Loss: deviance n_estimators: 20 random_state: 3	.64
Decision Tree Classifier	Criterion: entropy, gini Splitter: best,	Criterion: entropy Max depth: 4 Splitter: best	.65

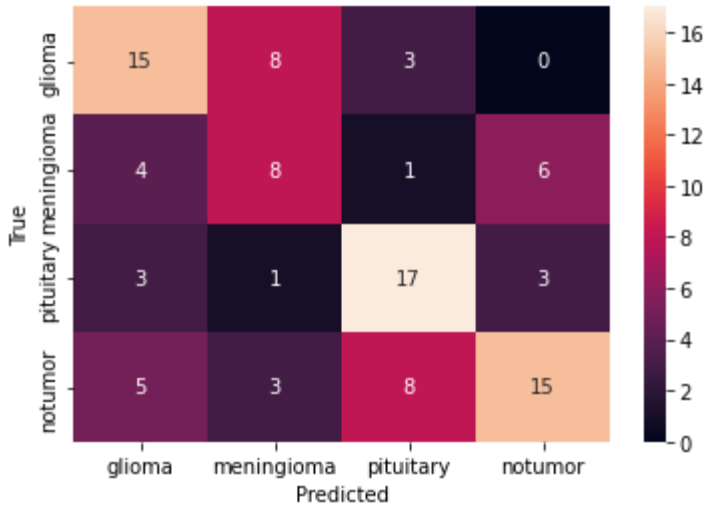
	random Max_Depth: 2,4,6		
Random Forest Classifier	N_estimators: 10-120 (increasing by 10) Criterion: gini, entropy Max_depth: 1-5, inclusive	Criterion: gini Max_depth: 2 N_estimators: 90	.77
Ridge Classifier	Alpha: 0, .25, .75, 1	Alpha: 0	.67

Confusion Matrices for Baseline Models

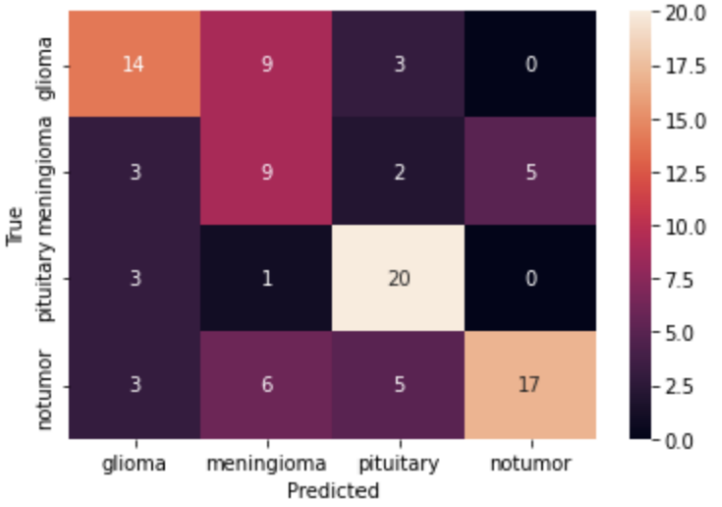
To further evaluate and to better understand my baseline models, I generated a confusion matrix for each model. This confusion matrix is essentially a grid in which the x-axis represents the tumor class predicted by the model and the y-axis represents the actual tumor class of the input image. Sections of the graph where both the predicted and true tumor class are indicative of correct predictions made by the model, and the larger the values in these sections, the more accurate the model is. According to our confusion matrices, our models were successful overall in distinguishing tumor from no tumor images, but they struggled to pick up differences between individual tumor types. In particular, there appeared to be large numbers of meningioma images misidentified as glioma cases by these models.



(A)


glioma
meningioma
pituitary
notumor
True
15 8 3 0
4 8 1 6
3 1 17 3
5 3 8 15
glioma
meningioma
pituitary
notumor
Predicted
16
14
12
10
8
6
4
2
0

(B)


glioma
meningioma
pituitary
notumor
True
14 9 3 0
3 9 2 5
3 1 20 0
3 6 5 17
glioma
meningioma
pituitary
notumor
Predicted
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

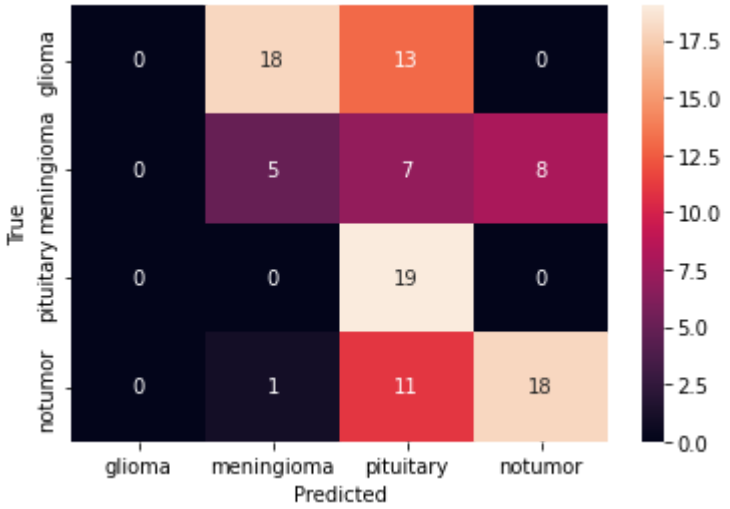
(C)


glioma
meningioma
pituitary
notumor
True
20 1 3 2
4 7 4 5
0 7 12 3
1 2 4 25
glioma
meningioma
pituitary
notumor
Predicted
25
20
15
10
5
0

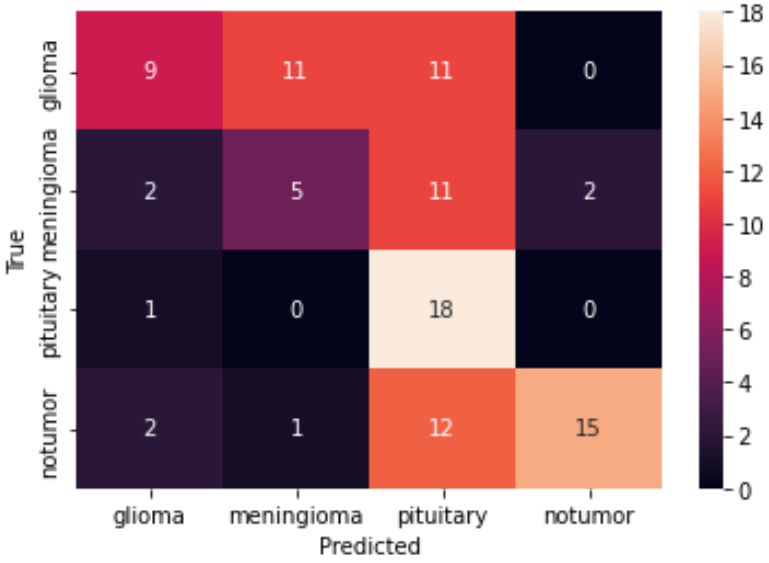
(D)


glioma
meningioma
pituitary
notumor
True
11 6 12 2
4 4 7 5
2 1 14 2
3 4 12 11
glioma
meningioma
pituitary
notumor
Predicted
14
12
10
8
6
4
2

(E)


glioma
meningioma
pituitary
notumor
True
0 18 13 0
0 5 7 8
0 0 19 0
0 1 11 18
glioma
meningioma
pituitary
notumor
Predicted
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

(F)


glioma
meningioma
pituitary
notumor
True
9 11 11 0
2 5 11 2
1 0 18 0
2 1 12 15
glioma
meningioma
pituitary
notumor
Predicted
18
16
14
12
10
8
6
4
2
0

(G)

Figure 2. Confusion Matrix Results for Models Trained with Optimal Parameters. (A) KNN Classifier (B) Logistic Regression Classifier (C) SVM (D) Gradient Boosting Classifier (E) Decision Tree Classifier (F) Random Forest Classifier (G) Ridge Classifier

CNNs:

Convolutional Neural Networks, a complex type of model often used for image analysis, include an input layer (where the images are received), several intermediate 'hidden' layers (where important features of the image are identified), and an output layer, which generates the classification results. In the intermediate layers, a grid of weights is passed over the image, and as it is applied to different parts of the image, the model identifies features of the image which are most prevalent to the classification task and adjusts its weights accordingly. There are several types of layers, and different combinations of layers, or different architectures, can yield different results. I experimented with several different CNN architectures and evaluated the accuracy for each model, noting that the highest accuracy was achieved with the model architecture of 1 convolutional layer, followed by 2 pooling layers, 2 more convolutional layers, another pooling layer, and a dense layer. The table below illustrates the architecture of the CNN model which achieved the greatest accuracy score of .9619.

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_11 (Conv2D)          (None, 256, 256, 32)      896

 max_pooling2d_9 (MaxPooling  (None, 128, 128, 32)     0
 2D)

 max_pooling2d_10 (MaxPoolin  (None, 64, 64, 32)       0
 g2D)

 conv2d_12 (Conv2D)          (None, 64, 64, 128)       36992

 conv2d_13 (Conv2D)          (None, 64, 64, 256)       295168

 max_pooling2d_11 (MaxPoolin  (None, 32, 32, 256)      0
 g2D)

 flatten_3 (Flatten)         (None, 262144)            0

 dense_3 (Dense)             (None, 4)                 1048580

=================================================================
Total params: 1,381,636
Trainable params: 1,381,636
Non-trainable params: 0
```

Figure 3. Model Architecture of the CNN with the highest accuracy score. Summary of model architecture: 1 Convolutional 2D layer, followed by 2 Max Pooling 2D layers, 2 more Convolutional 2D layers, another Max Pooling 2D layer, a flattening layer to convert the 2D grid of trained image features to a 1D vector, and a Dense layer

Saliency Map: CNN

Saliency maps are a method of interpreting a model, or understanding how a model makes its decisions. Saliency maps essentially manipulate features individually and calculate how changing that singular feature impacts the loss (or the gradient) of the model. Features of the image which cause the biggest change in loss when manipulated are identified as being most integral to the model's decisions and are highlighted by the image. The figure below illustrates saliency maps generated using my best CNN model. As illustrated by the saliency maps, the model appears to be basing its decision primarily using regions outside the brain. This is an interesting finding because it would seem counterintuitive for a model to perform well in classifying brain tumor MRI images while using features which seem unrelated to the brain tumors themselves.

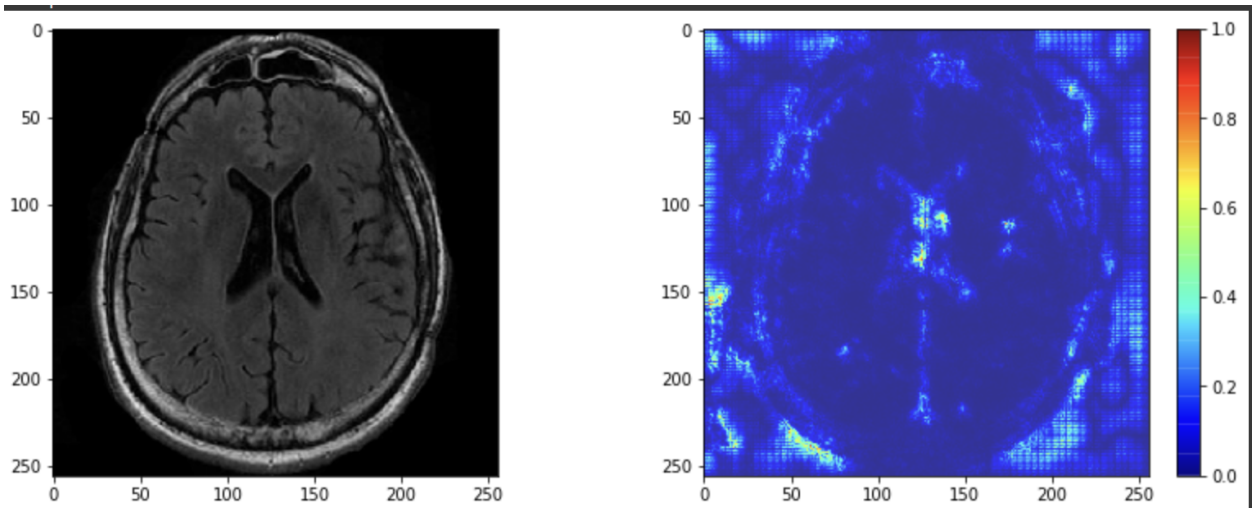


Figure 5. This figure shows a saliency map generated using the trained CNN model with the highest accuracy. The parts of the image with the highest weights (meaning that the model identified these features as the most important for classifying the images) are identified by warmer colors. As you can see, the warmest colors in this plot are slightly concentrated in the center of the image and are more prominently concentrated around the borders and edges of this image, away from the brain. This suggests that the model is largely reliant on features not included in the brain when making its decisions, although it may pick up on some key features within the brain.

Results:

Out of all the baseline models, the classifier which achieved the highest accuracy with its optimized parameters was the Random Forest classifier, with an accuracy of .77. The CNN model which demonstrated the greatest accuracy achieved an accuracy of .9619. The confusion matrices revealed that overall the models frequently misclassified meningioma

and glioma, often interchanging them. The saliency maps demonstrated a discrepancy between the model's high performance and its concerningly high use of features outside the brain to achieve this.

	Evaluation Metrics	Results/Observations
Baseline Models	Accuracy Score, Confusion Matrices	• Highest accuracy: Random Forest model (accuracy = .77) • Most models struggled to distinguish between meningioma & glioma
CNNs	Accuracy Score, Saliency Maps	• Relatively high accuracies (highest = .9619) • Saliency maps revealed high reliance on image features outside the brain

Discussion:

In this experiment, I developed several types of Machine Learning models with the goal of successfully classifying MRI images by tumor type, including several baseline classifiers and CNNs. The highest performing baseline model was a Random Forest Classifier with a .77 accuracy score, while the highest performing CNN model had an accuracy of .9619.

The confusion matrix results revealed that the baseline models frequently mistook meningioma for glioma and that the Decision Tree Classifier, Random Forest Classifier, and Ridge Classifier models often classified pituitary images as glioma or no tumor images. This suggests that these tumor types share similar histological features and that distinguishing between these types could present a unique challenge for Machine Learning models. Additionally, the models were overall highly successful in correctly identifying no tumor images, which suggests that while the models struggled to distinguish between tumor types, the model demonstrated more success in distinguishing whether a tumor was present.

Although my CNN models showed higher accuracies overall, upon generating saliency maps for my most accurate CNN I discovered that my CNN model was heavily reliant on features not included within the brain. Thus, I can conclude that while deep learning models hold potential to be valuable tools for interpreting medical images, it is important to consider how the models are making decisions before accepting these models as a success based on high accuracy alone. By investigating models to determine whether they utilize medically relevant features, healthcare providers and medical professionals can be more

confident that the models they choose to implement in the medical sphere are truly useful and accurate tools.

Citations

1) American Cancer Society. (2022). *Cancer Facts & Figures 2022| American Cancer Society*. Www.cancer.org. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2022.html

2) S Tandel, G., Biswas, M., G Kakde, O., Tiwari, A., S Suri, H., Turk, M., Laird, J. R., Asare, C. K., A Ankrah, A., N Khanna, N., K Madhusudhan, B., Saba, L., & Suri, J. S. (2019). A Review on a Deep Learning Perspective in Brain Cancer Classification. *Cancers*, *11*(1), 111. https://doi.org/10.3390/cancers11010111

3) Hayward, R. M., Patronas, N., Baker, E. H., Vézina, G., Albert, P. S., & Warren, K. E. (2008). Inter-observer variability in the measurement of diffuse intrinsic pontine gliomas. *Journal of neuro-oncology*, *90*(1), 57–61. https://doi.org/10.1007/s11060-008-9631-4

4) NHS. (2019). *Overview - MRI scan*. Nhs. https://www.nhs.uk/conditions/mri-scan/