

Capstone 3 Documentation

Business Problem: Diagnosing brain tumors is a challenging task and requires a professional to diagnose and classify the type of brain tumor. It is difficult for high volumes of scans to be looked at and human error can reduce the accuracy of diagnostics. We need to create a model that correctly classifies the different types of brain tumors with high accuracy in order to reduce false positives due to human error, reduce the cost of diagnostics, and to be able to handle higher volumes of brain scans. In this problem, we will specifically work on categorizing normal brains, pituitary tumors, glioma tumors, and meningioma tumors. The tumor classes are balanced but the normal class has less samples.

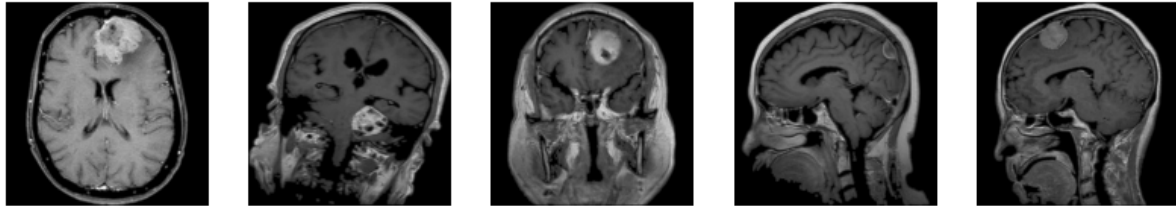
Constraints: There are about 1000 photos for each of the tumor classes and about 400 photos for normal brain scans. This is likely enough data to get reasonable results but more data may be needed for excellent results. Each of the classes also has several different angles of the tumor which may be confusing for the model and hurt performance. There is also only data for 3 tumor types, so this model would not be an exhaustive resource for diagnosing brain tumors but would only be useful for diagnosing the tumor types given in the data.

Stakeholders/benefited parties: This model would be used in medical facilities such as hospitals, clinics, physicians offices and perhaps other facilities. The model could also benefit people without health care who may not be able to afford getting checked and perhaps a diagnostic model that performs well enough could help those people to make more informed and data driven decisions about their health.



Samples of each tumor

meningioma_tumor meningioma_tumor meningioma_tumor meningioma_tumor meningioma_tumor



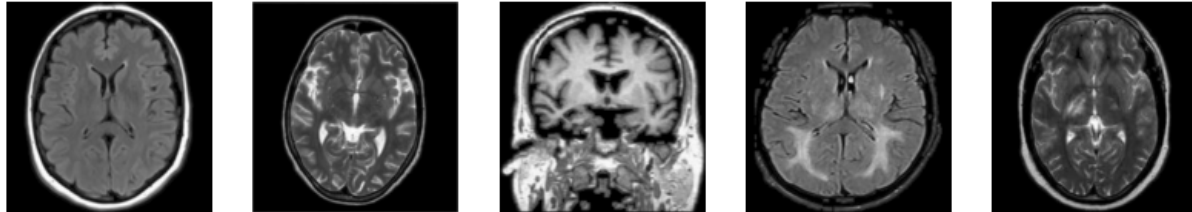
normal

normal

normal

normal

normal



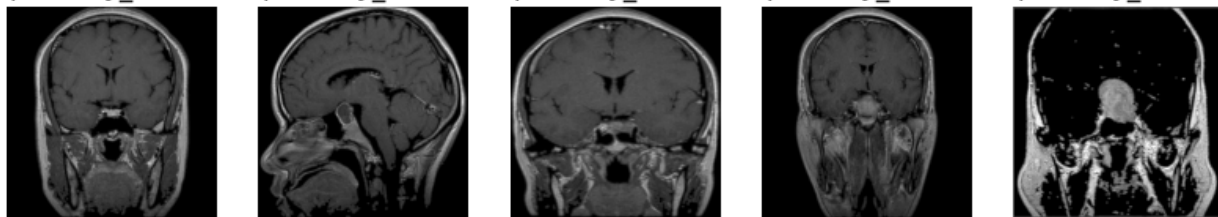
pituitary_tumor

pituitary_tumor

pituitary_tumor

pituitary_tumor

pituitary_tumor



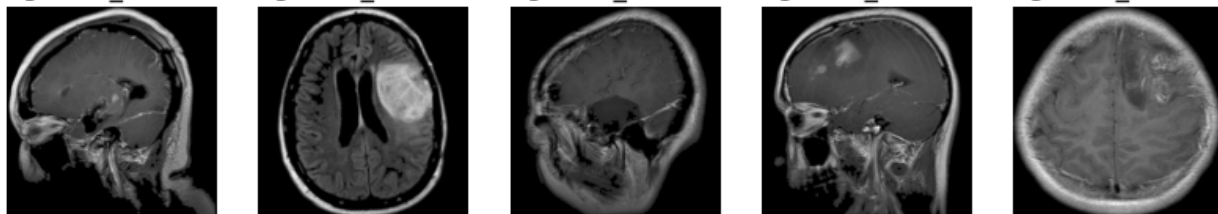
glioma_tumor

glioma_tumor

glioma_tumor

glioma_tumor

glioma_tumor



My Approach: My first goal was to process the images in a way where they would be in a readable format for a model. I resized each image to be 15 by 15 pixels but later realized that all of the images were already the same size. After this step, I flattened each image to create a 2d array with the pixel values of each image. Because the images were greyscale, I was able to use a 2d array instead of a 3d array. I wanted to establish a baseline model and used a support vector machine to start. Due to their strength and ability to compute in very high dimensional spaces, the SVC model performed quite well right off the bat with 86% accuracy. Because there are 4 classes, I decided to use accuracy as my performance metric and believe this will provide the most insight to the models overall performance. I also want to monitor the recall for each of the 4 categories because of the importance of avoiding false negatives in medical diagnostics. Now that I had a baseline performance to compare to, I decided to work with Keras deep learning library and create a neural net. Neural nets are commonly used and perform well on

image data and categorizing them into a class. With the neural net, I wanted to make sure I wasn't overfitting to the train data too much which is something that can happen quite easily with very powerful algorithms such as neural nets. Because of this, I decided to plot the training data accuracy against the test data accuracy to see how well the model was generalizing and if the two curves shared similar values. More similar values between the two would mean that the model is not overfitting too much and generalizing well to new data. With the neural net, I wanted to test different combinations of activation functions, layers, neurons in each layer and loss functions to see which yielded the best accuracy. To my surprise, I was having a difficult time beating the accuracy of the SVC base model and was getting similar or even worse performance as measured by validation data accuracy.

Model Performance: both the SVC model and neural net models performed best on the normal brain scans. final_model_CC which was a neural net had an accuracy of 83% but a high recall for normal scans of 97% with a precision of 87%. This means that it was catching 97% of normal scans but it was misdiagnosing some of the normal scans as having a tumor. This is preferable to me since it will miss as few positive tumor cases as possible even though it won't always categorize the tumor correctly. Some of the other models didn't have as good of performance on normal tumors but had higher metrics such as recall, precision and accuracy on the tumor categories. Each of the models has its strength in at least one of the categories and weaknesses in others. I would recommend using final_model_CC as a preliminary diagnostic to determine if a patient has a brain tumor or not since it can predict this with high accuracy. For patients who are flagged for having any type of tumor, they will be processed through a secondary algorithm. There are a few options for the second algorithm and an ensemble of second algorithms can also be used. The first option would be to choose the model with the best overall performance on tumor classification which would be model 2 which has f scores of 89, 89 and 92 for the tumor classes. The other option would be to pass the positive cases flagged by the first model through each of the models that performs best on a specific tumor type to get the most accurate classifications.

Further research: I would like to create synthetic data and split the classes into subcategories depending on the angle in which the photo was taken to improve model accuracy. I would need to create synthetic data in order to do this because there is not enough data currently and each class would be small and likely not capture the patterns and trends we are looking for. I would also like to create a model that is optimized for categorizing a scan as having a tumor or not having a tumor and then creating another model that is specifically for classifying what type of tumor was found in order to improve the current system being used.