Joelle Fitzgerald
HIDS 509 Final Project
Task Summary Report

The main objective of our final project was to train a model to accurately predict cancer type and identify highly important model features for detecting brain abnormalities when analyzing Magnetic Resonance Image (MRI) data. Segmented brain cancer MRI images from TCGA and Rembrandt database were used for this analysis. Rapid progression, high medical costs, and limited treatment options result in poor patient outcomes for brain cancer patients. Machine Learning (ML) algorithms tailored to detect specific features from MRI images revealing accurate information about tumor grade and characteristics, offer an opportunity for personalized medicine and improved patient outcomes through early detection and prognosis - working as a co-pilot with physicians and radiologists to provide patient care. To work to achieve an accurate ML model for detecting cancer type, there were 5 major tasks completed including data pre-processing, feature importance, model training, model optimization, and final result analysis.

Task 1

Task 1 involved the initial feature extraction using PyRadiomics. Pyradiomics is an open-source python package for the extraction of radiomics data from medical images. We first went through all 64 brain cancer patient files and used the PyRadiomics package to extract relevant features using the T1 modality image files and the cancer label files. It is important to note these files were already segmented. The feature extraction results were then appended to a csv file for Task 2.

Task 2

Reading in the TCGA csv file containing features for the training set, cancer type labels were added from the clinical data file. During data pre-processing, we found that the distribution of cancer labels were imbalanced and made a note to resample in Task 3.

Task 3

In Task 3, we first filtered out feature noise, such as the device manufacture information as this information is not as relevant for the purpose of our project. After this step, 107 features were retained. All records with the cancer type ‘Oligoastrocytoma’ were removed as the Rembrandt dataset does not include this cancer type label.

Task 4

Next in data preprocessing, we used SelectKBest method to evaluate features and select the top 10 most important features. We then scaled the whole dataset using StandardScaler and saved the scaled data into a file. After scaling, we applied a resampling pipeline to correct data imbalance: ADASYN upsampling, Repeated Edited Nearest Neighbours downsampling, and K-Means SMOTE oversampling to bring the amount of each label to an ideal number for training our model. Before training the models, the dataset was split into train and test sets with the ratio of 0.15.

We decided to train our data using 8 models. These models included, Support Vector Machine (SVM), Random Forest Classifier, Bagging Classifier, AdaBoost Classifier, K-Means Clustering, Linear Tree, Stacked Classifier, and a customer model. We chose to train our data on multiple models in order to find the most optimal result. We used cross-validation with 3 folds to train all models to prevent overfitting. When picking the best run to save, we manually inspected each test and train score and metric results when applying the models to the test set. Additionally, we also printed out feature importance when training the Random Forest Classifier and AdaBoost Classifier, for which they were the only algorithms that support this function. For the custom model, we used Bisecting K-Means to cluster the data points into 7 clusters, this number is obtained using the Elbow method to find the optimal number of clusters. We then used this information to train an AdaBoost model with Linear Tree model as the base estimator. Inside the Linear Tree base model, we used SGD Classifier as the base model.

Task 5

The same data pre-processing and reading in steps of Rembrandt data in Task 4, remained the same in Task 5. Several models were applied and trained on the data. We obtained all classification reports for each model type selected and run. The reports are included below.

SVM

	precision	recall	f1-score	support
Astrocytoma	0.00	0.00	0.00	24
GBM	0.00	0.00	0.00	8
Oligodendroglioma	0.18	0.88	0.30	8
accuracy			0.17	40
macro avg	0.06	0.29	0.10	40
weighted avg	0.04	0.17	0.06	40

Random Forest Classifier

	precision	recall	f1-score	support
Astrocytoma	0.62	0.21	0.31	24
GBM	0.40	0.50	0.44	8
Oligodendroglioma	0.18	0.50	0.27	8
accuracy			0.33	40
macro avg	0.40	0.40	0.34	40
weighted avg	0.49	0.33	0.33	40

Bagging Classifier

	precision	recall	f1-score	support
Astrocytoma	0.86	0.25	0.39	24
GBM	0.33	0.38	0.35	8
Oligodendroglioma	0.21	0.62	0.31	8
accuracy			0.35	40
macro avg	0.47	0.42	0.35	40
weighted avg	0.62	0.35	0.37	40

AdaBoost Classifier

	precision	recall	f1-score	support
Astrocytoma	0.68	0.71	0.69	24
GBM	0.33	0.50	0.40	8
Oligodendroglioma	0.33	0.12	0.18	8
accuracy			0.55	40
macro avg	0.45	0.44	0.43	40
weighted avg	0.54	0.55	0.53	40

K-Means

	precision	recall	f1-score	support
Astrocytoma	0.00	0.00	0.00	24
GBM	0.24	1.00	0.38	8
Oligodendroglioma	0.33	0.25	0.29	8
accuracy			0.25	40
macro avg	0.19	0.42	0.22	40
weighted avg	0.11	0.25	0.13	40

Linear Tree

	precision	recall	f1-score	support
Astrocytoma	0.57	0.17	0.26	24
GBM	0.32	0.88	0.47	8
Oligodendroglioma	0.18	0.25	0.21	8
accuracy			0.33	40
macro avg	0.36	0.43	0.31	40
weighted avg	0.44	0.33	0.29	40

Stacked Classifier

	precision	recall	f1-score	support
Astrocytoma	0.67	0.17	0.27	24
GBM	0.33	0.38	0.35	8
Oligodendroglioma	0.16	0.50	0.24	8
accuracy			0.28	40
macro avg	0.39	0.35	0.29	40
weighted avg	0.50	0.28	0.28	40

Custom Classifier

	precision	recall	f1-score	support
Astrocytoma	0.25	0.04	0.07	24
GBM	0.22	0.25	0.24	8
Oligodendroglioma	0.11	0.38	0.17	8
accuracy			0.15	40
macro avg	0.19	0.22	0.16	40
weighted avg	0.22	0.15	0.12	40

These results present an opportunity to improve detection models as a co-pilot for radiologists in detecting tumors and diagnosing brain cancer. From the classification reports above, AdaBoost demonstrated the best metrics for detecting cancer type when comparing models trained on the dataset available. However, like the other models trained above, a high precision score may indicate that the model is simply learning to predict one class very well - not giving a high accuracy for the overall performance of predicting all cancer types. Overall, all models above performed poorly possibly due to resampling and imbalanced data. By enhancing model feature selection and sampling techniques (possibly including more diverse data points), the models above may improve their accuracy scores also including precision, recall, and f1-score metrics.