Radiomics Analysis & Classification of Brain Cancer MRI Data:

Joelle Fitzgerald, Ansel Lian, Adam Li

The Data: TCGA and Rembrandt MRI Data

TCGA

Total of 167 samples:

21 Astrocytoma

91 GBM

26 Oligodendroglioma

29 Oligoastrocytoma

Rembrandt

Total of 64 samples:

24 Astrocytoma

8 GBM

8 Oligodendroglioma

24 with Missing Labels

Objective

Train a model to predict the cancer type with the top 10 most important features

Improve early detection of brain cancer to promote positive patient outcomes through early prognosis and tailored treatment

Quick Facts About Glioma

- The three cancer types we analyzed all belong to this category
- The name comes from the type of cell that turned into tumor: Glial cells
- It is the most common type of brain cancer, it accounts for 33% of cases
- Glioma comes in varying grades and growing rate
- It can start in either the brain or the spinal cord
- GBM, or glioblastoma, is a grade IV astrocytoma
- Astrocytoma and Oligodendroglioma are different glioma types

Task Overview

Task 1

Extract PyRadiomics features from MRI data (already segmented) of 64 brain cancer patients (REMBRANDT) and generate a multi-sample feature matrix

Task 2

Use TCGA Brain Cancer MRI dataset with radiomics features (already generated) to build/train a predictive ML model for chosen clinical attribute: **Cancer type**

Task 3

Apply our ML model to a test dataset - use that multi-patient PyRadiomics feature matrix for 64 REMBRANDT patients (generated in step 1) and predict the phenotype/clinical attribute : **Cancer type**

Task 4

Compare model predictions to ground truth data of known cancer type to find out the accuracy of your model and generate confusion matrix and other measures such as precision and recall

Resampling Techniques

ADASYN → RepeatedEditedNearestNeighbours → KMeansSMOTE

- Oversample: Astrocytoma, Oligodendroglioma →
- Undersample: GBM →
- Oversample: Astrocytoma, Oligodendroglioma

Final sample sizes: 98 Astrocytoma, 17 GBM, 99 Oligodendroglioma

Feature Selection

1. Manually removed unrelated features, e.g. diagnostics_Versions_Numpy, diagnostics_Mask-original_Hash, etc,.
2. 107 features left after initial selection
3. Sklearn’s SelectKBest for top 10 features
4. Scaled afterward uniformly across all features with StandardScaler

```
(['original_glrlm_GrayLevelNonUniformity',
  'original_glrlm_HighGrayLevelRunEmphasis',
  'original_glrlm_LongRunHighGrayLevelEmphasis',
  'original_glrlm_RunEntropy',
  'original_glrlm_ShortRunLowGrayLevelEmphasis',
  'original_glszm_GrayLevelNonUniformity', 'original_ngtdm_Busyness',
  'original_shape_MeshVolume', 'original_shape_SurfaceVolumeRatio',
  'original_shape_VoxelVolume'], dtype=object)
```

Training Models

Trained the following 8 models

SVM → Random Forest → Bagging → Ada Boost → K-Means → Linear Trees → Stacked → Custom

- 0.15 train test split ratio, fitted using cross validation with 3 folds
- RFC, Bagging, and Stacked had the best results
- SVM, and K-Means had the worst results

Evaluating Models

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 Ave.	0.40	0.40	0.34	40
Weighted Ave.	0.49	0.33	0.33	40

Evaluating Models

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 Ave.	0.47	0.42	0.35	40
Weighted Ave.	0.62	0.35	0.37	40

Evaluating Models

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 Ave.	0.45	0.44	0.43	40
Weighted Ave.	0.54	0.55	0.53	40

Evaluating Models

Linear Tree Classifier

	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 Ave.	0.36	0.43	0.31	40
Weighted Ave.	0.44	0.33	0.29	40

Feature Importance

Random Forest Classifier

	feature	importance
5	original_glszm_GrayLevelNonUniformity	0.226819
2	original_glrlm_LongRunHighGrayLevelEmphasis	0.178774
8	original_shape_SurfaceVolumeRatio	0.165986
3	original_glrlm_RunEntropy	0.145455
4	original_glrlm_ShortRunLowGrayLevelEmphasis	0.066358
1	original_glrlm_HighGrayLevelRunEmphasis	0.060471
6	original_ngtdm_Busyness	0.053653
0	original_glrlm_GrayLevelNonUniformity	0.042939
7	original_shape_MeshVolume	0.030724
9	original_shape_VoxelVolume	0.028820

	feature	importance
8	original_shape_SurfaceVolumeRatio	0.172608
2	original_glrlm_LongRunHighGrayLevelEmphasis	0.157769
3	original_glrlm_RunEntropy	0.141562
5	original_glszm_GrayLevelNonUniformity	0.133770
0	original_glrlm_GrayLevelNonUniformity	0.105052
4	original_glrlm_ShortRunLowGrayLevelEmphasis	0.085223
6	original_ngtdm_Busyness	0.063916
7	original_shape_MeshVolume	0.056745
9	original_shape_VoxelVolume	0.047562
1	original_glrlm_HighGrayLevelRunEmphasis	0.035793

	feature	importance
5	original_glszm_GrayLevelNonUniformity	0.185885
2	original_glrlm_LongRunHighGrayLevelEmphasis	0.147162
8	original_shape_SurfaceVolumeRatio	0.143689
3	original_glrlm_RunEntropy	0.116801
0	original_glrlm_GrayLevelNonUniformity	0.109980
1	original_glrlm_HighGrayLevelRunEmphasis	0.073592
4	original_glrlm_ShortRunLowGrayLevelEmphasis	0.070715
6	original_ngtdm_Busyness	0.067600
7	original_shape_MeshVolume	0.043253
9	original_shape_VoxelVolume	0.041324

Feature Importance

AdaBoost Classifier

```
                                      feature  importance
3                   original_glrlm_RunEntropy        0.48
8          original_shape_SurfaceVolumeRatio        0.48
2  original_glrlm_LongRunHighGrayLevelEmphasis        0.04
0       original_glrlm_GrayLevelNonUniformity        0.00
1     original_glrlm_HighGrayLevelRunEmphasis        0.00
4  original_glrlm_ShortRunLowGrayLevelEmphasis        0.00
5       original_glszm_GrayLevelNonUniformity        0.00
6                   original_ngtdm_Busyness        0.00
7                  original_shape_MeshVolume        0.00
9                 original_shape_VoxelVolume        0.00
                                      feature  importance
8          original_shape_SurfaceVolumeRatio        0.52
3                   original_glrlm_RunEntropy        0.48
0       original_glrlm_GrayLevelNonUniformity        0.00
1     original_glrlm_HighGrayLevelRunEmphasis        0.00
2  original_glrlm_LongRunHighGrayLevelEmphasis        0.00
4  original_glrlm_ShortRunLowGrayLevelEmphasis        0.00
5       original_glszm_GrayLevelNonUniformity        0.00
6                   original_ngtdm_Busyness        0.00
7                  original_shape_MeshVolume        0.00
9                 original_shape_VoxelVolume        0.00
                                      feature  importance
3                   original_glrlm_RunEntropy        0.48
5       original_glszm_GrayLevelNonUniformity        0.40
0       original_glrlm_GrayLevelNonUniformity        0.12
1     original_glrlm_HighGrayLevelRunEmphasis        0.00
2  original_glrlm_LongRunHighGrayLevelEmphasis        0.00
4  original_glrlm_ShortRunLowGrayLevelEmphasis        0.00
6                   original_ngtdm_Busyness        0.00
7                  original_shape_MeshVolume        0.00
8          original_shape_SurfaceVolumeRatio        0.00
9                 original_shape_VoxelVolume        0.00
```

PyRadiomics Features

1. ***First-order statistics:*** Distribution of gray levels within an image and include metrics such as mean, median, skewness, and kurtosis. Shown to be useful in distinguishing between different types of tissue and can provide information about the **overall intensity and distribution of gray levels** within an image.
2. ***GLCM features:*** The Gray Level Co-occurrence Matrix (**GLCM**) technique analyzes the spatial relationships between pixels with different gray levels. Features derived from the GLCM, such as contrast, **entropy**, and energy, provide information about the **texture and heterogeneity** of an image.
3. ***GLRLM features:*** The Gray Level Run Length Matrix (**GLRLM**) technique analyzes the distribution of runs of adjacent pixels with the same gray level intensity. Features derived from the GLRLM, such as Long Run High Gray Level Emphasis (LRHGLE), provide information about the **presence of irregular structures** in an image.
4. ***GLSZM features:*** The Gray Level Size Zone Matrix (**GLSZM**) technique analyzes the distribution of connected regions of pixels with the same gray level intensity. Features derived from the GLSZM, such as Gray Level **Non-Uniformity** (GLNU), provide information about the **heterogeneity** of an image.

Features of Interest

Original Shape Surface Volume Ratio

Original GLSZM Gray Level Non-Uniformity

Original GLRLM Gray Level Non-Uniformity

Original GLRLM Run Entropy

Original GLRLM Long Run High Gray Level Emphasis

Feature 1

Original Shape Surface Volume Ratio

Feature 1

Original Shape Surface Volume Ratio

Surface area compared to volume of the region/structure

Provides insights into brain development and potential abnormalities

Distinguish between benign (lower SVVR) and malignant (higher SVVR) tumor

Malignancy: irregular shape, rough surface, larger volume

Benign: more regular shape, smoother surfaces, smaller volume

Feature 2

Original GLSZM Gray Level Non-Uniformity

Feature 2

Original GLSZM Gray Level Non-Uniformity (Grey-Level-Size-Zone Matrix)

Homogeneity vs. heterogeneity of tissue structures

- Heterogeneous (varying grey level intensities) indicative of a mass

Varying grey-level intensities corresponding to areas of tissue with varying densities or textures compared to surrounding healthy tissue

Tumor characteristics, texture, and composition

Feature 3

Original GLRLM Gray Level Non-Uniformity

Feature 3

Original GLRLM Gray Level Non-Uniformity

Spatial relationship of grey-level values within an image and quantifies the distribution of runs of pixel values in a given direction and distance

- Tendency of pixels within an image to have similar or dissimilar gery values

Value high: the grey-level intensity of the pixels in the image varies significantly indicating high level of heterogeneity in tissue

Feature 4

Original GLRLM Run Entropy

Feature 4

Original GLRLM Run Entropy

Measures the randomness or complexity of the runs in the image - quantifies the degree of disorder in the distribution of runs of different lengths and gray levels in the image

Higher Run Entropy value indicates greater randomness in the runs of the image, whereas a lower value indicates a more regular pattern of runs

The Run Entropy feature is useful for detecting changes in texture patterns that may be indicative of pathological conditions in the tissue

Diagnostic marker, tumor stage, aid treatment plan

Feature 5

Original GLRLM Long Run High Gray Level Emphasis

Feature 5

Original GLRLM Long Run High Gray Level Emphasis

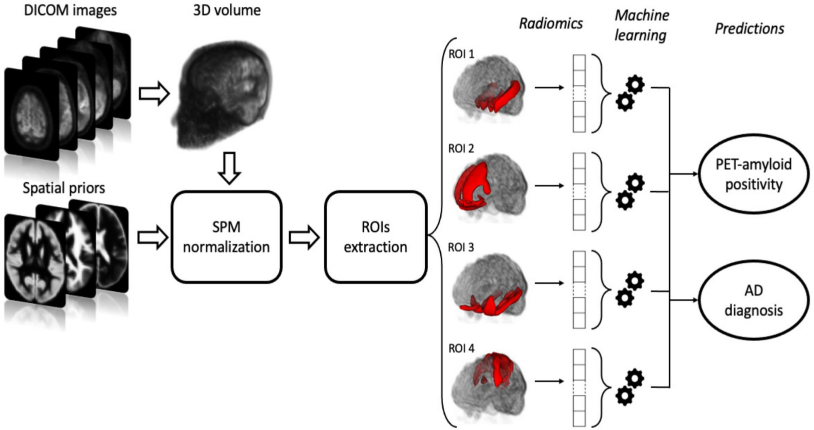
Measure of the tendency of an image to have long runs of high-intensity gey-levels

Brain MRI images, a higher LRHGLE value in certain regions may indicate the presence of abnormal tissue that have a higher gray level intensity and appear as irregular structures with longer runs in the image (tumor grade and size)

LRHGLE feature can be useful in aiding the diagnosis and monitoring of certain medical conditions

Conclusion

Discussion of Results



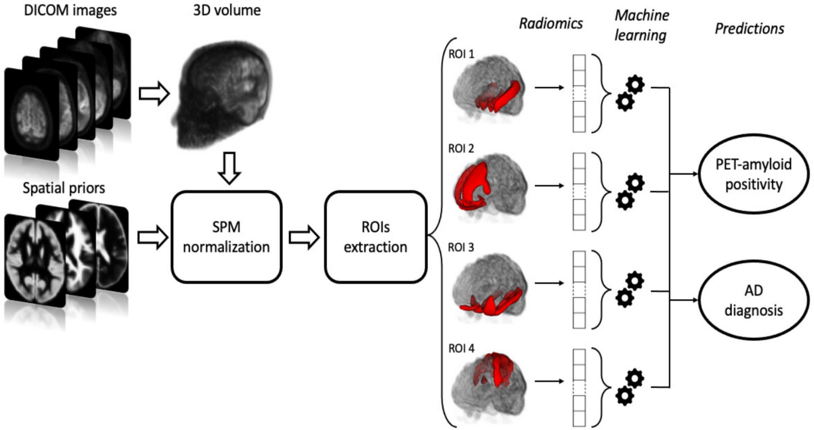
Main Objective: Train a model to predict the cancer type with the top 10 most important features

Purpose: Improve early detection of cancer/cancer type for brain cancer patients to promote positive patient outcomes through early prognosis and tailored treatment

Results: Present an opportunity to improve detection models to co-pilot radiologists in diagnosing brain cancer by enhancing model feature selection and sampling

Conclusion

Discussion of Results



AdaBoost Classifier: Highest metrics

High Precision Scores: Model is simply learning to predict one class very well

References

Becker, Anton S et al. “MRI texture features may predict differentiation and nodal stage of cervical cancer: a pilot study.” *Acta radiologica open* vol. 6,10 2058460117729574. 17 Oct. 2017, doi:10.1177/2058460117729574

“Brain Cancer and Gliomas.” *WebMD | WebMD*, 19 Nov. 2022, https://www.webmd.com/cancer/brain-cancer/malignant-gliomas

“Brain Tumor Types.” *Brain Tumor Types | Johns Hopkins Medicine*, 8 Nov. 2021, www.hopkinsmedicine.org/health/conditions-and-diseases/brain-tumor/brain-tumor-types.

Chaddad, Ahmad et al. “Radiomics in Glioblastoma: Current Status and Challenges Facing Clinical Implementation.” *Frontiers in oncology* vol. 9 374. 21 May. 2019, doi:10.3389/fonc.2019.00374

“Glioma.” *Mayor Clinic | Mayo Foundation for Medical Education and Research*, 10 Jan. 2023, https://www.mayoclinic.org/diseases-conditions/glioma/symptoms-causes/syc-20350251

Guillaume Thibault; Bernard Fertil; Claire Navarro; Sandrine Pereira; Pierre Cau; Nicolas Levy; Jean Sequeira; Jean-Luc Mari (2009). “Texture Indexes and Gray Level Size Zone Matrix. Application to Cell Nuclei Classification”. Pattern Recognition and Information Processing (PRIP): 140-145.

Rizzo, Stefania et al. “Radiomics: the facts and the challenges of image analysis.” *European radiology experimental* vol. 2,1 36. 14 Nov. 2018, doi:10.1186/s41747-018-0068-z

Thank You