Glioma Diagnosis between High Grade and Low Grade

This project focuses on analysing brain MRI images from the BRATS2020 dataset. Radiomic features were extracted from these images, specifically targeting regions with tumour masks. These features were utilized to train a Support Vector Machine (SVM) model for classification tasks.

Group members

Jarrad Harvey	21971144
Justin Duong	23179573
Vahkkshshan Sithsabasan	24065768

Data Partition

Here we got 369 data for each patient's MRI features with its Glioma grade Label, whether its HGG or LGG.

First, we are turning our *String Labels* to *Int Labels* with label encoder from sci kit, since string values when turned to represent data would be represented in float values. Since floats aren't accurate the data points would be varying with each even for the same label.

After that we are taking 10 of each HGG and LGG for our hidden test.

From the rest of the data, we can see a class imbalance where LGG is the minority. We used some techniques available to overcome this,

1. SMOTE Oversampling:
   - Here we used oversampling method from inbuilt functions from sci kit to create synthetic data of the minority class.

2. Stratified K- Fold:
   - with the train_test_split() we can turn on the stratify flag, unlike traditional K fold, stratified makes sure that depending on how many folds we want, each fold will contain a balanced class labels.

3. Class Weight Balance:
   - This flag is turned on our SVM model initialization. This assigns a higher weight to the minority class, so the function is more inclined towards the minority class.

We, eventually settled down with Class Weight Balance since all of it gave an almost equal level of accuracy

Features Used

We are using the radiometric features extracted using the radiomics library.
We are taking entire 3d volume of brain for each channel and use the combined mask to pass into the feature extraction method of radiomics. From there we analyse repeatability scores to check out the variance from repeated measurement or feature.

Then we take top 10 features from shape, intensity and texture. The following is the list of features employed in our analysis:

Shape features

- original_shape_Elongation
- original_shape_Flatness
- original_shape_LeastAxisLength
- original_shape_MajorAxisLength
- original_shape_Maximum2DDiameterColumn
- original_shape_Maximum2DDiameterRow
- original_shape_Maximum2DDiameterSlice
- original_shape_Maximum3DDiameter
- original_shape_MeshVolume
- original_shape_MinorAxisLength

Intensity features

- original_firstorder_90Percentile
- original_firstorder_Entropy
- original_firstorder_InterquartileRange
- original_firstorder_Maximum
- original_firstorder_MeanAbsoluteDeviation
- original_firstorder_Minimum
- original_firstorder_Range
- original_firstorder_RobustMeanAbsoluteDeviation
- original_firstorder_RootMeanSquared
- original_firstorder_Uniformity

Texture features

Gray level co-occurrence matrix

- original_glcm_Idmn
- original_glcm_Idn
- original_glcm_Imc1

Gray level dependence matrix

- original_gldm_DependenceEntropy
- original_gldm_DependenceNonUniformity
- original_gldm_DependenceNonUniformityNormalized

Gray level run length matrix

- original_glrlm_RunEntropy
- | original_glrlm_ShortRunEmphasis

Gray level size zone matrix

- original_glszm_SmallAreaEmphasis
- original_glszm_ZoneEntropy

Model Performance

Training and Validation

The SVM model with an RBF kernel (`C = 1.0`, `gamma = 1e-07`) was trained using 10-fold cross-validation. The average performance metrics across the folds were:

Validation accuracies	0.815	0.812	0.822	0.828	0.822	0.821	0.822	0.828	0.822	0.825
Accuracy scores	0.829	0.857	0.857	0.829	0.829	0.800	0.829	0.829	0.800	0.765

Hidden Test Set

The final model achieved an accuracy of 0.5 on the hidden test set, indicating areas for improvement with unseen data.

Feature Selection Considerations and Discussion

When working with radiomic features, ensuring their repeatability is important to maintain consistency and not much variance across different instances. However, just depending on repeatability of a feature is not advisable. Some features might not be relevant in classifying the labels, and more of the features which might not be related at all would end up making the model complex as well. Therefore, criteria such as relevance and redundancy reduction are important to be considered as well. Compactness and Saliency are what we call these characteristics.

By maintaining compactness we are basically trying to reduce the model complexity and prevent model from being overfitted as well. By Saliency, we meant using the relevant most prominent features that describe a data point best. This would ensure we have relevant data to give the best possible prediction while making a stable robust model.

Challenges Encountered

Throughout the classification process, several challenges were encountered. These challenges included:

1. *Class Imbalance*: Dealing with class imbalance, particularly when LGG class is significantly underrepresented as minority, posed challenges in model training and evaluation. Techniques such as oversampling methods or class weighting were used to figure the best outcome.

2. *Feature Selection*: Selecting the most relevant and informative features from a large set of radiomic features. Making sure to reduce feature dimensionality while making good predictive outcome was challenging.

3. *Model Interpretability*: SVM models are good at classification, yet it was hard to find a good decision boundary to represent the underlying data pattern.

4. *Hyperparameter Tuning*: Having to search for the best parameters with GridSearch for the SVM model with different kernels, Gamma values was tricky. Main challenge was time taken to find the best parameters and best model.

5. *Computational Resources*: It took huge amounts of time to process data for all 369 volumes. We at times had to work with few amounts of data in order to quickly fid bugs and do the code. In the end it still took a lot of time not just to extract features but to split train on stratified k fold with validation all the while searching for best model through Grid Search.