Glioma Classification Project – SVM Discussion

Mitchell Otley **(23475725)**  
James Wigfield **(23334375)**  
Nate Trew **(23120643)**

Data Collection

The data used to train the SVM is synthesized from the extracted conventional features (maximum tumor area, maximum tumor diameter and outer layer involvement), and our top ten intensity-based, shape-based, and texture-based radiomic features of each MRI volume in the dataset. The SVM model concatenates this data with data extracted from the *name\_mapping.csv* file, specifying the classification of each volume as to being High-Grade Glioma (HGG) or Low-Grade Glioma (LGG).

A screenshot of a computer

Description automatically generatedThe result is a csv table of the dimension 369x35 for the provided dataset (33 columns for the extracted features of each of the 369 MRI volumes, one column for the volume label and one column for the grading of the tumor).

1: Sample of combined\_features.csv

The conventional features and radiomic features are all obtained through the ‘Extract Conventional Features’ and ‘Extract Radiomic Features’ buttons in our MATLAB GUI.

Feature Selection

Conventional Features:

* Maximum Tumor Area
* Maximum Tumor Diameter
* Outer Layer Involvement

Intensity-Based Radiomic Features:

* Discretised Intensity Skewness
* Intensity Kurtosis
* Minimum Histogram Gradient
* Minimum Discretised Intensity
* Volume at 10% Intensity Fraction
* Volume Fraction Difference between Intensity Fractions
* Discretised Intensity Entropy
* Maximum Intensity
* Intensity Histogram Coefficient of Variation
* Global Intensity Peak

Shape-Based Radiomic Features:

* Minor Axis Length
* Smallest Axis Length
* Volume Density (Approximate Enclosing Ellipsoid)
* Volume (Mesh)
* Elongation
* Volume (Voxel Count)
* Major Axis Length
* Flatness
* Surface to Volume Ratio
* Spherical Disproportion

Texture-Based Radiomic Features:

* Information Correlation 1 - Merged
* Information Correlation 1 - Averaged
* Dependence Count Percentage
* Normalised Inverse Difference Moment - Merged
* Normalised Inverse Difference Moment - Averaged
* Normalised Inverse Difference - Merged
* Normalised Inverse Difference - Averaged
* Information Correlation 2 - Merged
* Information Correlation 2 - Averaged
* Run Entropy - Merged

Data Partitioning

As per the project specifications, 10 randomly selected HGG and LGG patients are assigned to a ‘hidden’ testing set. The remaining 349 patients are utilised in the training and validation of the SVM. However, of these 349 patients, only 66 patients are classified as having LGG. As the SVM is sensitive to unbalanced datasets, the resulting model would exhibit a bias towards classification of unseen samples to the class with a larger representation (in this case, HGG). To mitigate this, we reduce the dataset passed to the SVM to be equal between the binary classes. In essence:

* 10 LGG and 10 HGG patients assigned to testing set
* 66 LGG Patients assigned to training set
* 66 HGG Patients assigned to training set
* 217 HGG Patients unassigned (not used in training or testing of the SVM)

We don’t want 217 data samples being unused, so to account for this we train 1000 iterations of the SVM Classifier, each time using a different random sampling of 66 HGG patients of the available 283. We then select the model with the highest training accuracy as our preferred model to use on the testing set.

During the model training phase, we utilise *k-fold cross validation* to create a validation set. This method splits the training data into *k=5* same-sized groups in our model, wherein each group is used to perform a validation of the other groups of data in the training set. We chose *k=5* based on empirical recommendations.[[1]](#footnote-1)

Model Accuracy

A screenshot of a computer

Description automatically generatedWe opted to use MATLAB’s Classification Learner App to develop a model, and exported it to a function to incorporate with our data.   
Running the SVM Classifier with 1000 iterations yielded a model with a validation accuracy of 80.3% (106 out of 132 patients correctly classified). When this model was tested with the hidden dataset, the accuracy was 70% (14 out of 20 patients correctly classified).

2: Sample output from SVM Classifier

Impact of Feature Selection on Accuracy

The SVM accuracy is reliant on selecting the correct radiomic features, as the differences of these features in HGG and LGG determine how well the SVM can predict future cases, using these features. If poor selectors are used, the SVM will be unable to accurately identify the difference between an LGG patient and HGG patient.

As we are using 33 different features to train our model, it is likely that some of these features will have minimal impact on the outcome of the model, and are just increasing the dimensionality of the predictor space. This may be leading to overfitting of the data. To improve on this, we could consider using Principal Component Analysis to remove the redundant features, and only keeping enough components to explain a certain percentage of variance in the model (e.g. retaining 95% variance).

Our repeatability test is calculated as follows:

1. For each acquisition protocol of a volume, the radiomic features are calculated.
2. The mean value for each feature is calculated across the acquisition protocols, and from that the average standard deviation from the mean is calculated
3. The average standard deviation is normalised across different features, to ensure comparability. We use standard deviation as a measure of repeatability, as features with low standard deviations indicates a low variance of a radiomic feature across acquisition protocols.
4. The top ten features from Intensity-, Shape-, and Texture-based radiomics are selected, by selecting the ten features with the lowest average standard deviation across the 369 volumes.

Whilst using repeatability may be a satisfactory method for binary classification, it does have some shortcomings. Using highly repeatable features is useful for selecting features that are consistent across volumes, but does not guarantee that the features are distinguishable between classifications. For example, some radiomic features, like Volume, have a high repeatability, but don’t necessarily distinguish well between High-Grade and Low-Grade Glioma. This is because each MRI scan is taken at a different time of diagnosis, the time taken for the tumor to grow is not factored into the calculation.

To improve the accuracy of the SVM model, we could also consider features based on their saliency – the property that allow features of an image to be more prominent within visual clutter. Considering further criteria will allow for the selection of features that best classify the different grades.

Challenges Encountered

One issue we did encounter was deciding on which kernel function to use to train the dataset on (between Linear, Gaussian, Cubic, Quadratic, etc.). After training an SVM using each kernel function, we selected a linear kernel, as MATLAB documentation recommends it as the default for two-class classification.

1. [*https://machinelearningmastery.com/k-fold-cross-validation/*](https://machinelearningmastery.com/k-fold-cross-validation/) [↑](#footnote-ref-1)