SVM Classification Model

Data partition

The data is split into three parts

- Test Data, data to learn from
- Validation Data, larger db for checking against
- Hidden Data, final smaller benchmark to test from.

The overall data set consits of 368 total unique volumes
The SVM learnt on the Test data that consisted of all 368 volumes minus 20 hidden data. Then was validated against 100 random volumes. Then against the 20 hidden data set.

A small caviate of this is that the High Grade Glioma(HGG) and Low Grade Glioma(LGG) were not present in the same ratio in the test data with the test and validation data being appoximatly 65-70% High Grade Gliomas while the hidden set contained 50% HGG and 50% LGG. This made it more difficult for the SVM to do better on the hidden data as it expected more HGG than LGG.

Features used for training

The dataset contains the following features extracted from the images.

1. VolumeMesh3D
2. SurfaceAreaMesh3D
3. Sphericity3D
4. VolumeDensityAABB_3D
5. MajorAxisLength3D
6. MinorAxisLength3D
7. Elongation3D
8. Flatness3D
9. IntergratedIntensity3D
10. MeanIntensity3D
11. MedianIntensity3D
12. MinimumIntensity3D
13. MaximumIntensity3D

14. IntensityVariance3D
15. IntensitySkewness3D
16. IntensityKurtosisD
17. IntensityRange3D
18. IntensityInterquartileRange3D
19. RootMeanSquare3D
20. JointEntropyAveraged3D
21. AngularSecondMomentAveraged3D
22. ContrastAveraged33D
23. DissimilarityAveraged3D
24. ClusterTendencyAveraged3D
25. ClusterShadeAveraged3D
26. ClusterProminenceAveraged3D
27. InverseDifferenceAveraged3D
28. CorrelationAveraged3D
29. AutoCorrelationAveraged3D
30. maxTumorArea
31. maxTumorDiameter
32. outerLayerInvolvement
33. gliomaGrade

The first column *LabelID* is the volume number and is ignored and the last column, *gliomaGrade* is the target.

Accuracy of the model

We ran the SVM 10 times then averaged the result to ensure reliability of result

Training accuracy

The SVM was 89.971% accurate on the test dataset

Validation accuracy

The SVM was 82.3% accurate on the validation dataset

Testing accuracy

The SVM was 67% accurate on the hidden dataset

Challenges encountered during the classification process

A massive challenge we encountered was that the data had a massive inbalance of high grade glioma and low grade gliomas. The data has a lot more high grade glioma compared to low grade glioma, this causes issues with the SVM learning as it baised towards high grade glioma guesses. This means that when trying to guess the glioma grade in a perfect dataset such as the hidden dataset which is a 50/50 split in high grade gliomas and low grade gliomas.

There are solutions to mitigate these problems such as transforming the dataset so that there is a 50/50 split in classification. However doing this removes too much data to learn from and the SVM struggles. Another solution would be to have a massive dataset so it can learn very accuratly what a low grade glioma and a high grade glioma means, this is a problem for us however as it would take a long time to learn from and we also do not have access to a dataset big enough to learn from. An approach we were told in during our presentation was to duplicate and or slightly change the smaller low grade glioma dataset so that we had enough data to learn from.