Sydney University Masters of Data Science Capstone Project Final Report

Classification of brain tumor tissue images using Convolutional Neural Networks

Tim Johnson (tjoh6207@uni.sydney.edu.au)

Student Number: 198251447

Advisors: Tom Cai & Yang Song

Abstract

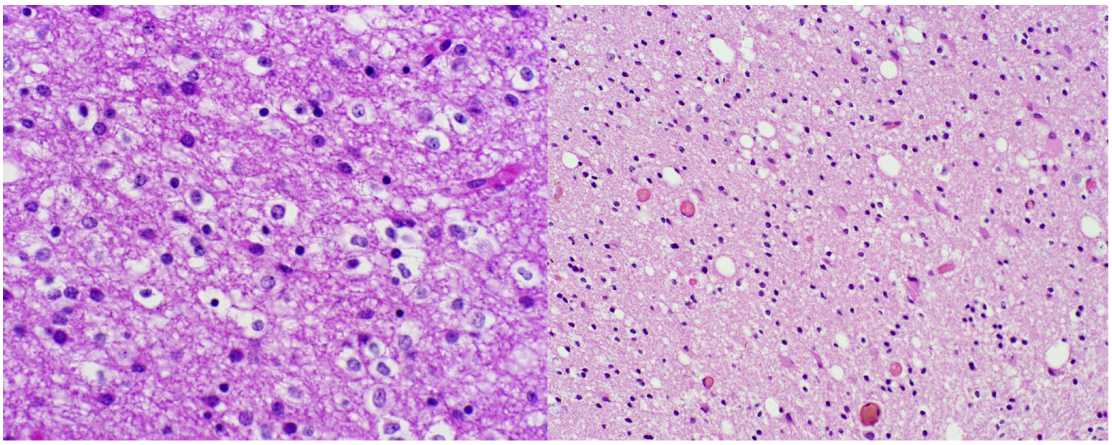
Low Grade Glioma (LGG) encompasses a highly diverse set of primary brain tumors, the classification of which suffers from high levels of inter-observer variation. In this study we examined the ability of features extracted from ResNet50, a state of the art pre-trained Convolutional Neural Network (CNN) to capture the richness of histological features of LGG and to classify the two types: Astrocytoma and Oligodendroglioma. The primary training set for the study was a 32 image set from the MICCAI 2015 Challenge of Computational Brain Tumor Cluster of Events (CBTC) with a second set of 104 images sourced from The Cancer Genome Atlas (TCGA) added to address overfitting concerns. We achieved an accuracy of 68% on the CBTC dataset using generic CNN-derived feature extraction. This result was 7% lower than the best result in the same challenge and similar to other approaches using hand crafted features. The result suffered from overfitting at an image level despite excellent cross validation performance at a tile level. This phenomenon was correlated with the observation that some images in the set achieved high, incorrect probabilities, illustrating detection by the classifier of some intra-class differences and inter-class similarities that were confirmed nueropathologically. Fine tuning the CNN weights suffered severe overfitting and a worse result was achieved than for generic features. Further research could involve repeating the experiment with a much larger image set, starting with histologically clear examples and gradually adding less classical shapes, in a more targeted attempt to address overfitting. Given the inter-observer discordance in the field, neuropathologist partnership may be highly beneficial in further investigations to confirm image labelling, map histological features to CNN-derived features and analyse the histological significance of feature vector clustering.

Introduction

Background Analysis and Literature Review

There are over 120 different types of brain tumors [1]. The most frequent type is Glioma, making up 40% of primary brain tumors [6]. Glioma represents a diverse category of tumor with as many different types as there are different types of glial cells from which they gain their genesis.

Of particular interest in this study is Low Grade Glioma (LGG) which includes types Astrocytoma, Oligodendroglioma and mixed Oligoastrocytoma. Although incidence rates of LGG are difficult to determine, Schiff et al [5] claim rates to be of the order of 15% of primary brain tumors, much smaller than for higher grade glioma, and therefore potentially more challenging to classify. Following is a figure showing classical formations of the respective tumor types wth a brief description of their histology.



Oligodendroglioma: Relative roundness and monotony to nuclei, Stippled chromatin, Perinuclear halos (fried egg appearance).

Astrocytoma: Modest hypercellularity, Nuclear pleomorphism, Relative nuclear hyperchromasia

Figure. Classical formations of Oligodendroglioma (left) and Astrocytoma (right)

According to Nader and Schiff [5], LGG is technically considered to include cancer grades I and II but there are considerably different histology and prognostic outcomes for tumors within these grades. Additionally, as Barker et al [7] point out, The Cancer Genome Atlas (TCGA) includes grade III gliomas amongst their low grade samples.

Differentiating between Astrocytoma and Oligodendroglioma tumor types is important because, as Wesseling et al [2] point out, it has significant implication for therapy and prognosis. Furthermore, as Jansen et al [15] note, tumors with Oligodendrial components have better outcomes than those with Astrocytic components. However, as Smits [4] points out, Astocytoma and Oligodendroglioma both present non-classical shapes that are sometimes difficult to distinquish from one another.

The World Health Organisation (WHO) provides the reference standard for classification and grading of tumors of the Central Nervous System (CNS), the most recent version of which was released in 2016. Astrocytic and Oligodendroglial tumor types from the 2016 version have been assembled below from papers by Tatter [17] and Louis et al [26], combined with general characteristics from Mayfield Brain and Spine [1].

Table. Type and Grading of selected CNS tumors according to the 2016 CNS WHO

Grade	Tumor Type	Characteristics
I	Pilocytic astrocytoma Subependymal giant cell astrocytoma Pleomorphic xanthoastrocytoma	Slow growing Almost normal appearance Least malignancy, Long term survival
II	Diffuse Astrocytoma (IDH-mutant) Diffuse Astrocytoma (IDH-wildtype) Diffuse Astrocytoma (NOS) Oligodendroglioma (IDH-mutant and 1p/19q-codeleted) Oligodendroglioma (NOS) Oligoastrocytoma (NOS)	Relatively slow growing Slightly abnormal appearance Can invade nearby tissue Sometimes recurs as a higher grade
III	Anaplastic Astrocytoma (IDH-mutant) Anaplastic Astrocytoma (IDH-wildtype) Anaplastic Astrocytoma (NOS) Anaplastic Oligodendroglioma (IDH-mutant and 1p/19q-codeleted) Anaplastic Oligodendroglioma (NOS) Anaplastic Oligoastrocytoma (NOS)	Actively reproducing abnormal cells Abnormal appearance Infiltrates normal tissue Tend to recur, often at higher grade
IV	Glioblastoma multiforme	Rapidly reproducing abnormal cells Very abnormal appearance Necrosis in the centre

		Form new blood vessels for growth

Ceccarelli et al [12] point out that, despite the WHO standard, histopathological classification of Oligodendroglioma has been difficult to define and is subject to poor reproducability between pathologists. Martin J. van den Bent [20] sites Bruner et al in stating “some degree of disagreement was present in 42.8% [of pathological diagnoses of glioma in their study], which was considered serious in 8.8%”. Coons et al [14] suggest this discordance is due to “subjective histologic criteria” used in the classification and grading process.

The author of the Oligodendroglioma Wikipedia page [11] gives more specific commentary relating to “controversy” between East and West coast trained neuropathologists in the US related to the classification of Oligodendroglioma in non-classical shapes. With some neuropathology facilities in the US diagnosing Oligodendroglioma substantially more often than others. Buckland [24] confirms similar experiences apply for different facilities in Sydney, Australia where rates of Oligodendroglioma diagnosis are known to differ by up to 100%.

Jansen et al [15] point out the subjective nature of histological criteria for mixed Oligoastrocytoma classification as the cause of substantial inter-pathologist disagreement. Oligoastrocytoma includes distinct histological characteristics of both Astrocytoma and Oligodendroglioma and is known to demonstrate behaviour of either type based on the molecular profile of the patient. However, a common diagnostic mistake, as Buckland [24] observes, is to use this mixed type for borderline classifications of astrocytoma and oligodendroglioma.

Louis et al [25] note that, in addition to histological phenotypes, for the first time the 2016 WHO CNS classifications include elements associated with the patient’s molecular profile. In particular, identification of 1p/19q codeletion results in better clarity for LGG classification when combined with histological techniques.

Histological diagnosis of a tumor type is a two-stage process. Firstly, a preliminary diagnosis of a tumor location, type and size can be made by X-ray computed tomography (CT) or magnetic resonance imaging (MRI) scans as part of a neurological examination. Confirmation of the tumor diagnosis and sub-typing then requires neuropathological examination of a tumor biopsy.

The primary motivation of this study is to improve the efficiency and speed of histological classification after biopsy by making use of fully automatic techniques performed by computer. In particular, no custom feature selection is used in this approach. The hope is that automation may eventually eliminate the subjective nature of LGG diagnosis. Given the numerous diagnostic challenges and noise in this field, and the fact that the study of computer based methods are in relative infancy, there is clearly more work to be done. However, recent computer vision advancements made with Convolutional Neural Networks (CNN) make a study that includes those techniques timely as an important step along the way.

Related Research

Barker et al, in their paper on automated classification of LGG and GBM [7], note that the key challenges to computer-based analysis of pathology images are twofold:

1. The enormous size of whole slide images (WSI) used
2. The diversity of tissue regions of the slide not related to the disease

They propose a technique to overcome these challenges by cutting up the images into small tiles, and deriving the localized features within tiles in a two-stage process moving from course feature selection to finer-grained feature selection. As part of this process, feature vectors are grouped into clusters and representative tiles selected and used for input into the fine grain feature selection stage. In this paper, a similar approach is used to tile images and cluster their feature vectors to identify best representative tiles for the image.

Though the feature engineering process outlined by Barker et al in [7] is automated, it is somewhat complicated and requires specialised domain expertise related to the features being engineered. In this paper, taking inspiration from Spanhol et al [8], a Convolutional Neural Network (CNN) is used to derive features from the tiles in an attempt to produce a fully automated approach to the binary classifation problem under consideration. Transfer learning from the ResNet50 model trained on the ImageNet dataset is used to avoid compute resource challenges associated with training the neural network. Features are derived from ResNet50 in two ways: 1) directly from the second top layer of the network without tuning and 2) after fine tuning of the network with our histological images.

Once features are extracted, classifiers (SVM, Random Forest and XGBoost) are trained to make predictions at a tile level using n-fold cross validation with grid search to determine optimal parameters. Predictions made at an image level are based on a similar voting scheme as that outlined by Barker et al in [7].

The first dataset utilized in this study was used for the MICCAI 2015 challenge of Computational Brain Tumor Cluster of Events (CBTC). Results are evaluated in comparison to papers from that challenge, in particular Song et al [9] and Joel Carlson [10], and characteristics of that dataset explored thoroughly.

Additional data was collected from TCGA to investigate overfitting of data at an image level.

Methods and Materials

Dataset Description

The data used in this study is taken from two sources:

1. MICCAI 2015 challenge of Computational Brain Tumor Cluster of Events (CBTC)
2. The Cancer Genome Atlas (TCGA)

Table. Characteristics of the CBTC dataset

Number of Examples			File format
Total	Astrocytoma	Oligodendroglioma	Aperio SVS whole slide images
32	16	16	
Description	Aperio SVS files are potentially very large. In this dataset, some files are over 1.5Gb. They contain multiple levels of resolution from the original 40x image with downsampling factors for the other levels varying: 4, 8, 16, 32, 64. Images used for testing purposes were the level 1, 4x downsampled images. The distribution of labels in the data are balanced between Astrocytoma and Oligodendroglioma.		

Table. Characteristics of the TCGA dataset

Number of Examples					File format
Total	Astrocytoma		Oligodendroglioma		Aperio SVS whole slide images
	Grade II	Grade III	Grade II	Grade III	
104	17	35	33	19	
Description	TCGA images are sourced from patients with grade II and III glioma of types Astrocytoma, Oligodendroglioma and mixed Oligoastrocytoma. A single image was selected from each patient and 52 examples of Astrocytoma and Oligodendroglioma were selected to maintain a balanced datatset. Images used for testing purposes were the level 1, 4x downsampled images.				

Methodology

The processing in this study is perfomed in several stages as follows:

Stage 1. Image Preparation

Stage 2. Feature Extraction using a Convolutional Neural Network

Stage 3. Tile Clustering and Selection

Stage 4. Machine Learning

Stage 5. Inter-Imageset Testing

Stage 1: Image Preparation

Whole Slide Images (WSI) were processed in stages. A summary of the image processing stages for the first dataset are listed in the figure below. A similar process was used for the second dataset.

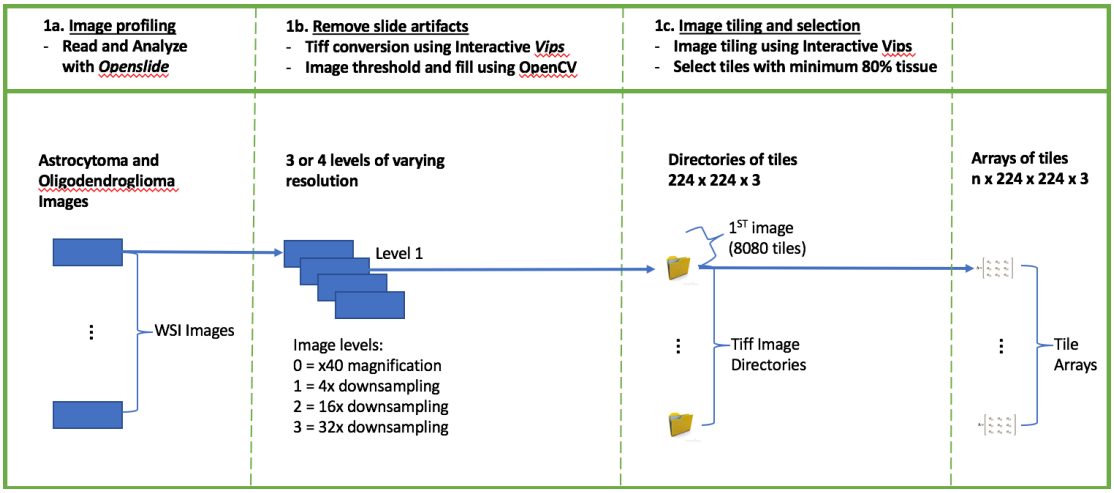


Figure. Image Processing Steps

1a. Image profiling

The python OpenSlide module was used to read and profile images. In particular, image resolution details were considered in order to decide on appropriate resolution levels and tiling dimensions. The following tables show details of dimensions of the 4 levels in the main Whole Slide Image set for this study.

Table. Whole Slide Image Size Details for CBTC Dataset

Level 0		**Level 1**		**Level 2**		**Level 3**	
X40 Magnification		**Downsampling: 4x**		**Downsampling: 8 or 16x**		**Downsampling: 32x or 64x**	
Height (pixels)	**Width (pixels)**	**Height (pixels)**	**Width (pixels)**	**Height (pixels)**	**Width (pixels)**	**Height (pixels)**	**Width (pixels)**
Astrocytoma images:							
89640	70989	22410	17747	5602	4436	2801	2218
147723	69077	36930	17269	9232	4317	2308	1079
65228	28836	16307	7209	4076	1802		
178024	58022	44506	14505	11126	3626	2781	906
95924	51658	23981	12914	5995	3228	2997	1614
74821	75577	18705	18894	4676	4723	2338	2361
165949	84949	41487	21237	10371	5309	2592	1327
176501	85030	44125	21257	11031	5314	2757	1328
95924	23008	23981	5752	5995	1438	2997	719
27887	37919	6971	9479	1742	2369		
47962	34524	11990	8631	2997	2157		
84413	83008	21103	20752	5275	5188	2637	2594
94006	79535	23501	19883	5875	4970	2937	2485
70448	3402	17612	8505	4403	2126	2201	1063
151560	76709	37890	19177	9472	4794	2368	1198
27608	29571	6902	7392	1725	1848		
Oligodendroglioma images:							
84413	67031	21103	16757	5275	4189	2637	2094
80576	78159	20144	19539	5036	4884	2518	2442

59759	51040	14939	12760	3734	3190	1867	1595
119520	57706	29880	14426	7470	3606	3735	1803
71943	68955	17985	17238	4496	4309	2248	2154
107521	45460	26880	11365	6720	2841	3360	1420
73861	84188	18465	21047	4616	5261	2308	2630
113191	42668	28297	10667	7074	2666	3537	1333
111552	63155	27888	15788	6972	3947	3486	1973
41832	39671	10458	9917	2614	2479		
100720	32678	25180	8169	6295	2042	3147	1021
26858	16353	6714	4088	3357	2044		
146764	82254	36691	20563	9172	5140	2293	1285
124701	46574	31175	11643	7793	2910	3896	1455
16307	19556	4076	4889	2038	2444		
172664	84999	43166	21249	10791	5312	2697	1328

Image resolution – Level 1 images (the second highest resolution) were used as the optimal choice to put a cap on resource requirements compared with the massive level 0 images and a loss of resolution and hence signal at lower levels. Levels 1 to 3 were also processed and the train and test harness had the ability to choose between levels. However, as the study progressed it became evident that lower resolution images were of limited value for our purposes.

1b. Removal of Slide Artifacts

The interactive *Vips openslideload* function was used to convert SVS files to tiff. OpenCV was then used to remove background blank slide areas from consideration, thereby allowing classifiers to operate only on foreground tissue. Image thresholding and flood filling was used to differentiate between image background and foreground and to colour background pixels black.

1c. Image Tiling and Selection

A decision was made to use tiling dimensions of 224 x 224 pixels since that is the default size for the ResNet50 CNN used for feature extraction. *Vips dzsave* was used to create tiled versions of the images. Vips is fast and requires minimal memory for its purpose. All tiles were stored to disk as jpg files.

Both background-removed and original images were tiled, the former serving as an index into the tiles of the latter. Based on a ratio of foreground to background pixel values on background-removed images, tiles that constituted 80% of tissue were selected from original images and used for further testing. This final set of tiles is written to disk as a numpy array for input into the ResNet.

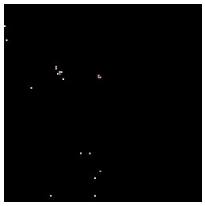
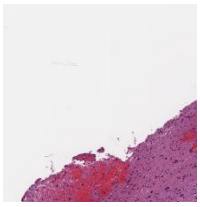
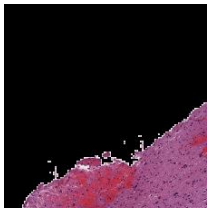
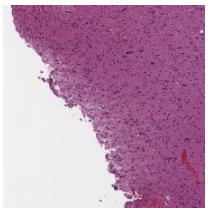
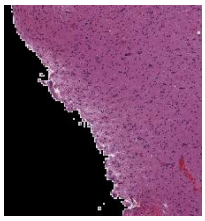
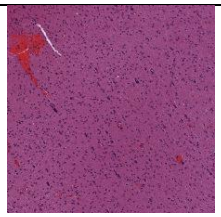
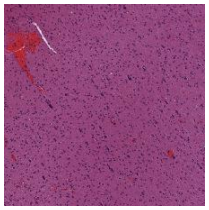
Original Image	With Background Differentiated in Black	Calculated Tissue Percentage
		**1.7%**
		**28.8%**
		**91%**
		**100%**

Figure. Example tiles, original (left) and with background removed (right). The bottom two tiles with 80+% tissue were retained

Stage 2: Feature Extraction using the ResNet50 Convolutional Neural Network (CNN)

Resnet50, a pretrained convolutional neural network, was used to derive features from images under test. Keras with a Tensorflow backend was used to download the CNN. Resnet50 is a recent iteration of the famous Alexnet neural network but is substantially deeper, with higher accuracy on image classification tasks, reaching better than human performance on the ImageNet dataset. Resnet50 is a 50 layer version of the winning 152 layer submission in the 2015 ImageNet Large Scale Visual Recognition Competition [18] as shown in the figure below.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
		3×3 max pool, stride 2				
conv2_x	56×56	[3×3, 64; 3×3, 64] ×2	[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_x	28×28	[3×3, 128; 3×3, 128] ×2	[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
conv4_x	14×14	[3×3, 256; 3×3, 256] ×2	[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512; 3×3, 512] ×2	[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure. Diagram from He et al [18] showing different versions of ResNet that won the 2015 ILSVR competition. ResNet50 circled in red.

CNN Transfer learning

Making use of transfer learning from the ResNet50 network is one of the main novel approaches under test in this study in an effort to find a fully automated approach to feature extraction of histological images. Transfer learning makes it possible to gain access to the capabilities of a CNN trained against millions of images in the ImageNet database that may otherwise not be possible, since contemporary ImageNet based CNNs take weeks to train with multiple GPUs [22]. In the first instance, as shown in the figure below, generic features were extracted by instantiating the entire convolutional part of the model up to but excluding the fully-connected layers. The model was run using our tiles as input. The 4D tensor output from the ResNet50 represents the features of our tiles. ResNet50 yields a 2048 dimension feature vector in the last dimension (tensorflow format). All features were written to disk as a numpy array.

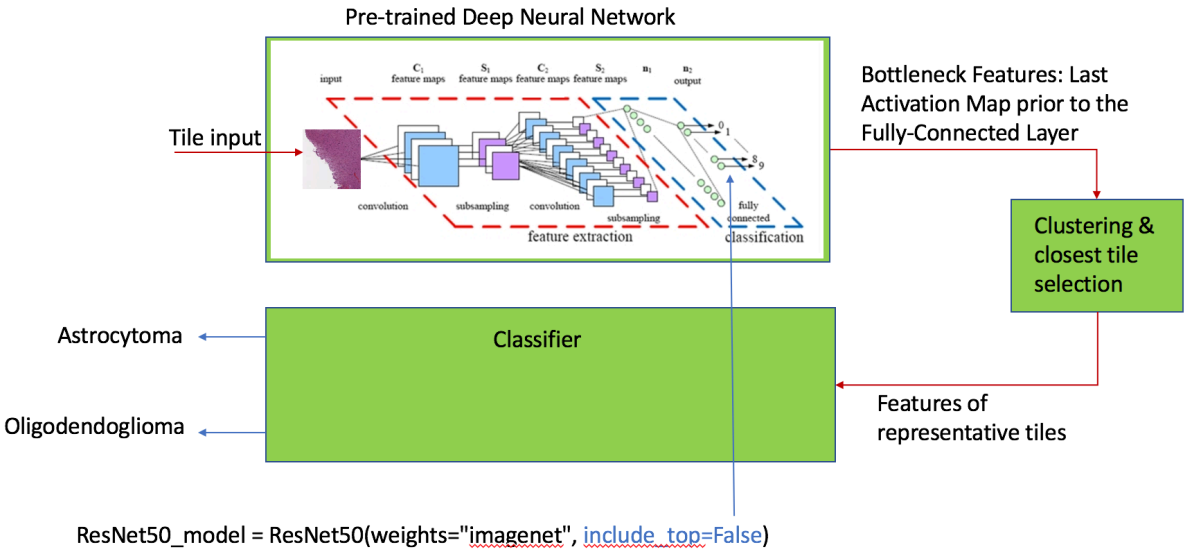


Figure. Transfer Learning: Feature extraction from bottleneck layer used for classifer training

Timing for feature extraction was of the order of 15 minutes per level 1 image. Blank tile removal substantially reduced this process to an overnight run on a Macbook Pro I7 with 16Gb of RAM for 32 level 1 images plus levels 2 and 3.

Fine tuning

In the second instance, fine tuning of the ResNet was performed. Fine tuning entails re-training of the upper couple of layers of the CNN with our own images in the hope that a richer, more relevant set of features can be extracted. A richer feature set should provide better modelling of similarities and differences between cancer types which might translate into better tile level results across images. However, the additional richness of feature set must be balanced against overfitting concerns particularly for a small dataset.

To avoid data leakage at this stage, fine tuning must be performed in conjunction with machine learning validation testing, ensuring that the network is only fine tuned with images not held out for testing. To in turn avoid impractically long machine learning runs including iterative feature extraction and clustering within the validation harness, tiles were selected ahead of time from each image in two ways 1) randomly and 2) based on clustering of generic (non-fine tuned) features.

Stage 3: Feature Clustering

Kmeans was used for the clustering stage. As shown on the right hand side of the

figure above, the feature vectors output from stage 2 were clustered and tiles closest to the cluster centroid selected and written to disk as a numpy array. Feature vectors of selected tiles could then be used to train a classifier to predict the class.

Optimal cluster numbers were explored through silhouette analysis which provides a graphical representation of the cluster density and separation in comparison to neighbouring clusters.

The purpose of clustering is twofold: One, to avoid the potentially intractable processing problem of using all tiles in all images. Instead, selecting a relatively small subset of representative tiles for the image both enabled image processing on a CPU and also allowed the as much variance of the image as possible to be preserved. And, two, to handle the issue of tile labelling.

Tile labels

Since the ground truth of tile labels is not known, in this study we labelled them according to the images from which they are taken. This approach is potentially problematic for tumors that exhibit a diversity of structures and textures spread across the image, or indeed a paucity of tumor cells spread across the image. Hence, the tile clustering and most representative tile selection technique has been adopted to address this issue as proposed by Barker et al [8].

Stage 4: Machine Learning

Classifiers (SVM with poly kernel, XGBoost, Random Forest) were trained and models tested to perform tile level predictions. Most tile predictions was used to determine image level classification.

Grid Searches using 5-fold cross validation were performed across the parameter space for each of the algorithms and final models determined. The parameters used to train classifiers are shown in the table below.

Table. Parameters used to train classifiers

Algorithm	Best parameters
XGB	{'reg_alpha': 0.1, 'learning_rate': 0.1, 'objective': 'binary:logistic', 'max_depth': 7, 'reg_lambda': 0.5, 'n_estimators': 200, 'gamma': 0.01}

RandomForestClassifier	{'max_depth': None, 'n_estimators': 1000, 'min_samples_split': 1}
Support Vector Machines (svm.SVC)	{'kernel': 'poly', 'C': 1000, 'gamma': 0.001}

Image Level Cross Validation

At an image level, a cross validation harness was established that allowed the choice of train/test split and rotation through all test images to maximize available training images. Leave 2 out cross-validation was used for larger runs equating to 16-folds on the 32 image CBTC set. More commonly, 4-fold cross validation was used for testing, particularly for the larger dataset including the TCGA images.

Validation Performance metrics

For the purposes of this paper accuracy of classification on the available test images is the overarching performance metric. Tile level accuracy is also an important indicator as is variance as defined in the following table.

Table. Validation Evaluation Measures

Image Level Accuracy	$\frac{correct\ image\ predictions}{all\ images}$	Overall proportion of correct image classifications
Tile Level Accuracy	$\frac{correct\ tile\ predictions}{all\ tiles\ in\ image}$	Proportion of correct tile classifications with an image
Variance	$s^2 = \frac{\sum_{\{i=1\}}^{n} \sqrt{(y_i - \bar{y})^2}}{n-1}$	Variance amongst readings. Square of standard deviation (s)

Stage 5: Inter-Imageset Testing

A testing framework was setup to test the CBTC images against a model trained with the TCGA images.

Test Performance metrics

In testing the entire CBTC dataset against models trained with the TCGA dataset, precision, recall and a balanced F1 measure were also considered in addition to accuracy and variance.

Table. Test Evaluation Measures

Image Level Accuracy	$\frac{correct\ image\ predictions}{all\ images}$	Overall proportion of correct image classifications
Tile Level Accuracy	$\frac{correct\ tile\ predictions}{all\ tiles\ in\ image}$	Proportion of correct tile classifications with an image
Variance	$s^2 = \frac{\sum_{\{i=1\}}^{n} \sqrt{(y_i - \bar{y})^2}}{n-1}$	Variance amongst readings. Square of standard deviation (s)
Precision	$\frac{correct\ predictions\ for\ class}{predictions\ for\ class}$	Proportion of predictions correct for class
Recall	$\frac{correct\ predictions\ for\ class}{number\ in\ class}$	Proportion of class predicted correctly
Balanced F Measure (F1)	$\frac{2\ (Precision\ .\ Recall)}{Precision + Recall}$	Combined effectiveness measure balancing precision and recall

Results and Discussion

MICCAI 2015 Challenge Dataset

Image Level Accuracy and Variance

The baseline average accuracy achieved in predicting Astrocytoma versus Oligodendroglioma was 68% with low variance as shown in the table below.

Table. Average Accuracy and Variance across 10 runs of 16 folds

	Overall Image Level	Astro Tile Level	Oligo Tile Level
Accuracy:	68%	53%	54%
Variance:	0.0006	0.0483	0.0319

This result was achieved across 10 runs with a cross validation harness using 16 folds rotating through test images across the folds. With 32 images, each fold held out 2 images and trained against the remaining 30. There are large run times

involved in training the model 160 times so other results in this study are generally with fewer runs, though generally fold numbers were kept as high as possible.

The tile level accuracy required to achieve this image level result was only of of the order of 50-55% on average, which is reasonable given the nature of two class classification problems. However, notice the high variance of the tile level accuracy scores within each class. A 95% tolerance interval would yield a range of 48.6 – 57.4% for Astrocytoma and 50.4 – 57.6% for Oligodendroglioma. It would appear that the average for the cross validation testing belies a large variation in scores *between images* within validation sets.

Overfitting during image level Testing

The following figure shows the high tile level accuracy achieved during grid search with 5-fold cross validation.

```
[CV]  n_estimators=200, objective=binary:logistic, gamma=0.01,
reg_lambda=0.3, max_depth=7, learning_rate=0.2, reg_alpha=0.1,
score=0.872798
```

Figure. Tile Level Accuracy achieved during grid seach with 5-fold Cross Validation (no images held out)

The fact that the classifier is able to achieve this level of accuracy indicates that it is able to discriminate signal across the tiles. However, the signal is sufficiently different at an image level to cause overfitting that reduces accuracy dramatically.

The table below shows average tile level accuracy derived during image level cross validation testing for:

- Tiles in the training images
- Tiles in the test images.

Table. Average Tile level Accuracy during 16-fold Image Cross Validation Testing

	Average Tile Level Accuracy	Test Harness
Train	78%	5-fold tile level Cross Validation
Test	53%	16-fold image level Cross Validation

The 78% average accuracy was achieved for tiles in training images in the validation set (5-fold cross validation). The 53% average accuracy was achieved for held out images, illustrating the dramatic overfitting being experienced. To better understand the substantial differences amongst images of the same labels, the accuracy for each individual image was investigated.

The table below shows accuracy per image across 10 runs. Images marked in red have less than 30% accuracy with a high degree of consistency.

Table. Tile level test accuracy for all images (10 runs, 16 folds)

Image Accuracy	0.69	0.66	0.69	0.69	0.66	0.69	0.66	0.69	0.66	0.66	6.72	0.0003
Run	1	2	3	4	5	6	7	8	9	10	Ave	Var
Astro:												
1	0.68	0.61	0.68	0.71	0.61	0.61	0.65	0.66	0.64	0.68	0.65	0.0004
2	0.64	0.59	0.59	0.59	0.61	0.63	0.67	0.61	0.67	0.64	0.62	0.0003
7	0.52	0.53	0.6	0.61	0.52	0.52	0.51	0.61	0.52	0.54	0.55	0.0005
8	**0.1**	**0.03**	**0.11**	**0.04**	**0.03**	**0.02**	**0.08**	**0.04**	**0.02**	**0.04**	**0.05**	**0.0003**
9	0.62	0.6	0.6	0.64	0.66	0.65	0.62	0.64	0.65	0.65	0.63	0.0001
11	0.59	0.59	0.59	0.7	0.7	0.6	0.76	0.76	0.68	0.68	0.67	0.0014
14	0.73	0.6	0.62	0.64	0.78	0.63	0.65	0.73	0.63	0.68	0.67	0.0010
15	0.67	0.61	0.64	0.66	0.57	0.64	0.55	0.63	0.52	0.61	0.61	0.0007
17	0.74	0.74	0.69	0.61	0.68	0.72	0.75	0.83	0.83	0.72	0.73	0.0013
18	0.36	0.41	0.44	0.49	0.41	0.45	0.43	0.44	0.49	0.27	0.42	0.0012
19	0.56	0.54	0.53	0.67	0.55	0.59	0.61	0.57	0.56	0.59	0.58	0.0005
23	0.54	0.56	0.44	0.49	0.46	0.46	0.45	0.52	0.51	0.46	0.49	0.0005
26	0.3	0.35	0.42	0.27	0.32	0.56	0.32	0.28	0.31	0.28	0.34	0.0023
28	0.55	0.75	0.67	0.56	0.58	0.56	0.55	0.56	0.55	0.56	0.59	0.0013
29	**0.29**	**0.18**	**0.19**	**0.22**	**0.19**	**0.2**	**0.16**	**0.19**	**0.19**	**0.26**	**0.21**	**0.0005**
31	**0.1**	**0.13**	**0.13**	**0.11**	**0.08**	**0.1**	**0.11**	**0.15**	**0.15**	**0.08**	**0.11**	**0.0002**
Oligo:												
3	0.78	0.73	0.77	0.77	0.75	0.81	0.78	0.87	0.72	0.72	0.77	0.0006
4	0.81	0.77	0.86	0.82	0.81	0.8	0.83	0.84	0.78	0.8	0.81	0.0002
5	0.7	0.62	0.62	0.75	0.7	0.65	0.5	0.71	0.68	0.7	0.66	0.0014
6	0.61	0.49	0.54	0.6	0.54	0.5	0.54	0.54	0.54	0.52	0.54	0.0004
10	0.8	0.75	0.75	0.73	0.67	0.79	0.76	0.69	0.77	0.79	0.75	0.0005
12	0.49	0.51	0.72	0.54	0.51	0.56	0.53	0.5	0.43	0.51	0.53	0.0016
13	**0.28**	**0.28**	**0.29**	**0.25**	**0.21**	**0.29**	**0.25**	**0.29**	**0.29**	**0.21**	**0.26**	**0.0003**
16	0.62	0.56	0.56	0.56	0.57	0.58	0.62	0.61	0.56	0.56	0.58	0.0002
20	0.61	0.47	0.54	0.55	0.49	0.49	0.43	0.48	0.42	0.47	0.50	0.0010
21	**0.27**	**0.24**	**0.35**	**0.31**	**0.35**	**0.26**	**0.35**	**0.35**	**0.34**	**0.31**	**0.31**	**0.0005**
22	**0.19**	**0.21**	**0.19**	**0.24**	**0.24**	**0.21**	**0.22**	**0.15**	**0.24**	**0.15**	**0.20**	**0.0003**
24	0.77	0.76	0.64	0.67	0.73	0.68	0.84	0.79	0.75	0.79	0.74	0.0011
25	0.75	0.69	0.79	0.68	0.76	0.7	0.81	0.79	0.72	0.73	0.74	0.0006
27	0.33	0.4	0.36	0.3	0.33	0.34	0.33	0.37	0.33	0.36	0.35	0.0002

30	0.61	0.57	0.64	0.53	0.6	0.57	0.6	0.62	0.57	0.58	0.59	0.0003
32	0.56	0.59	0.61	0.58	0.59	0.59	0.59	0.59	0.58	0.59	0.59	0.0000

This table alludes not just to histological differences between images, but to differences that bear a strong similarity to the opposite class. One would expect images not previously seen, non-classical shapes for example, to have probabilities close to 50%. Instead, we see here for highlighted images, the classifier confidently predicting opposite labels with low variance (consider for example images 8 and 31). Some possible explanations include:

1. Noise across classes, including irrelevant histological features or non-cancerous tissue in some tiles, distracting the classifier.
2. Signal in these images that are indeed similar to the opposite class. For example, mixed oligoastocytoma images might potentially cause the classifer to behave in this way or other structures such as microcalcifications which are typical of both cancer types, but not specific to either.
3. Features of a cancer subtype that are not always present and may or may not be present in the images in a small dataset
4. The generic CNN-derived features are not sufficiently rich to capture the complexity and heterogeneity of some LGG subtypes as suggested by Hou et al [21].
5. Incorrectly labelled images due to subjective histologic criteria used in the classification process [14] and [24].

Image Variation investigation

To view the image level variation more closely, a kind of intentional overfitting process was performed whereby images classified with “strongly incorrect probabilitites” (defined for this test to be those with less than 30% accuracy) were iteratively removed from the test set, leaving only “well performing” images (defined here to be images with accuracy greater than 30%). The following table shows results after two iterations of low performing image removal across 4 runs of 4 folds.

Table. Tile level test accuracy for “well performing” images (4 runs, 4 folds)

Run	**1**	**2**	**3**	**4**	**Average**	**Variance**
Image Accuracy	**0.92**	**0.92**	**0.92**	**0.88**	**0.91**	**0.0003**
Astro:					Astro Var:	0.0013

1	0.8	0.81	0.94	0.84	0.85	0.0031
2	0.81	0.77	0.81	0.79	0.80	0.0003
7	0.7	0.69	0.75	0.71	0.71	0.0005
9	0.83	0.76	0.76	0.76	0.78	0.0009
11	0.75	0.75	0.75	0.74	0.75	0.0000
14	0.87	0.91	0.9	0.91	0.90	0.0003
15	0.7	0.72	0.71	0.69	0.71	0.0001
17	0.89	0.89	0.92	0.94	0.91	0.0004
18	0.38	0.36	0.37	0.35	0.37	0.0001
19	0.64	0.62	0.66	0.8	0.68	0.0050
26	0.49	0.35	0.35	0.48	0.42	0.0046
28	0.88	0.88	0.86	0.87	0.87	0.0001
Oligo:					Oligo Var:	0.0025
3	0.96	0.91	0.95	0.91	0.93	0.0005
4	0.84	0.84	0.89	0.91	0.87	0.0010
5	0.71	-	0.64	0.55	0.63	0.0032
6	0.63	0.57	0.63	0.57	0.60	0.0009
10	0.74	0.74	0.79	-	0.76	0.0004
12	0.7	0.66	0.66	0.66	0.67	0.0003
16	0.53	0.54	0.62	0.49	0.55	0.0022
20	0.8	0.79	0.78	0.83	0.80	0.0003
24	0.84	0.93	-	0.95	0.91	0.0017
25	0.82	0.8	0.8	0.77	0.80	0.0003
27	0.63	0.62	0.6	0.63	0.62	0.0002
30	0.57	0.6	0.56	0.59	0.58	0.0002
32	-	0.59	0.61	0.71	0.48	0.0211

The final result shown in this table is 91% accuracy. This result was arrived at with removal of 3 Oligodendrioglioma images and 4 Astrocytoma images: 8,23,29,31,13,21,22. The reduction in variance throughout this process is also highlighted (in green).

In a hypothetical situation where these strongly incorrect images did not exist, the classification problem would be somewhat easier. Since these images are highly influential in the result, a better understanding of this group could lead to better insights into the problem. They were therefore gathered into an isolated set and input into the cross validation harness. The result is an accuracy of 83% as shown in the table below. This accuracy improves to 100% when image 8 highlighted (in red) is removed from the train/test set.

Table. Tile level test accuracy for "poor performing" images (4 runs, 4 folds)

Run	1	2	3	4	Average	Variance
Image	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	**0.0000**

Accuracy						
Astro:					**Astro Var:**	
8	**0.16**	**0.14**	**0.16**	**0.16**	**0.16**	**0.0001**
23	0.84	-	0.9	0.9	0.88	0.0008
29	0.94	0.94	0.93	-	0.94	0.0000
31	NaN	0.9	-	0.76	0.83	0.0033
Oligo:					**Oligo Var:**	
13	0.74	0.79	0.67	0.69	0.72	0.0029
21	0.59	0.59	0.65	0.65	0.62	0.0012
22	0.75	0.6	0.75	0.75	0.71	0.0056

As a result of this analysis, some observations can be made. This group of 32 images can reasonably be split into 5 subsets of images as follows:

2. Subset 1: [1,2,7,9,11,14,15,17,18,19,26,28] – Astrocytoma
3. Subset 2: [3,4,5,6,10,12,16,20,24,25,27,30,32] – Oligodendroglioma
4. Subset 3: [23,29,31] – Astrocytoma
5. Subset 4: [13,21,22] – Oligodendroglioma
6. Subset 5: [8] – Astrocytoma

Subsets 1 and 2, comprising 25 of the original CBTC 32 images, are distinct and readily able to be differentiated by the classifer. Based on this result, it would appear that the generic CNN-derived features are in fact able to capture the richness of these particular images and distinquish them with 91% accuracy as shown in tables above. This conclusion is supported by the high accuracy of the tile level cross validation set. Of interest is the observation that 25/32 represents a 78% success rate on the original 32 image CBTC image set, not dissimilar to the 75% accuracy achieved by Yang et al [9]. It may be possible that this high water mark actually sets an upper limit to the accuracy possible with this set.

Subsets 3 and 4 have characteristics that, tested against subsets 1 and 2, cause stongly incorrect predictions by the classifier. However, when tested together, the classifier is able to distinquish them with 100% accuracy.

Subset 5, slide 8, is strongly predicted to belong to the opposite class in both cases.

Mixed OligoAstrocytoma

To test the idea that the latter three subgroups may be mixed Oligoastrocytoma images, 43 images from the TCGA site *labelled as mixed Oligoastrocytoma* were tested against different subsets of the 32 CBTC images. The tables below show

results with two different training sets:

1. Subset 1 and Subset 4: “Well behaved” astro images and “badly behaved” oligo images
2. Subset 3 and Subset 2: “Badly behaved” astro images and “well behaved” oligo images

Table. Test of Mixed Oligoastrocytoma images against models trained with subsets 1,4 and 2,3, respectively

Mixed Oligoastrocytoma image predictions	Training set	
	Subset 1 (astro) Subset 4 (oligo)	Subset 3 (astro) Subset 2 (oligo)
Subset 1	43	
Subset 4	0	
Subset 3		0
Subset 2		43

The results of this test are definitive in showing that the “badly behaved” images in the CBTC set cannot reasonably be considered to be mixed images. In both cases, the actual mixed images were predicted to be the “well behaved” images.

Neuropathology

There are several different sets of histological characteristics that could be identified by a neuropathologist in making an LGG diagnosis.

Neuropathologist, Michael Buckland, was consulted to examine the labelling of our images and discuss diagnosis complexity [24]. He noted a number of borderline diagnoses within our dataset, particularly amongst Oligodendroglioma cases, and importantly identified different histological featues in developing a diagnostic view for images of the same class. For example, the following diverse set of features were all observed in our image set, not necessarily together in the same image and not in all cases:

1. Astrocytoma:
    - Nuclear pleomorphism
    - Relative nuclear hyperchromasia
    - Abundant eosinophilic cytoplasm

- Glial processes.

2. Oligodendroglioma:
    - Relative roundness and monotony to nuclei
    - Stippled chromatin
    - Perinuclear halos
    - Calcification
    - Chicken-wire vasculature

In addition to borderline diagnoses, there was one slide in particular, slide 4 shown below, diagnosed as mixed oligoastrocytoma including both distinct astro and oligo features.

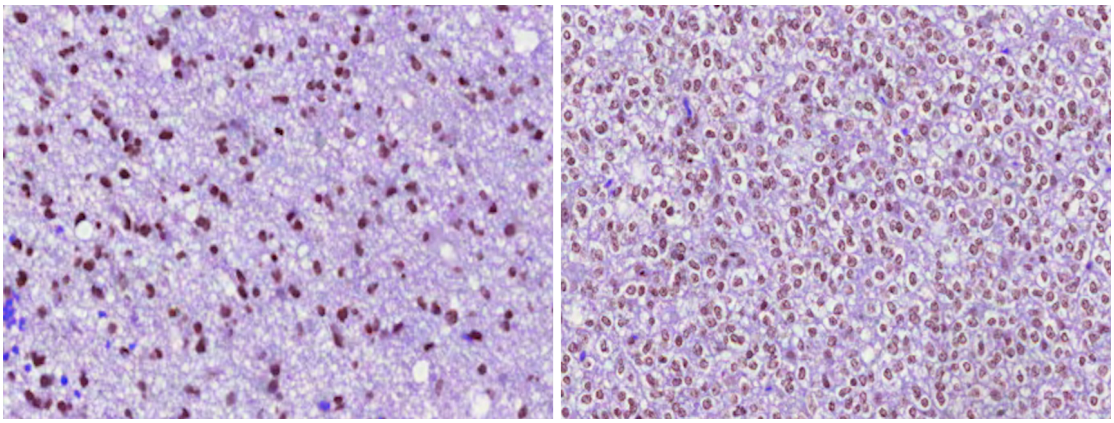


Figure. Slide 4 is Oligoastrocytoma. Different portions of the slide demonstrate distinct Astrocytoma (left) and Oligodendroglioma (right)

There were other slides of quite poor quality that may have obfuscated regularity in our images, especially for oligodendroglioma (which could result in oligo images being incorrectly labelled as astro).

This feedback served to support the thesis that multiple subgroups of each LGG subtype exist in our testing, therefore accounting for the overfitting experienced. It also served to support the idea that an accuracy of greater than 75% may be unlikely on this image set.

CNN Fine Tuning

CNN fine tuning was performed on the original CBTC dataset and cross validation testing again performed. The thesis being that fine tuning the ResNet with

histological images should deliver feature richness more relevant to these images and thereby potentially deliver better image predictions. However, only chance level accuracy was achieved.

Results were heavily impacted by overfitting which may indicate that the dataset is too small to reasonably make use of fine tuning [22]. The tables below show overfitting experienced after fine tuning the top two blocks of the ResNet50 (19 layers) over 3 epochs. Notice very high training accuracy but a consistently low validation accuracy. Reducing the number of layers to 2 and epochs to 2 improved the validation accuracy, as shown in the second table below, but the image level result was no better than chance.

Table. Tuning Top 2 blocks (layers 31-49) of ResNet50 on CBTC dataset with 4 epochs

Epoch	Train accuracy	Train loss	Validation accuracy	Validation loss	Image accuracy
1	0.89	0.25	0.49	1.63	
2	0.92	0.21	0.49	1.53	
3	0.93	0.19	0.49	1.33	
4	0.94	0.18	0.49	1.10	.49

Table. Validation accuracy after tuning layers 48-49 of ResNet50 on CBTC dataset with 2 epochs

Epoch	Train accuracy	Train loss	Validation accuracy	Validation loss	Image accuracy
1	0.78	0.45	0.79	.44	
2	0.81	0.42	0.79	.43	.51

The table below shows CNN validation results on the larger TCGA dataset with only layer 49 being retrained. The overfitting is clearly addressed at a tile validation level but the image level result still did not improve on chance.

Table. Tuning layer 49 of ResNet50 on TCGA dataset with 3 epochs

Epoch	Train accuracy	Train loss	Validation accuracy	Validation loss	Image accuracy
1	0.88	0.26	0.64	1.03	
2	0.90	0.24	0.83	.35	
3	0.90	0.23	.85	.30	.51

Future research could focus on fine tuning the ResNet including addition of dropout layers and using data augmentation techniques. It should ideally also make use of a

much larger image set. Furthermore, given the dissimilarity of histological images to the ImageNet image set, another possibility could be to test by extracting features from a lower layer in the Resnet50. In this way, more primative CNN features could be used for testing and may be more appropriate than those in higher layers of the ResNet.

Clustering Effectiveness

Clustering was used as a tile selection technique to avoid processing all tiles from within the image and to retain maximum possible feature variance. Given that objective, Kmeans clustering was performed on tile feature vectors derived from ResNet50 and closest tiles per cluster selected and saved to disk as a numpy array.

The number of clusters (n_clusters) was chosen to be 10 as a suitable number to maximise the likelihood of sufficient image feature variance. The 10 tiles closest to the cluster centroids were selected per cluster. Therefore, 100 tiles were selected per image.

Different cluster numbers were investigated between 2 and 10 and silhouette scores per cluster were derived. Average silhouette scores for different values of n_clusters were found to be highest for lower numbers of clusters and gradually declined as the value of n_clusters was increased. The figures below shows silhouette diagrams and plots against the first two principal components for n_clusters values of 2 and 10 for one of the Astrocytoma images.

Figure. Silhouette analysis for Image 1 with two clusters. The average silhouette_score is : 0.653355

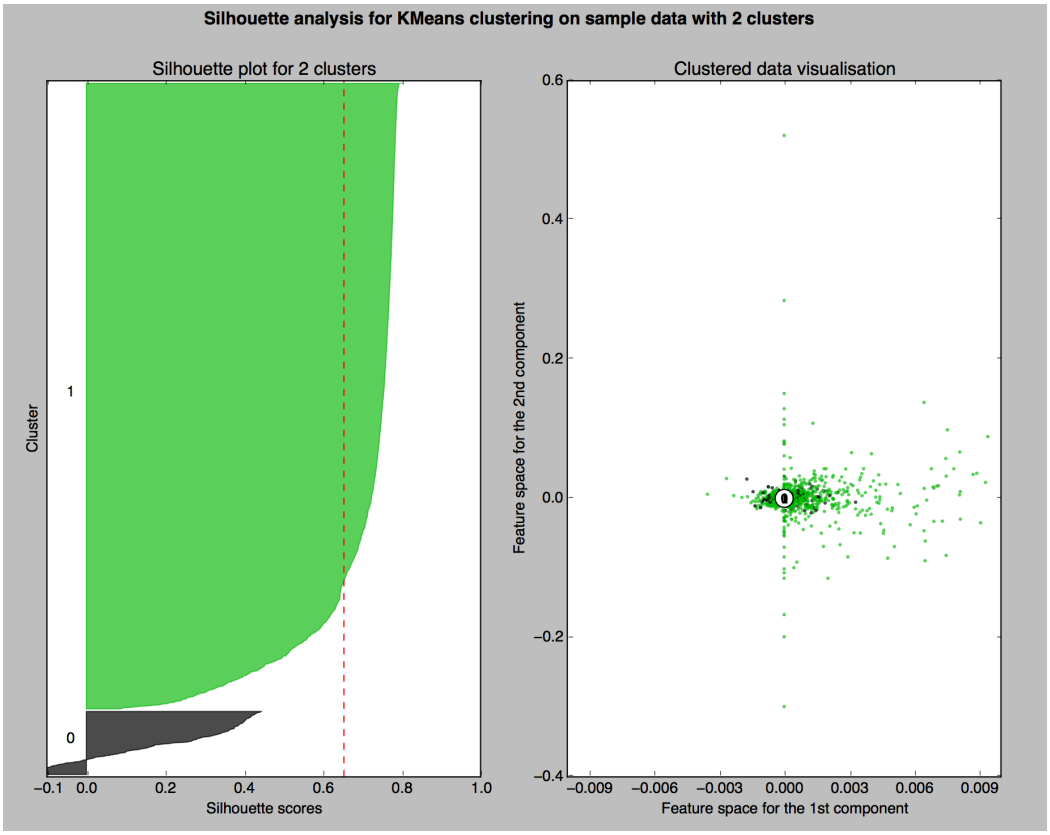
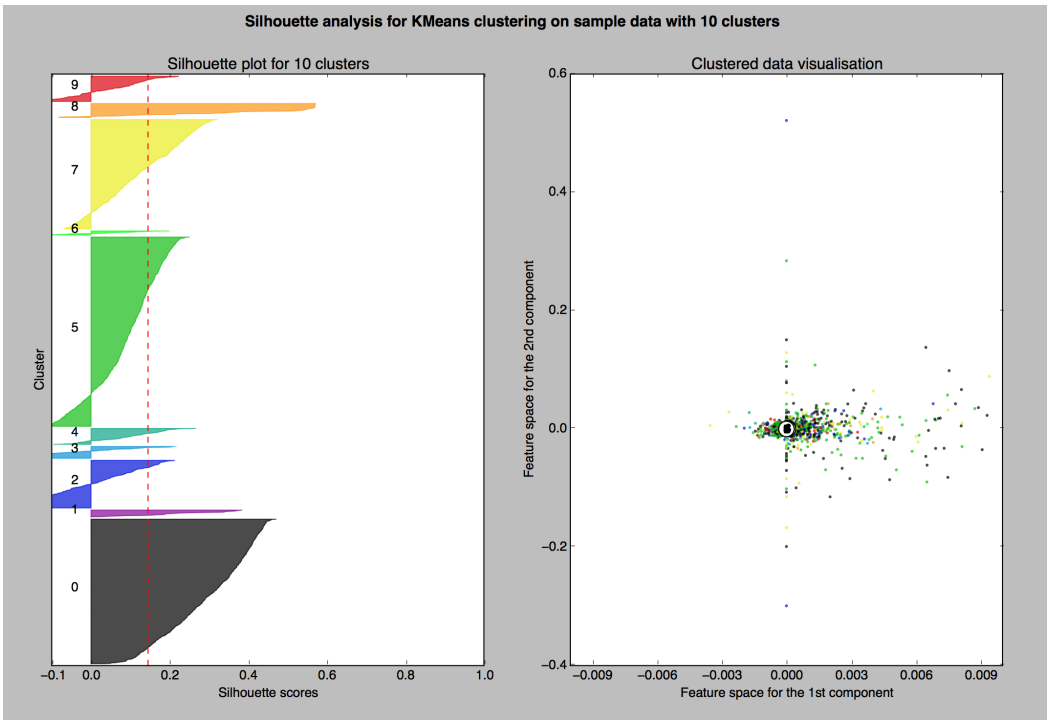


Figure. Silhouette analysis for Image 1 with ten clusters. The average silhouette_score is : 0.144996



The figures above show silhouette diagrams with a low, positive average, non-uniform clusters with values ranging from over .6 to negative .1. Additionally, the feature space of the principal components shown in the right panel of the figures above is not particularly well utilized. The conclusion from this diagram, and others like it for other images, implies that the clusters are relatively poorly separated.

It is arguable therefore that although clustering served its purpose as a means of overcoming an intractable processing problem, it may not have served its purpose in terms of its role in tile labelling and maintaining image variance.

Cross validation testing was performed to examine the value of clustering relative to random tile selection. The test made use of 100 randomly selected tiles per image rather than 100 tiles closest to cluster centroids. The results are shown in the tables below.

Table. Tile selection through clustering (10 runs, 16 folds)

	Overall Image Level	Astro Tile Level	Oligo Tile Level
Accuracy:	67%	53%	54%
Variance:	0.0006	0.0386	0.0319

Table. Random tile selection (10 runs, 16 folds)

	Overall Image Level	Astro Tile Level	Oligo Tile Level
Accuracy:	63%	48%	60%
Variance:	0.0011	0.0483	0.0185

Based on these results, clustering improves accuracy by a small margin over random tile selection. However, these results are within a tolerance of 2 standard deviations from each other and so don't represent a statistically significant difference at a 95% confidence level.

Of note is that a small improvement in performance for astrocytoma images for clustering outways a substantial performance decease for oligodendroglioma images. A marginally better result overall is produced possibly due to the improvement in borderline Astrocytoma predictions. It is also arguable that clustering is working better for Astrocytoma perhaps due to greater heterogeneity of Astrocytoma histology relative to Oligodendroglioma.

Testing was performed with different values of n_clusters but no performance improvement was noted. Principal Component Analysis was also tested with no difference in performance noted.

Additionally, during the course of this study clustering was attempted as a blank tile removal strategy, and this seemed to be effective. The blank tile cluster had very high silhouette scores, meaning they could be readily removed, and visual inspection of the tiles in this cluster confirmed they were blank. However, the time taken to extract features using the ResNet50 CNN was increased substantially with this approach and it proved more time efficient to remove the tiles from consideration prior to the feature extraction stage.

One final observation related to clustering is the commonplace occurrence of a single cluster with a high silhouette score (8 in the figure above). Similarly, there was usually a single large cluster (0 in the figure above). It is possible that large or well formed clusters may include either highly valuable tiles or indeed valueless tiles, those not including cancer signal, for example. Testing was performed removing these clusters with no improvement noted. Future work studying the relationship between tile clusters and histological features would be of immense value, particularly in collaboration with neuropathologist researchers. Hou et al [21] introduced a novel approach to addressing the tile labelling problem that could also be considered in future work.

The Cancer Genome Atlas (TCGA) Dataset

In an effort to overcome the overfitting experienced at an image level, an additional 104 images were obtained from the TCGA site, processed and features extracted making use of ResNet50 as outlined for the CBTC images. Six small images were removed since the number of tiles in these images was lower than the number of tiles beng used for testing. The combined dataset of images for testing in this way became 130.

The results for validation testing on this 130 image dataset are shown in the table below. The results shown are for a single run of 4 folds.

Table. Average Accuracy and Variance across 1 run of 4 folds

	Overall Image Level	Astro Tile Level	Oligo Tile Level
Accuracy:	54%	52%	50%
Variance:	0.0366	0.0342	0.0362

An extra 104 images did not address the overfitting problem, in fact it may have made it worse. It is plausible that even more image variability has been introduced with these images which offsets the benefit of the additional data. A natural follow-on to this study is to systematically investigate the impact of overfitting on results as image quantities are incremented.

As described for CBTC images, images with strong, incorrect predictions were removed through at iterative process. The results in columns 2 and 3 of the table below were derived for the new combined image set.

Table. Average Accuracy and Variance across 2 runs of 10 folds

	Full 130 image set (as above)	Reduced "Well performing" 102 image set	"Poor performing" 28 image set
Accuracy:	54%	79%	88%
Variance:	0.0366	0.0000	0.0006

Once again, removal of a handful of images dramatically improves performance and reduces variance.

Performance Measurements of Inter-Imageset Testing

The additional image set also made it possible to hold out the CBTC image set for testing while using the TGCA set to train the model. Results are shown in the following table.

	Accuracy	**Precision**	**Recall**	**F1 Score**
Overall	62.5%			
Astrocytoma		.75	.375	.5
Oligodendroglioma		.58	.875	.7

As shown in the table below, removing poor performers from the TGCA *training set* marginally improves performance on the CBTC test set by 3 percentage points but also enhances astrocytoma recall and oligodendroglioma precision. Most likely, this is due to the removal of Astrocytoma images being strongly and incorrectly predicted as Oligodendroglioma. The approach of removing images with high probabilities for the opposite class is not proposed as a general technique to improve results. However, it serves in this case as a substitute for a lot more data and for focusing research.

	Accuracy	Precision	Recall	F1 Score
Overall	65.6%			
Astrocytoma		.73	.5	.59
Oligodendroglioma		.62	.81	.7

Discussion and Future Work

In this study, we have conducted a set of experiments that serve to shed light on relevant questions with regard deep learning and its use in classification of low grade glioma in whole slide images.

CNN-derived features

Firstly, we demonstrated that transfer learning from a CNN (ResNet50) can be used successfully to derive features in a completely automated manner from whole slide images, and classify those images with a comparable level of success to other studies of LGG classification. The average result of 68% accuracy on the CBTC dataset was of the same order of magnitude as other studies making use of this same dataset (Yang et al [9] achieved 75% accuracy while Carlson [10] achieved 68.75%).

The lack of success in driving greater degrees of accuracy may indicate that generic transfer learning of CNN-derived features may not be sufficiently rich to capture the complexity and heterogeneity in signal amongst images. Hou et al [21] indeed believe this to be the case. Additional research could focus more on fine tuning CNN features and/or combining them with hand crafted features in an attempt to improve accuracy of classification in this difficult area. However, the benefit of automatically-

derived, objective CNN-derived features cannot be over-stated in this field. We believe that, longer term, hand crafted features will need to be eliminated in order to in turn eliminate the subjectivity and effort involved in LGG classification.

Additionally, given the extreme overfitting experienced with fine tuning, an opposite tack could be taken in exploring generic features extracted from a lower layer in the CNN as proposed by Kaparthy [22].

Overfitting

Indeed, the main challenge to present itself in this study did not appear to be feature richness, rather it was overfitting. Of immense interest was the dramatic reduction in test tile accuracy in comparison with training tile accuracy. The high levels of training tile accuracy lend support to the idea that even generic CNN features are able to differentiate signal given enough data. We believe that more data than the additional 104 images sourced from TCGA could solve the non trivial challenge of variability between images of the same class. It may be that additional images at first do more harm than good by increasing image variability until sufficient images are added, particularly if those images contain numerous borderline diagnoses. A promising avenue of future research could entail attempting to address overfitting with much more data made up of only classical histological shapes of LGG subtypes. Data augmentation techniques may be useful in assembling the required dataset size. Once a critical mass is achieved, less than classical shapes can be added progressively.

Image Variability and Strongly incorrect predictions

The detailed logistics of image level variation showed that a number of images were strongly and consistently predicted to belong to the opposite class. Removal of these images from the primary dataset did, as would be expected, eliminate classifier confusion with subsequent very high accuracy (91% cross validation accuracy on the CBTC dataset).

Though not suggested as a general technique to improve prediction accuracy, this approach served its research purpose to explore the overfitting problem. The approach exposed clear subgroups within classes that were confirmed by neuropathologist review. The clear solution to addressing this variability is more data and definite labels.

Of value in ensuring correctly labelled images could be leveraging molecular profiling in addition to histological features. In particular, future work could benefit from consideration of identification of 1p/19q codeletion in addition to histological features to ensure reliable Oligodendoglioma labelling.

Clustering and Tile Labelling

Another avenue for future work could be clustering which was shown to perform only marginally better than random tile selection. A more detailed investigation into feature vector clusters and their relevance to histological features could be valuable as would implementing the novel approach to overcome lack of ground truth for tile labels proposed by Hou et al [21]. Such studies may also benefit from visualizing the CNN-derived features and mapping them to histological features of LGGs.

Conclusion

Based on the results achieved in this study, the potential of CNN in the field of Low Grade Glioma appears to be strong. The result of 68% LGG subtype classification accuracy on the small, balanced dataset in this study was achieved without any hand crafted feature engineering and was consistent with other non-CNN based studies on the same dataset. A strong case was made for future work to engage with a much larger image set and to investigate histological mappings of CNN features and clusters.

Finally, it should be kept in mind, that the subjective nature of typing and grading of LGG may set limits to success at this point in time. The literature is rife with commentary regarding the subjective classification and high inter-observer disagreement experienced in relation to low grade glioma. Working alongside a pathologist in this work could therefore be highly productive. Furthermore, as results from Convolutional Neural Networks improve in this area, the real value may lie in adding objectivity to a field currently struggling with subjectivity.

References Sited

[1] "Brain tumors: an introduction". *Mayfield Clinic, Brain & Spine, 2016.*

[2] Pieter Wesseling, Johan M. Kros, Judith W.M. Jeuken. "The pathological diagnosis of diffuse gliomas: towards a smart synthesis of microscopic and molecular information in a multidisciplinary context". *Diagnostic Histopathology, Volume 17, Issue 11, Pages 486–494, 2011*

[3] Dimitri P. Agamanolis M.D. "Neuropathology: An illustrated interactive course for medical students and residents, Tumors of the Central Nervous System". *North East Ohio Medical University, 2017*

[4] Marion Smits, M.D., PhD. "Imaging of oligodendroglioma". *Br J Radiol. April 2016; 89(1060), 2016*

[5] Nader Pouratian, David Schiff . "Management of Low-Grade Glioma". *Curr Neurol Neurosci Rep. 10(3): 224–231, 2010*

[6] "Brain Tumor Information, Oligodendroglioma and Oligoastrocytoma". *American Brain Tumor Association, 2014*

[7] Jocelyn Barker, Assaf Hoogia, Adrien Depeursinge, Daniel L. Rubin. "Automated classification of brain tumor type in whole-slide digital pathology images using local represenative tiles". *Med Image Anal. 30:60-71 2016*

[8] Fabio A. Spanhol, Luiz S. Oliveira. "Breat Cancer Histopathological Image Classifation using Convolutional Neural Networks". *IEEE Neural Networks (IJCNN), 2016*

[9] Yang Song, Fan Zhang, Weidong Cai. "Tumor Classication from Digital Pathology Image with Patch-based Local Approximation". *CBTC, MICCAI, 2015*

[10] Joel Carlson. "Combined Radiology and Pathology Classification". *CBTC, MICCAI, 2015*

[11] Wikipedia: Oligodendroglioma

[12] Ceccarelli et al., "Molecular profiling reveals biologically discrete subsets and

pathways of progression in diffuse glioma". *Cell. 164(3): 550–563, 2016*

[13] Figarella-Branger, Bouvier C., "Histological classification of human gliomas: state of art and controversies". *Bull Cancer. 92(4):301-309, 2005*

[14] Kolles H, Niedermayer I, Feiden W., Pathologe. "Grading of astrocytomas and oligodendrogliomas". *NCBI. 19(4):259-268,1998*

[15] Jansen et al. "Prediction of oligodendroglial histology and LOH 1p/19q using dynamic [18F]FET-PET imaging in intracranial WHO grade II and III gliomas". *Neuro-Oncology 14(12):1473–1480, 2012*

[16] Coons, Johnson PC, Scheithauer BW, Yates AJ, Pearl DK., "Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas". Cancer. 79(7):1381-93, 1997

[17] Stephen B. Tatter., "The new WHO Classification of Tumors affecting the Central Nervous System". *Neurological Service, MGH*

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. "Deep Residual Learning for Image Recognition". *Cornell University. arXiv:1512.03385v1, 2015*

[19] Yan-Nung Chen and Mark Chang., "Applied Deep Learning: Convolutional Neural Networks". *Linkedin Technology, 2016*

[20] Martin J. van den Bent., "Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective". *Acta Neuropathol. 120(3): 297–304, 2010*

[21] Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, Joel H. Saltz. "Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification". *Cornell University. arXiv:1504.07947, 2015*

[22] Andre Kaparthy. "Lecture notes for: CS231n Convolutional Neural Networks for Visual Recognition". *Standford University, 2016*

[23] Anusua Trivedi. "Deep Learning Part 2: Transfer Learning and Fine-tuning Deep Convolutional Neural Networks". *Data Revolutions, Microsoft, 2016*

[24] Private consultation with Michael Buckland, Clinical Associate Professor Pathology, School of Medical Sciences, The University of Sydney., 2017

[25] Louis et al., "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary" *Acta Neuropathol, Volume 131, Issue 6, pp 803–820, 2016*

[26] David N. Louis et al., "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary". *Acta Neuropathol, Springer, 2016*

Appendix 1: Computing Platforms

All processing was carried out on a Macbook Pro with the following resources:

- MacBook Pro (Retina, Mid 2012)
- Processor: 2.3 GHz Intel Core i7
- Memory: 16 GB 1600 MHz DDR3
- Disk: 2T+

Appendix 2: Instructions for accessing and running code

The code for this study is contained in a Jupyter notebook Masters_Project_Notebook.ipynb. A zipped version of the notebook has been uploaded to Github and can be downloaded by clicking on this link:

https://github.com/tijohnso/Usyd_masters/raw/master/Masters_Project_Notebook.ipynb.zip

The code is written as much as possibl so that decisions can be taken in interactive cells such as path specification, option settings, image to be processed and so on, without effecting the operation of the functions.

Steps to run:

1. Unzip the submitted file which contains the Jupyter notebook
2. Set the *images_root_dir* directory. This is the root directory for all images referenced within the code. Subdirectories *level0*, *level1*, *level2*, *level3* will be created in this root directory and respective Tiff files of relevant SVS levels will be stored here. With current defaults (levels = [1]), only level 1 will be used.
3. Set the *results_dir* directory. This is used for saving results.
4. It is assumed that .svs files for CBTC and TCGA images are residing in locations specified in the cell entitled “Source SVS Files and Directories”. Overwrite the relevant paths as applicable for:

   CBTC_A = '/Volumes/2T_HD/Masters_Project/astrocytoma/*.svs'

   CBTC_O = '/Volumes/2T_HD/Masters_Project/oligodendroglioma/*.svs'

   TCGA_raw_data_dir = '/Volumes/2T_HD/Data/'

   TCGA_dirs = '/Volumes/2T_HD/Masters_Project/data/'

   patient_info = "/Volumes/2T_HD/Masters_Project/data/Patient_info.txt"
5. Step through the notebook from the top – bearing in mind some cells may not need to be run
6. The cell “TCGA: Copy Raw images into patient directories” creates directories with name of the first 12 characters of the raw image file. Eg: “TCGA-HT-7902”. If you have already processed raw TCGA images into patient directories with this name, you will not need to run this cell.
7. Some of these cells take a long time to run. You should be able to execute all cells in the first 3 stages at one time and wait for them to process (e.g. overnight).
8. Once each of stages 1, 2, 3 have been run, required numpy arrays will have been saved to disk. Cells should not need to be run again (unless for example you want to use different cluster numbers for example).
9. Some of the results in the project report require interactive use of the code, changing image indexes and so on to perform different tests. Functions should not need to be changed.