Automated Gleason Score Prediction from histopathology Images of Prostate Cancer Micro Arrays

Tushar Kataria

***Abstract*—Histopathology images remains one of the leading prognostic markers for diagnosis, tracking and treatment of Prostate Cancer. Prostate cancer tissue micro arrays with H&E (Hematoxylin and Eosin) staining images are used to estimate the prevalance of the diesease in the patient. Stages of prostate cancer are differentiated by Gleason Grading given to a patient's sample by the pathologist. Grading a sample requires highly trained pathlogist and due to the subjective nature and heterogenity of the cells present in the images, there is significant variability amongst pathologist accessments for a single patient. In recent years deep learning has shown significant improvement in analysis of medical images. In this study we will be applying deep learning to automate Gleason Grade scoring. Two paths were explored :- 1) Segmentation of Patches 2) Classification of Patches of the sample to different grades. Code available at GitHub Directory. Gleason Score project.**

I. Introduction

Prostate cancer is the second leading cause of cancer deaths in America. There is a 12 percent chance that a person is going to be diagonsed with prostate cancer in their life time. So accurate diagnosis, tracking disease progression and plan of treatment are very important. Histopathology images is the leading prognostic marker for diagnosing prostate cancer but due to the intra-variability amongst pathologist accessment of a patients samples, misdiagnosis and mistreatment are possibilities which cannot be ruled out.

Gleason score for each cell or pattern is given on a scale from 1(well diffrentiated) to 5(poorly differentiated). The final gleason score to sample is given by adding scores of 2 most predominant(Primary and Secondary) cells/tumors found in the sample. So for example if a patients sample as cell patterns of only gleason grade 3, then gleason grade will be 6, but if the patient has cell pattern of 3 and 5, then gleason score is 8. Detecting cell sections in the slide for different gleason score decides the final Gleason Grade. Based on the final Gleason Score, there are five grades for progression of the disease:-

1) Grade 1: Gleason Score 6 or lower, low grade cancer, Cancer might grow slowly if it grows at all.
2) Grade 2: Gleason score 3 + 4 = 7 , medium grade cancer, Cancer might grow slowly.
3) Grade 3: Gleason score 4 + 3 = 7 , medium grade cancer, Cancer might grow at a moderate rate.
4) Grade 4: Gleason score 8 , high grade cancer, Cancer might grow moderately but has a potential to grow faster.
5) Grade 5: Gleason score 9 to 10, high grade cancer. Cancer likely to grow faster.

With the above description it is clear that detecting gleason score 3,4 and 5 are very important and, differentiating them because a difference of one grade can mean difference between medium grade and high grade cancer. In this study we will explore some possible directions for automating the Gleason Grading of a patients sample. Because of the huge size of the training images(approximately 5120 x 5120) and very few training samples, any segmentation model for this input size will have a huge number of parameters and very hard to train on a 4GB graphics card which is available on CADE Machines. Downsampled images might be trainable but downsampled images might loose details which might be useful in differentiating between different gleason score. So different techniques were tried to solve the problem which can be divided into 2 categories:-

1) Segmentation Type Techniques :- 256x256 patched from training set were sampled with and without overlap, sampling the mask's at the same stop as well. These resulted in creating thousands of samples of 256x256 rather an a few hundred of size 5120x5120. U-NET [5] type architectures for different number of initial channels were trained. Some Extentions to U-NET architectures were also trained like with resnet-18 and resnet-34 acting as encoder backend of the U-Net architecture. Other extension and detail experiments will be listed in the section III.
2) Classification Type Techniques :- As we discussed in the section above it is very important to find patches on the slides with specific gleason score. It's not that important that we find the exact demarcations of each type of cells present in the slide, because the final Gleason score depends on the most prominent cells present. So Training a classification model to recognize patches of each type of cell will also provide a reliable estimate of the final Gleason score. This type of technique was tried in [2]. More Details in Section IV.

Due to the limitation of the Graphics Card size 4GB, most of the dataset's created for experiments(as explained above) can only run for a few epochs, sometimes for segmentation datasets it couldn't even complete 1 epoch and run into resource constraint. So a distributed training was used to train the model. Any dataset which had huge amount of input data was split in multiple training and validation subsets. After training on one subset the model and optimizer states were saved and then reloaded again for training on different subsets. This was done for all segmentation experiments and Classification experiments with non-zero overlap.

In Section II, the datasets used and data preprocesing for segmenationa and classification are explained. Section III explaines all the segmentation experiments and results. Section IV details all the classification experiments and results. Section V is a dicussion on Future Work.

II. Datasets

Primarily 2 datasets were used for experimentation. 1) Gleason Score Grand Challenge Dataset [4] [6] 2) Harvard Dataset [1]

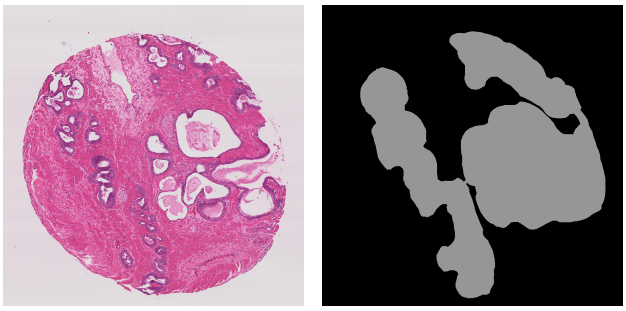
A. Gleason Score Grand Challenge Dataset:- Dataset I

We have 244 traing images of varying sizes but the most common size of the image is 5120 x 5120. We have segmentation mask count from each pathologist 1 to 6 are 244, 141, 242, 244, 246, 65 respectively. Each pixel is gleason Score between 1 to 5. For this dataset most of the pixels are given gleason score of 1,3 and 4. There are almost no mask with pixel value 2 and there are very less examples of gleason score 5. Masks with pixel value 0 are background. On the challenge page it is advised to create a single mask using all these mask using majority voting. But there are a few images where none of the pathologist agree on the gleason segmentation resulting in bad final masks with every pixel treated as backgound. We have removed these samples from the dataset because these samples are not reliable and can cause confusion for the trained model. An example of the dataset is shown in Figure 1.
To get an estimate of how much agreement is there between pathologist for segmentation present, we found out the average cohen's kappa coefficient between each pathologist over all the training samples common between them (Cohen kappa also includes background label). Scores are listed in the table below I. Cohen kappa are symmetric, so table doesn't need to be filled fully.

Pathologist	2	3	4	5	6
1	0.223	0.452	0.52	0.49	0.50
2	-	0.23	0.266	0.269	0.259
3	-	-	0.497	0.568	0.447
4	-	-	-	0.587	0.578
5	-	-	-	-	0.55

TABLE I: Cohen's Quadratic Kappa Coefficient for Grand Challenge Dataset

We can clearly see from the table I that there is significant variability between pathologist accessment of a single sample. This is why automating gleason score prediction is important that will remove the personal bias from the output prediction, giving reliable estimate of the Gleason Grade. For this dataset testing segmentation mask or gleason score of the slide are not available. So for all the experiments val dataset values metrics will be reported. I tried submitting results on Gleason Grade page but it displayed no results on the leaderboard.

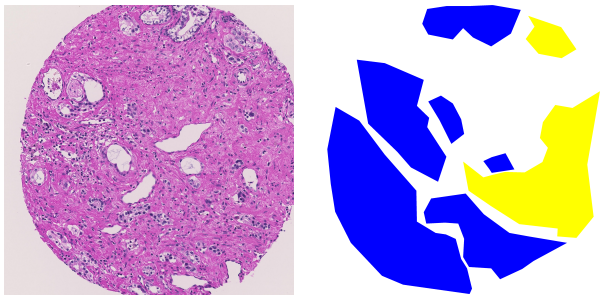


(a) Image (b) Mask Image

Fig. 1: Training Sample from Dataset I

B. Harvard Prostate Cancer dataset: Dataset II

The second dataset is Harvard prostate cancer dataset [1] which has 641 training samples with their segmentation mask. Test set consist of 241 samples with segmentation mask obtained from 2 pathologist. All the training and testing images are of size 3100 x 3100. Cohen kappa agreement for the 2 annotators segmentation mask is 0.52538381. For this dataset 2 Testing segmentation masks are available for different pathologist. Testing data was used only once at the end of the experiments to check metrics so as not to bias any hyperparameter search. More details on the dataset can be found in [2]. An example of the dataset is shown in Figure 2.



(a) Image (b) Mask Image

Fig. 2: Training Sample from Dataset 2

C. Data preprocessing for Segmentation

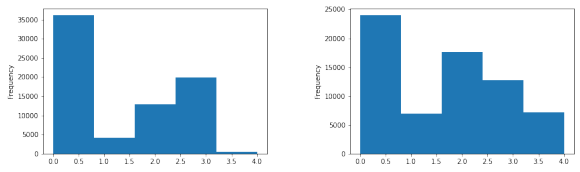
2 Dataset were created using this big dataset. Image Patches and corresponding mask of size 256x256 were sampled from training images with a overlap of 0 and 128 pixels . A mask label was given to each sample, label of most frequent pixel value in mask. These label value was used to create Training and Validation sets using stratified sampling. Stratified sampling was used to create Training and Validation sets with similar imbalance of each class. All the sampled images created were saved as '.png' images so as to create a lossless dataset.

D. Data Preprocessing for Classification

Patches for size 256x256 were sampled from training images with a overlap of 0, 128 and 192 pixels . If the mask of the patch had more than 95% pixels of the same grade or score, it was kept in the training set and assigned that grade, all other patches were discarded. The histogram of resulting labels for the dataset thus created is shown in figure 3. As we can clearly see a large number of samples are background patches. Label 5 had very less number of samples in the training set. With overlap, the number of samples for each class increased but the ratio still remained similar. DataSet II is more balanced than Dataset I. Training and Validation sets were created with a 80-20 split usind stratified sampling, so that both training and validation sets are equally imbalanced.
Now using the above sampling technique and overlap one might say that validation sets have a higher chance of seeing similar patches from the same slide, so these models might not generalize well to unseen data. So 2 different datasets were also created where training and validation sets have no common slides between them.

III. Segmentation Experiments

The most popular segmenation model for medical datasets is U-Net [5]. This model was compared with a few variations in



(a) Dataset I (b) Dataset II

Fig. 3: Histogram of Labels for resulting Classification datasets.

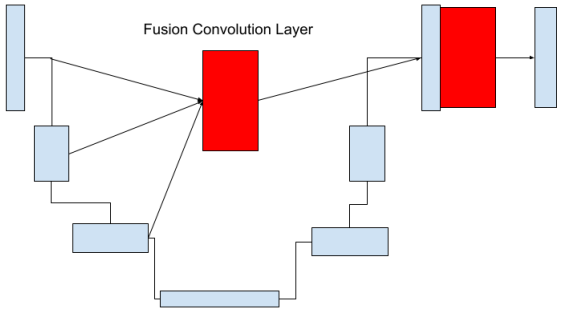


Fig. 4: Unet-Fusion Architecture

architecture, U-Net acts as a baseline in these experiments. The different variations tried are :-

1) **Unet with Resnet18 Backbone**:- The unet encoder stage is replaced by trained layers of resnet 18. Skip connection and other details are kept the same.
2) **Unet with resnet34 Backbones**:- Similar to the variation above, intead resnet 34 layers were used.
3) **Unet with Modified skip connections** :- The encoder features instead of just plainly transferred to decoder side are passed through a convolutional layer.
4) **Unet with fusion Extension**:- The Skip Connections are removed. Instead all the different subencoder features are fused using convolutional layers and upsampling as used in [9]. The features thus created are concatenated at the final decoding stage just after the final upsampling stage of the unet architecture. Architectural diagram is shown figure 4.

These five models are compared with each other using F1-Score and IOU with a threshold of 0.5. The F1 Score Reported are by an F1 score implementation on Tensors using a threshold of 0.5, this was done to speed up the training. Using F1-Score from sklearn was slowing down the Training process.

A. Dataset I

Training F1-score & IOU-score and Validation F1-score & IOU-score on the dataset are shown in Table II and Table III. Overlapping patches didn't make a difference in the final segmentation metrics. U-Net type architecture with resnet-18 and resnet-34 backbone work best amongst all the models compared. Unet with modified skip connections and u-net fusion architecture perform better than simple u-net.

B. Dataset II

Testing F1-Score, IOU score when compared with given testing pathologist segmentation is shown in Table IV. Again with this dataset using pretrained backbone of resnet in U-net architecture performed better than other architectures. Models trained were more aligned with 1st pathologist more than the 2nd pathologist.

Model	**TrainF1**	**TrainIOU**	**ValF1**	**ValIOU**
U-net	0.8050	0.6776	0.8175	0.6971
U-resnet-18	0.8398	**0.7479**	**0.8568**	**0.7900**
U-resnet-34	**0.8409**	0.7474	0.8440	0.7634
U-Modified-skip	0.8040	0.6770	0.8345	0.7255
U-Fusion	0.8007	0.6825	0.8272	0.7138

TABLE II: Segmenation Metrics for Dataset I when no overlap between patches. Above are the values for best trained models.

Model	**TrainF1**	**TrainIOU**	**ValF1**	**ValIOU**
U-net	0.7962	0.6664	0.8146	0.6873
U-resnet-18	**0.8374**	**0.7469**	**0.8674**	**0.8187**
U-resnet-34	0.7938	0.6829	0.8324	0.7464
U-Fusion	0.8122	0.6974	0.8618	0.7629
U-Modified-skip	0.8144	0.6920	0.8348	0.7316

TABLE III: Segmenation Metrics for Dataset I when 128 pixel overlap between patches

Model	**TestF1-1**	**TestIOU-1**	**TestF1-2**	**TestIOU-2**
U-net	0.8057	0.6788	0.7781	0.6418
U-resnet-18	0.8324	0.7216	**0.8034**	**0.6837**
U-resnet-34	**0.8328**	**0.7219**	0.8033	0.6828
U-Fusion	0.8083	0.6831	0.7813	0.6474
U-Modified-skip	0.8243	0.7069	0.7957	0.6685

TABLE IV: Segmenation Metrics for Dataset II when no overlap between patches.

Model	**TestF1-1**	**TestIOU-1**	**TestF1-2**	**TestIOU-2**
U-net	0.7488	0.6027	0.7293	0.5780
U-resnet-18	0.7740	0.6361	0.7463	0.6005
U-resnet-34	0.7982	0.6695	0.7722	0.6360
U-Fusion	0.7791	0.6422	0.7565	0.6130
U-Modified-skip	0.7659	0.6247	0.7449	0.5980

TABLE V: Segmenation Metrics for Dataset II when no overlap between patches. Using Dice loss.

IV. Classification Experiments

For Classification experiments the following models were used i) resnet18 ii) resnet34 [3] iii) Mobilenet [7] iv) MNASNet [8]. Data augmention for these experiment was random horizontal, vertical flip and random affine with both rotation and translation. Without Data augmentation these models were overfitting the data. The metrics used are accuracy and ROC AUC score for mutliclass classfication using marco averaging.

A. Dataset I

Training Accuracy, Validation accuracy and Validation ROC are listed in Table VI and Table VII. Table VI shown the result when patches are sampled with no overlap between each other. Training and validation dataset are created from all patches thus sampled. Table VII show results when patches are sampled with 128 pixels overlap. Overlap changes the training and validation accuracy by a fairly significant margin. Looking at the confusion matrix the model was easily able to classify background, Gleason Grade 1, and Gleason Score 5. But almost all models had most difficulty in classifying patches of Gleason Score 3 and Gleason Score 4. For both overlapping anf non-overlapping cases, resnet-34 pretrained model works best.

Model	Train Accu	Val Accu	Val ROC
resnet18	0.8976	0.8755	0.9738
resnet34	**0.9689**	**0.8775**	**0.9690**
MobileNet	0.9177	0.8748	0.9600
MNASnet	0.8624	0.8584	0.9490

TABLE VI: Validation and Train Accuracy When patches for datset can be taken from the same slides between Training and Validation sets. Dataset I. No overlap between patches.

Model	Train Accu	Val Accu	Val ROC
resnet18	0.9302	0.9258	0.9885
resnet34	**0.9392**	**0.9348**	**0.9909**
MobileNet	0.8389	0.8543	0.9333

TABLE VII: Validation and Train Accuracy When patches for datset can be taken from the same slides between Training and Validation sets. Dataset I. Patched Overlap for 128 pixels.

B. Dataset II

Testing Accuracy, Testing ROC for each pathologist are listed in Table VIII when patches are sampled without overlap. The training and validation accuracies are quite high when compared to testing accuracies in the Table. MNAS-net is the best performing model based on accuracy and also ROC. With overlap the data is listed in Table X. The changes in accuracy due to overlap doesn't changes that much.

When the whole dataset is divided before sampling patches, the classification testing accuracies are listed in Table IX. With this new sampling method of patches it was expected that models trained will perform better on the Testin data. But that was not the case for all models. Only resnet34 model performed well.

Model	Test1 Accu	Test2 Accu	Test1 ROC	Test2 ROC
resnet18	0.6881	0.6130	0.8456	0.8039
resnet34	0.6792	0.6102	0.8380	0.8000
MobileNet	0.6923	0.6210	0.8594	0.8249
MNASnet	**0.6952**	**0.6249**	**0.8727**	**0.8500**

TABLE VIII: Testing Accuracies and ROC for different pathologist when slides were common between Training and Validation sets. Dataset II. 'Test1' corresponds to testing accuracy compared with pathologist 1, and 'Test2' corresponds to testing accuracy when compared with pathologist 2.

Model	Test1 Accu	Test2 Accu	Test1 ROC	Test2 ROC
resnet18	0.6144	0.5214	0.7097	0.6703
resnet34	**0.7022**	**0.6600**	0.7774	0.7445
MobileNet	0.6348	0.5743	**0.7936**	**0.7565**
MNASnet	0.6628	0.6278	0.7835	0.7437

TABLE IX: Testing Accuracies and ROC for different pathologist When no slides were common between Training and Validation sets. Dataset II.'Test1' corresponds to testing accuracy compared with pathologist 1, and 'Test2' corresponds to testing accuracy when compared with pathologist 2.

C. Observations

The following observation were made during the training and testing process for these datasets :-

1) Overlapping patches didn't help classification models in getting to a better accuracy but helped in segmentation models.

Models	No Overlap		Overlap 128 pixel	
	Test Accu	Test ROC	Test Accu	Test ROC
resnet18	0.6881	0.8456	**0.7122**	0.8639
resnet34	0.6792	0.8380	0.6982	0.8581
MobileNet	0.6923	0.8594	0.7118	0.8724
MNASnet	**0.6952**	**0.8727**	0.7106	**0.8825**

TABLE X: Testing Accuracy When slides were common between Training and Validation sets but different overlap. Dataset II. Pathologist 1.

Models	No Overlap		Overlap 128 pixel	
	Test Accu	Test ROC	Test Accu	Test ROC
resnet18	0.6130	0.8039	**0.6437**	0.8314
resnet34	0.6102	0.8000	0.6313	0.8231
MobileNet	0.6210	0.8249	0.6419	0.8450
MNASnet	**0.6249**	**0.8500**	0.6379	**0.8559**

TABLE XI: Testing Accuracy When slides were common between Training and Validation sets but different overlap. Only Pathlogist 2. Dataset II.

2) Classification models like resnet-18 and resnet-34 tend to overfit the training data when compared with mobilenet and MNAS-net. MNAS-net and mobile-net has less training accuracy compared to resnet-18 and resnet-34 but performed well on testing data.
3) resnet based backbone for u-net architecture performs better than simple u-net. Extensions of u-net architecture also performed better in terms of F1-Score and IOU score.
4) Main 2 classes for misclassification of patches were Gleason grade 3 and Grade 4.
5) Using cross-entropy loss or Dice loss for segmenation models didn't perform so well. BCE loss for segmentation gave much better results.

V. Future Work

Due to limitation of time some ideas that weren't implemented and can be explored in future study can be :-

1) exploring more backbones for segmentation encoder of u-net architecture. Maybe Mobile net or MNAS-net which performed

Models	No Overlap		Overlap 128 pixel	
	Test Accu	Test ROC	Test Accu	Test ROC
resnet18	0.6144	0.7097	**0.6231**	**0.8052**
resnet34	**0.7022**	**0.7774**	0.5875	0.8050
MobileNet	0.6348	0.7936	0.4222	0.7167
MNASnet	0.6628	0.7835		

TABLE XII: Testing Accuracy When slides were no common slides between Training and Validation sets but different overlap. Only Pathlogist 1. Dataset II.

Models	No Overlap		Overlap 128 pixel	
	Test Accu	Test ROC	Test Accu	Test ROC
resnet18	0.5214	0.6703	0.6129	**0.7676**
resnet34	**0.6600**	0.7445	**0.6332**	0.7658
MobileNet	0.5743	**0.7565**	0.4218	0.6885
MNASnet	0.6278	0.7437		

TABLE XIII: Testing Accuracy When no slides were common between Training and Validation sets but different overlap. Dataset II. Pathologist 2.

fairly well on classification type dataset, might also work well for segmentation models.
2) Data augmentation for segmentation were not explored in this study becuase patch segmentation gave huge amounts of data to work with, but with data augmentation better trained models will be achieved.
3) More experiments on Fusion type architecture.
4) Trying different loss functions like Focal loss or Jaccard loss.
5) More experiment on patch type classification with different classification architectures.
6) Pretraining a model on one dataset and then retraining on other dataset, this way we can get the advantage of both datasets and maybe get a better generalizable model.

REFERENCES

[1] Eirini Arvaniti, Kim Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter Wild, Jan Hendrik Rüschoff, and Manfred Claassen. Replication Data for: Automated Gleason grading of prostate cancer tissue microarrays via deep learning., 2018.

[2] Eirini Arvaniti, Kim S Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter J Wild, Jan H Rueschoff, and Manfred Claassen. Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports*, 8(1):1–11, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Davood Karimi, Guy Nir, Ladan Fazli, Peter C Black, Larry Goldenberg, and Septimiu E Salcudean. Deep learning-based gleason grading of prostate cancer from histopathology images—role of multiscale decision aggregation and data augmentation. *IEEE journal of biomedical and health informatics*, 24(5):1413–1426, 2019.

[5] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[6] Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, R Storey Wilson, et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical image analysis*, 50:167–180, 2018.

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019.

[8] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile, 2019.

[9] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural Information Processing Systems*, 2020.