Similarity Learning: The Key to Unsupervised Gleason Scoring

Ong Zhi Lin, Charlene[0000−0003−0250−0415]

School of Computer Science and Engineering, Nanyang Technological University
ongz0070@e.ntu.edu.sg

Abstract. Could we devise an approach to discriminate between different Gleason patterns of the prostate gland biopsy, by *training the model to learn from the images themselves*? Current Gleason grading is complex and has high interobserver and intraobserver variability. In this paper, we propose the use of unsupervised Convolutional Siamese Networks and demostrate that we can train the model to learn meaningful embeddings representing the different Gleason patterns. By reducing the embeddings to a single value for each whole slide image using principal component analysis, we found that there is a statistically significant difference in the means of the first principal components for slides of ISUP score 0 with 2 and 5, at an alpha level of 0.05, using the Welch's T-test. Our results demostrate that the use of similarity learning can be a powerful tool for developing unsupervised Gleason scoring and potentially aid in improved diagnosis outcomes for patients.

Keywords: Contrastive Learning · Prostate Cancer · Digital Pathology

1 Introduction

Prostate cancer is one of the main causes of cancer deaths in men around the world. [12] Current diagnostic approaches for prostate cancer involve the use of tissue biopsy which remains the gold standard for diagnosis. In order to grade the severity of the prostate cancer based on the biopsy, the Gleason grading is developed based on the structural patterns of the prostate tumour cells and is the most frequently used histological scoring methodology.

The Gleason grading can be classified into three different categories, Gleason score 3, 4 and 5, and in increasing order of severity. [16] The various Gleason patterns are shown on Fig. 7. In order to grade the whole slide images (WSIs), the Gleason scores usually comprised of 2 grades, [4] which consist of the most frequently occurring pattern and the second most frequently occurring pattern. However, the second most frequently occurring pattern must account for at least 5% of the tissue, unless it is the highest grade which must always be part of the score. In order to better determine patient treatments, the Gleason grading have also been converted to a International Society of Urological Pathology (ISUP) grade for each biopsy, as shown on Table 1.

However, high interobserver and intraobserver variability is common in the grading of Gleason scores. Ozkan et al [16] explored the interobserver variability of two pathologists and found that the κ with respect to the Gleasom sum was 0.43, and the concordance on Gleason pattern 4 and 5 tumours was only 18%.

In view of the challenges in histopathology, deep learning have been shown to provide state of the art approaches that can rival or even surpass pathologists' scoring. [15] Most of these approaches involved supervised techniques that require thousands of annotations. However, to the best of our knowledge, there is currently limited or no work done to explore the use of unsupervised deep learning techniques to learn meaningful representations on patches from H&E images of prostate glands and study these representations according to the different Gleason scores. This would be important to learn a representation innate to the images and alleviate the problem of high pathologists' variability.

Prior unsupervised techniques are applied in other modalities or looked into the classification of tumour vs non-tumour tissue. [3] In particular, Bulten et al [3] utilised convolutional adversarial autoencoders to classify prostate tissue as either tumour or non-tumour tissue without any labeled data. In this work, we aim to show that unsupervised learning, in particular using similarity learning, offers more discrimminative value as to the *actual Gleason patterns.*

Inspired by methodology first proposed by Gildenblat et al [7], we propose the use of a Convolutional Siamese network with DenseNet-121 backbone in this study. We trained the model using two different losses, the contrastive and the triplet loss, and compared the model results with a weakly supervised model where WSI ISUP scores are utilised during the sampling of the patches. In this study, we utilised 298 H&E images from Radboud Medical Centre, which are part of the 2020 Prostate cANcer graDe Assessment (PANDA) challenge. Using principal component analysis (PCA), we reduced the dimensions of the embeddings into 3 components and showed that these embeddings hold great discrimminative value for the Gleason patterns. We further extended the work to look at slide-level ISUP scoring and evaluated the means of the first principal components of the mean embedding from samples with different ISUP scores.

Table 1: Conversion of Gleason score to ISUP score

Gleason score	ISUP score
6	1
7 (3+4)	2
7 (4+3)	3
8	4
9 - 10	5

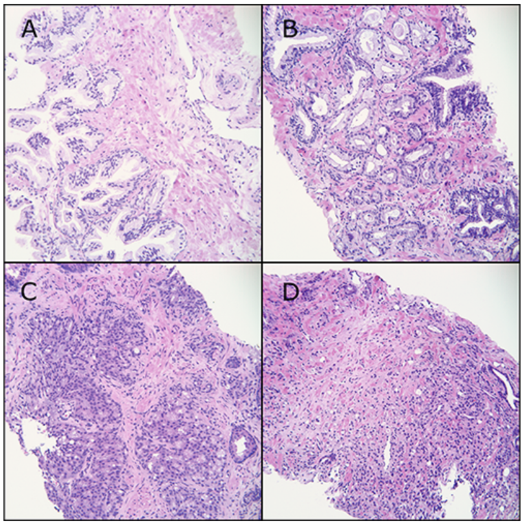

Fig. 1: Gleason patterns on H&E Images (A) Healthy glands (B) Gleason pattern 3 (C) Gleason pattern 4 (D) Gleason pattern 5 [4]

2 Background

2.1 Similarity Learning

Siamese networks are a well-known framework for similarity learning. They are first introduced in 1990s by Bromley et al [2] for the detection of forged signatures. It consists of two identical artifical neural newtorks which share the same weights and which both learn a feature representation of their inputs. [6] By comparing the output generated by the Siamese network, the semantic similarity between two inputs can be compared.

Common losses used in the Siamese networks are the contrastive and the triplet losses. [14] The contrastive loss [8] works by reducing the distance between a pair of inputs which have the same label, or the positive pair, and increasing the distance for the pair of inputs with a different label, or a negative pair, and is formulated as follows.

$$L(A, B) = (1 - y)(\| f(A) - f(B) \|_2)^2 + (y)max(0, m - \| f(A) - f(B) \|_2)^2 \quad (1)$$

where y is the ground truth label that indicates if the two inputs are a positive, 0, or a negative pair,1, m is the margin, $f(A)$ and $f(B)$ are the two outputs of the Siamese neural network from the inputs. On the other hand, the triplet loss allows for direct comparison between the negative and positive pair. This idea of relative comparison is first introduced by Schultz et al in 2003. [18] This is done by comparing an anchor image with another image of the same class, the positive pair, and another image with a different class, the negative pair, and stipulating that the L2 norm of the absolute distance between the negative pair must be greater than the L2 norm of the absolute distance between positive pair. [9] By introducing a margin, the triplet loss can be formulated as follows.

$$L(A, P, N) = max(0, \parallel f(A) - f(P) \parallel_2 - \parallel f(A) - f(N) \parallel_2 + m) \tag{2}$$

where m is the margin, $f(A)$, $f(P)$ and $f(N)$ are outputs of the Siamese network for the anchor, positive and negative input.

There have also been other state of the art losses proposed over the years, such as ProxyNCA [14] and Soft Triple Loss[17].

2.2 Related Work on Use of Siamese Networks

Siamese networks are usually used to learn latent representations for supervised learning. Tellez et al [19] investigated the use of Siamese networks for unsupervised learning by using positive pairs consisting of patches from the same WSI location but with different augmentation, and using negative pairs consisting of neighbouring patches or patches from different WSI locations. This is used to generate compressed representations which are used to train a CNN-based classifier. [19]

In another work by Gildenblat et al [7], they proposed a novel self-supervised methodology using spatially adjacent tiles as positive pairs and distant tiles as negative pairs to train a Siamese network to generate meaningful image representations.

In addition to learning feature representations, Siamese networks have also been widely used in few-shot learning. Koch et al [11] proposed the use of Convolutional Siamese network for character classification of alphabets from Omniglot dataset. Koch et al compared the performance of Convolutional Siamese networks with other methods, and found that Convolutional Siamese networks is only slightly below human error rates.

Cano et al [5] also utilised Convolutional Siamese Networks for classification of different samples of tissue in breast cancer histopathology images. and obtained a 90.83% accuracy.

3 Datasets

The Prostate cANcer graDe Assessment (PANDA) Dataset [4] consists of 11,000 whole slide images of digitized H&E stained biopsies from two centers, namely the Radboud University Medical Center (Radboudumc) and the Karolinska Institute. Both the Gleason Score and the International Society of Urological Pathology (ISUP) grade are provided.

Region-based annotations are provided by the organisers for both the datasets from Radboudumc and Karolinska Institute. For the Radboudumc Dataset, the annotations can be considered as weakly supervised labels as they are semi-automatically generated by several deep learning algorithms and contain noise. However, the labels of the Radboudumc dataset are more localised to the diseased regions, as the individual prostate glands are labeled. On the other hand, for the Karolinska dataset, regions are labelled. The number of categories of region-based labels in the Radboudumc dataset is also greater than the Karolinska dataset, and is shown in Table 2.

Table 2: Region-based Labels in Radboudumc Dataset

Values	Explanation
0	Background (non-tissue) or unknown
1	Stroma (connective tissue, non-epithelium tissue)
2	Healthy (benign) epithelium
3	Cancerous epithelium (Gleason 3)
4	Cancerous epithelium (Gleason 4)
5	Cancerous epithelium (Gleason 5)

Hence, we decided to utilise the Radboudume dataset in this study. Due to computational constraints, only 298 images from the Radboudume dataset are used in the model development and evaluation.

4 Proposed Approach

In this section, we would describe our proposed approach, which involves data preprocessing and our model architecture.

4.1 Data Preprocessing

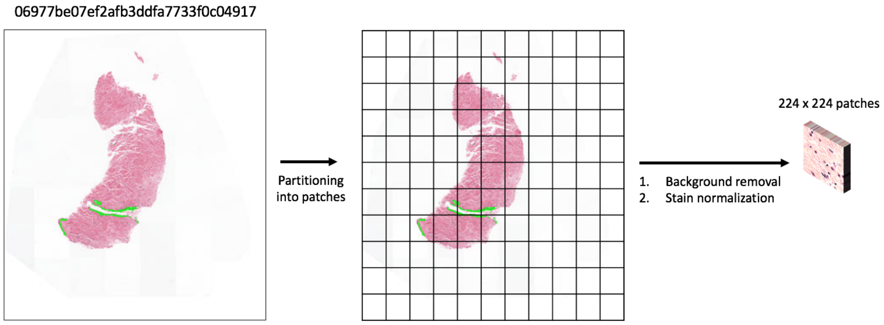

Fig. 2: Processing Workflow of WSIs

The WSIs are first partitioned into patches of 224 pixels by 224 pixels. The patches are then converted to grayscale images for background removal and then a simple intensity threshold of 0.8 is applied. As majority of the WSIs consists of the white background, patches with less than 90 percent of tissue are discarded [20]. Next, stain normalisation is applied [13]. The input to the model are 224 pixels by 224 pixels stain-normalised RGB images.

In the given dataset, the mask labels are given as pixel-level annotations. It must be noted that a single patch could contain two or more different labels. In order to prepare the patch-level labels for evaluation, four different approaches are employed, as shown in Table 3.

Table 3: Variants of Mask Labels Tested

Variants	Description
V1	Patch Label According to Most Frequently Occurring Label
V2	Patch Label According to Highest Gleason Score Occupying at Least 5 % of Tissue
V3	Patch Label According to Most Frequently Occurring Label, Cancerous vs Non-Cancerous Tissue
V4	Patch Label According to Highest Gleason Score Occupying at Least 5 % of Tissue, Cancerous vs Non-Cancerous Tissue

For variants V1 and V3, the most frequently occurring value is taken to be the patch label. In view that it could be interesting to also investigate the differentiation between non-cancerous and cancerous tissue, this most frequently occurring value is then reassigned to 0 if it is 0,1,2, which indicates non-cancerous tissue or 1, if it is 3,4,5, which indicates presence of cancerous tissue in the variant V3.

For variants V2 and V4, we considered that the most frequently occuring label might not be clinically relevant. Instead, the highest Gleason score which occupies at least 5% of the tissue in the patch is taken as the patch label. This is in line with the Gleason scoring methodology, where the highest grade must always be part of the score, and the second most frequently occurring pattern must account for at least 5% of the tissue. In consideration that labels in Radboudumc dataset are weak labels and are susceptible to noise, we employ a stricter criteria requiring the highest Gleason score to be at least 5% of the patch to be counted as a patch label. Similar to V1 and V3, the difference between V2 and V4 is that the patch label is then reassigned to 0 if it is 0,1,2 or 1 if is 3,4,5 in variant V4.

It must be noted that these patch labels are not utilised for training at all in both the unsupervised and weakly unsupervised methods. However, they are crucial in evaluating the trained model and comprehending the results. It can be seen from Fig. 3, 4, 5 and 6 that the distribution of the patch labels vary significantly depending on how the labels are determined. In general, it can be seen that variants V2 and V4 capture a greater percentage of cancerous tissue and are perhaps more clinically relevant labels to use.

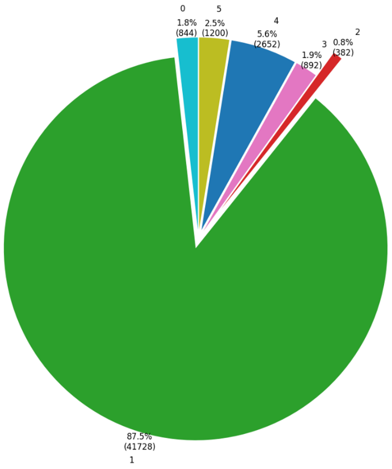

Fig. 3: Distribution of Patch Labels According to V1

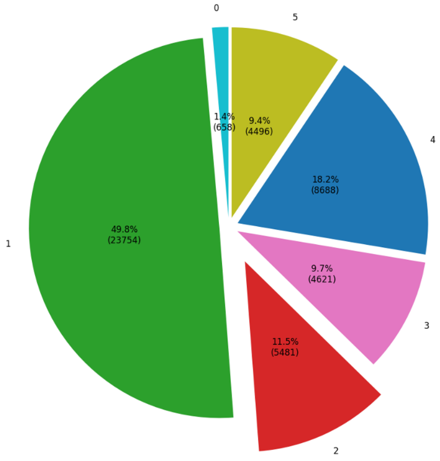

Fig. 4: Distribution of Patch Labels According to V2

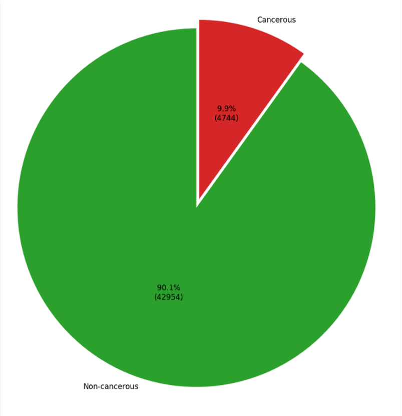

Fig. 5: Distribution of Patch Labels According to V3

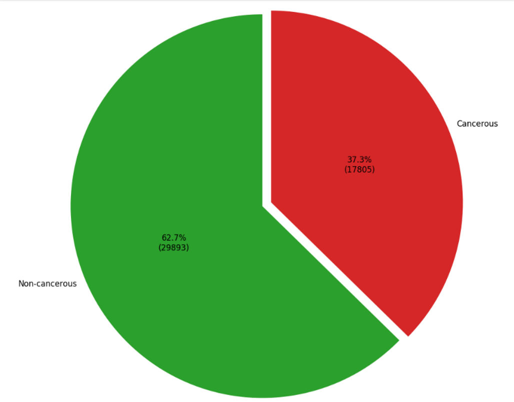

Fig. 6: Distribution of Patch Labels According to V4

4.2 Model Architecture

We propose a Siamese network with a Densely Connected Convolutional Network (DenseNet) backbone. The DenseNet is first proposed by Huang et al in 2018 [10] and directly combines features by concatenating the features of the different layers. It has been shown to achieve state-of-the-art performances in benchmark datasets such as CIFAR datasets. In our approach, we utilised a pretrained DenseNet with 121 layers. Inspired by Gildenblat et al [7], we adopt a similar approach that neighbouring patches should be more similar to each other, compared to distant patches in the image, due to spatial continuity of WSIs. We stipulate that similar to the case in [7], the neighbouring patches can be considered as the same class and distant patches can be considered as different classes. The pretrained DenseNet-121 is also replaced with a last fully connected layer with an output of size 128.

In this work, we tested 3 different variants of Convolutional Siamese networks, as shown on Table 4. Siamese-UC and Siamese-UT have the same sampling strategy of random sampling of similar and dissimilar pairs from near and distant patches in the same WSI. However, Siamese-UC consists of 2 branches of DenseNet and a contrastive loss is used. On the other hand, Siamese-UT consists of 3 branches of DenseNet and a triplet loss is used.

We also compared the unsupervised Convolutional Siamese networks with Siamese-WC, where weakly supervised labels are used to guide the random sampling. While the similar pairs are obtained from near tiles of the same WSI, dissimilar pairs are obtained from different WSIs with different ISUP score. Two branches of DenseNet and contrastive loss are used in this case.

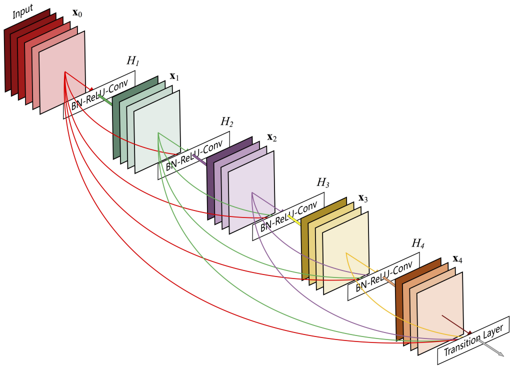

Fig. 7: DenseNet Model Architecture [10]

Table 4: Variants of Convolutional Siamese Networks

Variants	Description
Siamese-UC	Random sampling of neigbouring and distant patches from same WSI, Contrastive loss
Siamese-UT	Random sampling of neighbouring and distant patches from same WSI, Triplet loss
Siamese-WC	Random sampling of neighbouring patches from same WSI, distant patches from different WSIs with different ISUP scores, Contrastive loss

5 Implementation

Siamese-UC and Siamese-UT are trained for 20 epochs, while Siamese-WC is trained for 17 epochs [1], on a NVIDIA Tesla P100 GPU. The model at the 19th epoch is taken for Siamese-UC, while the model at 7th epoch is taken for Siamese-UT due to overfitting after 7 epochs. The model at the 17th epoch is taken for Siamese-WC. In order to learn meaningful semantic representations, data augmentation is included, consisting of color jitter, random horizontal flips, random vertical flips and random rotations, similar to Gildenblat et al [7]. This is to account for the fact that the orientation of the biopsy does not have any meaning. An Adam optimizer is also used with default parameters from Pytorch.

The criteria used for determining near vs distant patches is adopted from that proposed by Gildenblat et al [7], where near patches must lie within a Euclidean distance of 1792 pixels while distant patches must be separated by more than 6720 pixels in Euclidean distance. As each patch has one or more similar patches and one of more dissimilar patches, the number of possible pairs is gargantuan. In order to combat this issue, we devise a sampling strategy where we iterate through all the patches and randomly sample either a similar or a dissimilar patch with equal probability. This means that there could have different pairs in different epochs, also reducing the risk of overfitting. To prevent the likelihood that a patch may not have either a similar or a dissimilar patch, e.g. patches of isolated tissue, a quality check is done to ensure that all patches in the training dataset have both similar and dissimilar patches.

In total, 208 WSIs are randomly sampled from the whole dataset and used in the training and validation dataset. This is further split into a 80% - 20% train-validation split ratio.

6 Results

6.1 Evaluation Metrics

We utilise the common evaluation metrics such as recall at *k* and Normalised Mutual Information measure (NMI) in similarity learning. [14] Both metrics

[1] A server-related issue happened at the 18th epoch

are commonly used in similarity learning to benchmark results. The recall at k represents the percentage of the test dataset where the same label is in the k nearest neighbours. NMI evaluates the clustering quality after clustering the embeddings with a clustering technique, for examples, K-means clustering and can be expressed as follows [14]:

$$NMI(W,C) = 2 * \frac{I(W,C)}{H(W) + H(C)} \tag{3}$$

where $W = w_1, w_2, w_3, ..., w_k$ represents the embeddings that are assigned to different clusters through, for example, k-means clustering, and $C = c_1, c_2, ..., c_m$ represents the embeddings that are assigned with the actual class labels.

In order to compare the results with Gildenblat et al [7], the global Average Descriptor Distance Ratio (ADDR) is calculated. The average L2 distance between patches which belong to the same ground truth patch labels, d^+, is calculated.

$$d^+ = \frac{\sum_{n=1}^{N} \parallel \boldsymbol{x_1} - \boldsymbol{x_2} \parallel_2}{N} \tag{4}$$

where N represents the number of data points, and $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are two different embeddings with the same ground truth patch labels.

Similarly, the average L2 distance between patches which belong to different classes, d^-, is also calculated. $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ belong to two different classes in this formulation.

$$d^- = \frac{\sum_{n=1}^{N} \parallel \boldsymbol{x_1} - \boldsymbol{x_2} \parallel_2}{N} \tag{5}$$

The ADDR is then defined as follows.

$$ADDR = \frac{d^-}{d^+} \tag{6}$$

6.2 Comparison between Different Model Variants

In this section, we compared the results from the different model variants, Siamese-UC, Siamese-UT and Siamese-WC. 90 WSIs are used in the testing dataset. In the computation of the NMI and recall at k, same results should be obtained regardless of iterations as there are same representation of the embeddings in the high-dimensional space. For the computation of the L2 distances, it could vary as random sampling is done to select similar and non-similar pairs according to patch labels.

Comparison of NMI As we considered 4 different variants of patch labels in this work, we evaluated the NMI based on all 4 patch labels for the 3 different model variants. In total, we have 12 different values of NMI.

It can be seen that Siamese-UT has the highest NMI. It must be noted that the learning for the Siamese-UT converges very quickly and starts overfitting after 7 epochs.

Table 5: NMI of Different Model Variants

Label Variants	Siamese-UC	Siamese-UT	Siamese-WC
V1	0.080	**0.082**	0.053
V2	**0.146**	**0.146**	0.074
V3	0.069	**0.095**	0.070
V4	0.154	**0.200**	0.121

Comparison of ADDR It can be seen that Siamese-UT also performs the best in terms of the ADDR. The ADDR of all the variants are not high, compared to [7], where they showed a ADDR of 1.5. This could be due to the different patterns of pathologies present in prostate glands vs breast tissue.

Table 6: ADDR of Different Model Variants

Label Variants	Siamese-UC	Siamese-UT	Siamese-WC
V1	1.03	**1.17**	**1.17**
V2	1.11	**1.21**	1.12
V3	1.02	**1.15**	**1.15**
V4	1.11	**1.21**	1.12

Comparison of Recall at k Compared to ADDR or NMI, the Recall at k is a more significant measure to understand the value of our approach. We reported the global recall at k and recall at k for the different labels, for the 3 different model variants. As seen from NMI and ADDR, the definition of V2 and V4, i.e. taking the highest Gleason score that occupies at least 5% of the tissue, appears to be more promising and also more clinically relevant.

From Table 7 and 8, the Siamese-UC model has the best recall at 1 for most of the label categories, despite having lower ADDR and lower NMI except for label V2 compared to Siamese-UT and Siamese-WC. Surprisingly, Siamese-WC does not perform as well as Siamese-UC and Siamese-UT. The sampling strategy using WSIs of different ISUP scores may fail to pick out subtle features which could be critical in differentiating the different patches.

In addition, despite the low ADDR and NMI, it can be seen that the recall at k is impressive. Based on the distribution of the test dataset, 1.5% of the labels are classified as 0, 48.9% as 1, 10.1% as 2, 8.2% as 3, 22.3% as 4 and 9.0% as 5 for label V2. For label V4, 60.5% of the labels are classified as 0 and 39.5% as 1. The recall at 1 values for Siamese-UC is significantly higher than the corresponding percentage in the dataset, suggesting that the representations have important semantic value.

Table 7: Comparison of Recall at k for V2

Model	Siamese-UC				Siamese-UT				Siamese-WC			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
Global	**54.8**	69.1	81.2	90.6	51.3	65.5	78.8	88.6	50.8	66.3	79.5	89.6
0	16.0	27.8	38.7	45.8	16.5	23.6	37.3	44.8	**21.2**	27.8	36.3	45.8
1	**76.9**	88.0	93.7	97.4	75.5	87.1	93.8	97.1	73.2	87.2	94.3	98.0
2	**23.8**	38.7	58.3	78.7	19.6	32.1	51.0	70.9	19.4	34.5	53.3	73.6
3	**22.3**	37.4	55.4	75.6	16.3	27.5	45.8	68.9	19.1	33.2	51.8	72.5
4	**45.3**	65.0	82.0	92.8	38.7	60.2	80.1	91.6	38.6	59.2	78.7	91.2
5	**29.4**	46.5	66.9	82.9	24.2	40.7	62.4	80.0	27.9	42.5	63.1	80.7

Table 8: Comparison of Recall at k for V4

Model	Siamese-UC				Siamese-UT				Siamese-WC			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
Global	**75.2**	87.6	94.6	98.0	72.6	86.2	94.0	97.6	72.0	86.4	94.2	98.0
0	**79.2**	89.4	95.4	98.5	77.2	88.3	94.9	98.2	77.2	89.6	96.1	98.9
1	**69.1**	84.9	93.3	97.4	64.9	83.0	92.5	96.6	64.1	81.6	91.5	96.6

Visualisation of Embeddings In order to validate the results from recall at k, we visualise the embeddings from the 3 model architectures for label V2 and V4. Dimensionality reduction is done using PCA to reduce the embeddings to 3 components, for the three model variants. Amazingly, we observe that the embeddings with the same labels have been clustered nearer together *without any supervisory information at all* for Siamese-UC and Siamese-UT. For Siamese-UC, in particular, we observe a nice spectrum from 1 to 5 for the embeddings from Siamese-UC in Fig, 8 validating our results. It must be noted that label 0 represents background or unknown tissue.

We conclude that Siamese-UC has the best discriminative value for the different Gleason patterns.

Welch's T-test for Slide-Level Analysis We further evaluate the utility of embeddings from Siamese-UC by exploring *slide-level based information*. The mean embedding for each WSI is taken and the first principal component is taken and plotted on Fig. 9. Normality of the samples from each ISUP group is tested using the Shapiro-Wilk test at a significance level of 0.05. Samples from ISUP score 1 and 3 are removed from the analysis as they deviate from the normal distribution. After testing for normality, we conducted the Welch's T-test for ISUP grade 0, 2, 4, and 5 to test the hypothesis that the any two populations of first principal component from the different ISUP grades have equal means, at a significance level of 0.05. We observe that we can reject with 95% confidence that the mean of the 1st PCA component from the samples of ISUP score 0 is equal to 2 or 5.

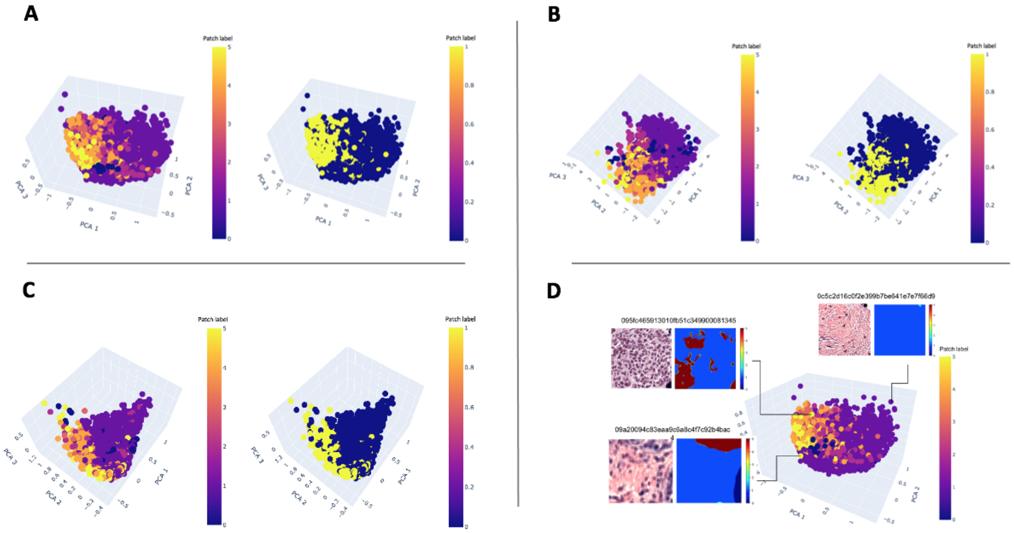

Fig. 8: Visualisation of Embeddings from A.Siamese-UC B.Siamese-UT C.Siamese-WC D.Visualisation of Patches from Embeddings from Siamese-UC with Label V2
(For A, B, C, Left: Label V2, Right: Label V4)

7 Conclusion and Future Work

We propose the use of unsupervised Convolutional Siamese Networks adapted from [7] to learn deep meaningful representations that offer great discriminative value to the Gleason patterns. We demostrate that these embeddings have potential to act as distinguishing markers for samples from different ISUP scores.

To characterise patch-level embeddings as a single vector to capture slide-level information, the mean of these embeddings are taken. In reality, the ISUP score is more weighted towards embeddings representing patches of higher Gleason patterns, as higher Gleason patterns will be counted into the Gleason score irrespective of the percentage. However, as the aim is to show this relationship in a completely unsupervised manner, prior knowledge as to the actual patch labels is not included in this analysis.

In addition, we observe that there is a much greater percentage of non-cancerous tissue compared to cancerous tissue should we only consider the most frequent label of the patch, as seen in Fig. 5. It might be interesting to characterise a model to detect what is normal tissue, before using it to detect abnormal tissue to account for the class imbalance. In particular, Alaverdyan et al [1] proposed a model consisting of Siamese network with stacked Convolutional

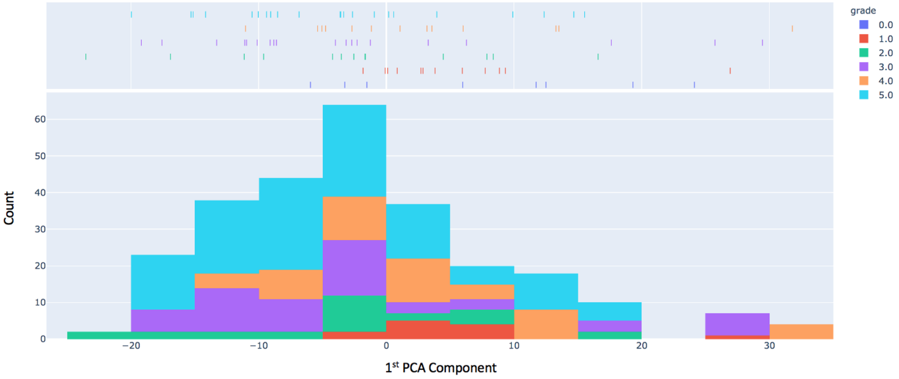

Fig. 9: Distribution Plot of First Principal Component of Samples from Different ISUP Scores

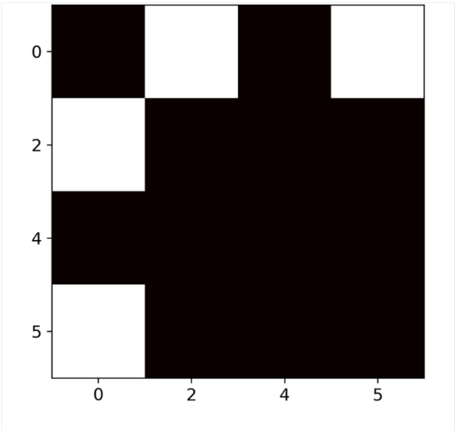

Fig. 10: Heatmap Comparing Means of First Principal Component of Samples of Different ISUP Scores using Welch's T-test with Alpha=0.05
(White: Null hypothesis can be rejected with 95% confidence level that the mean of the populations from the corresponding groups are equal.)

Autoencoders as subnetworks that is trained on healthy MRI dataset, followed by the use of Support Vector Machines for outlier detection. Future work could touch on similar approaches to characterise normality to account for huge class imbalance.

References

1. Alaverdyan, Z., Jung, J., Bouet, R., Lartizien, C.: Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening. Medical Image Analysis **60**, 101618 (2020)
2. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a “siamese” time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence **7**(04), 669–688 (1993)
3. Bulten, W., Litjens, G.: Unsupervised prostate cancer detection on h&e using convolutional adversarial autoencoders. arXiv preprint arXiv:1804.07098 (2018)
4. Bulten, W., Litjens, G., Pinckaers, H., Ström, P., Eklund, M., Kartasalo, K., Demkin, M., Dane, S.: The PANDA challenge: Prostate cANcer graDe Assessment using the Gleason grading system (Mar 2020). https://doi.org/10.5281/zenodo.3715938, https://doi.org/10.5281/zenodo.3715938
5. Cano, F., Cruz-Roa, A.: An exploratory study of one-shot learning using siamese convolutional neural network for histopathology image classification in breast cancer from few data examples. In: 15th International Symposium on Medical Information Processing and Analysis. vol. 11330, p. 113300A. International Society for Optics and Photonics (2020)
6. Chicco, D.: Siamese neural networks: An overview. Artificial Neural Networks pp. 73–94 (2020)
7. Gildenblat, J., Klaiman, E.: Self-supervised similarity learning for digital pathology. arXiv preprint arXiv:1905.08139 (2019)
8. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06). vol. 2, pp. 1735–1742. IEEE (2006)
9. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition. pp. 84–92. Springer (2015)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
11. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop. vol. 2. Lille (2015)
12. Litwin, M.S., Tan, H.J.: The diagnosis and treatment of prostate cancer: a review. Jama **317**(24), 2532–2542 (2017)
13. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. pp. 1107–1110. IEEE (2009)
14. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 360–368 (2017)

15. Nagpal, K., Foote, D., Liu, Y., Chen, P.H.C., Wulczyn, E., Tan, F., Olson, N., Smith, J.L., Mohtashamian, A., Wren, J.H., et al.: Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. NPJ digital medicine **2**(1), 1–10 (2019)
16. Ozkan, T.A., Eruyar, A.T., Cebeci, O.O., Memik, O., Ozcan, L., Kuskonmaz, I.: Interobserver variability in gleason histological grading of prostate cancer. Scandinavian journal of urology **50**(6), 420–424 (2016)
17. Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R.: Softtriple loss: Deep metric learning without triplet sampling. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6450–6458 (2019)
18. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. Advances in neural information processing systems **16**, 41–48 (2003)
19. Tellez, D., van der Laak, J., Ciompi, F.: Gigapixel whole-slide image classification using unsupervised image compression and contrastive training (2018)
20. Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016)