A Voting-Based Encoding Technique for the Classification of Gleason Score for Prostate Cancers

Zobia Suhail[1], Arif Mahmood[2], Liping Wang[3], Paul N Malcolm[4], and Reyer Zwiggelaar[1]

[1] Department of Computer Science, Aberystwyth University, Aberystwyth, United Kingdom
[2] School of Computer Science and Software Engineering, University of the Western Australia, WA, Australia
[3] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, United Kingdom
[4] Norfolk and Norwich University Hospitals NHS Foundation Trust, Norwich, United Kingdom

Abstract. We present a novel approach for classifying the Gleason score for prostate tumours based on MRI data. Proposed approach uses three scores: 2, 3 and 4-5 (representing Gleason scores 4 and 5 as one single class). Patches are extracted from annotated MRI data for each of the class. Raw image patches have been used as features, instead of extracting manual hand-crafted features. Each patch is encoded using a dictionary and the encoded feature vector is then used for classification. A voting-based encoding approach is used to transform data from the image domain to more discriminative class-specific representations. Initial investigation demonstrated excellent results (Classification Accuracy equal to 85% and Area Under the ROC Curve (AUC) of 0.932) for 3-class Gleason score classification for prostate tumours.

1 Background

Prostate cancer is considered to be one of the most common life threatening disease in old aged men, where 85% of cases are found in the patients after age of 65 years [4]. The survinal rate of patients is strongly influenced by the detection of cancer at early stage [11,5]. According to a study in the United States, prostate cancer represents about 1/3 of the newly diagnosed cancer cases and comes in the top 3 most common types of cancer occuring in men [12]. After prostate cancer diagnostics, pathologists define the aggressiveness of the cancer by using Gleason scores. The Gleason score ranges from numbers 1 to 5, where 1 means almost normal tissues. A higher Gleason score indicates a higher probability that the cancer will grow and spread. In clinical practice, the sum Gleason score is also used where the Gleason scores of two cancerous areas are added up (and as such this ranges between 1 and 10). Classification of Gleason scores based on the image processing technique can be helpful in clinical practice for improving

the treatment process and to avoid further patients intervention.
Several methods have been proposed in the past for image-based Gleason score classification of prostate cancer [2,16]. Fehr *et al.* [2] used first and second order texture features for the classification of high (>7) and low (<6) sum Gleason scores. They used multi-parametric MRI images for feature extraction. By using machine-learning based classification methods, they reported 93% classification accuracy for cancers present in both the peripheral and transition zones. Tiwari *et al.* [16] proposed a hierarchical machine learning-based classification of sum Gleason scores, where Gleason score 6 and 7 (3+4) were combined as low grade class and 7 (4+3) and 8 were combined as high grade. They combined T2-w MR images with MR spectroscopy images. Using their Semi Supervised Multi Kernel Graph Embedding strategy they reported areas under the ROC curve equal to 0.84±0.07 for classifying voxels based on high and low grade sum Gleason scores. Doyle *et al.* [1] proposed a graph-based morphological and texture features for the classification of Gleason grade 3 and 4 versus benign epithelium and benign stroma. They reported classification accuracy of 76.9% for the classification of Gleason grade 3 and 4. They concluded that texture and graph-based features were significant in classifying different tissue classes. Tabesh *et al.* [14] used several features (color histograms, fractal features, etc.) for classification of the Gleason grade as high or low grade. By using an independent test set they reported an accuracy of 77.6% for classification of Gleason score as high versus low grade. In their later work [15], prostate cancer was diagnosed alongwith the Gleason score classification into high and low grade. They combined colour, texture and morphometric features. After feature selection and using several classifiers, they reported 81% accuracy.
Recently, dictionary based approaches have become popular for solving problems in various medical images, for example for prostate and breast cancer classification and segmentation [8,10,13]. A texton-based approach was used in [8] for the classification of Gleason score as grade 3 and 4. They used filter responses for generating a texton dictionary, where a Random Decision Forest (RDF) was used for clustering. A texton histogram was generated after assigning each pixel to the closest textons, that were then used as features for a SVM classifier.
In contrast to most of the existing methods which performs two class classification, we proposed a novel method for the classification of Gleason scores into one of the three classes. In order to avoid data in-balance, small patches from the tumour areas are extracted and used as features for the overall model. Additionally, a soft assignment technique is used to generate features for each of the extracted patches. Initial results reveal excellent performance for the 3-class classification problem for Gleason scores in prostate cancer.

2 Methodology

In this section, we describe complete framework of the proposed approach for classification of prostate tumour by Gleason score.

2.1 Data-Set Preparation

The dataset that has been used in the experiments are segmented prostate cancer areas within the peripheral zone of T2-w MRI. The dataset is annotated by expert radiologists from the Norfolk and Norwich University Hospital. For the initial step the segmented area corresponding to the tumour region in T2-w MRI has been extracted. The associated classes are defined as Gleason 2, 3, whereas tumours belonging to Gleason 4 and 5 were combined as one class and is represented as 4-5 (throughout the paper classes are used as 2, 3 and 4-5). Four cases were extracted for each of the defined class (i.e. 2, 3 and 4-5). A sample tumour region extracted can be seen in Figure 1.

(a) (b) (c) (d)

Fig. 1: Extracting the prostate area from T2-w MRI corresponding to the annotation: (a) T2-w MRI slice, (b) annotation provided by the radiologist, (c) the corresponding masked area bounded by the tumour size, (d) the extracted tumour region from the T2-w MRI.

After extracting the tumour region, patches of size 5×5 were extracted from each tumour region covering each of the Gleason scores. Only patches completely within the tumour region were extracted (see Figure 2). The relevant information that is being used are the raw image data corresponding to the 5×5 patch extracted from the tumour area.
The number of cases that were used for the experiments are 4 per each class (2, 3 and 4-5), therefore in total we had 12 cases from different patients. As T2-w MRI is 3-dimensional data, there can be multiple slices covering tumour. We selected one slice for each patient, corresponding to the largest tumour area annotation provided by the radiologist. The largest tumour from multiple annotated regions was selected so that the maximum number of patches could be extracted from each patient in order to increase the sample size.
The data is divided into 2-distinct sets; i.e. data used for model generation and the data that is used for model evaluation. The split has been done at a patient level (not at the patch level), so that the patient cases that are included in generating the model are not included for model evaluation. Patches extracted from 3 cases for each of the Gleason-score classes (2, 3 and 4-5) were selected as data for model generation, whereas patches belonging to the remaining cases from each class were used for model evaluation.

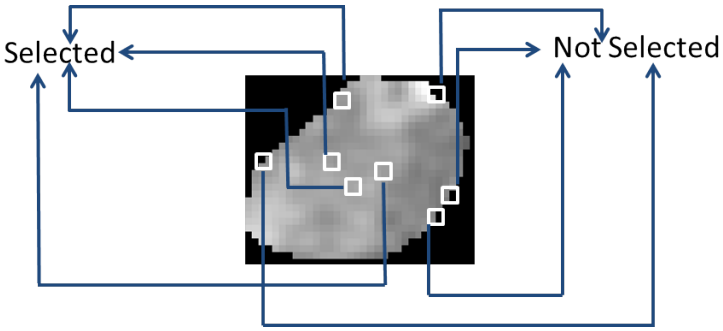

Fig. 2: The process of data extraction from the annotated tumour areas.

2.2 Dictionary Generation

After extracting the patches, the next step is to generate the model that is used for feature generation. For dictionary generation, K-Means clustering was used, where the extracted patches from the MRI data were used for model generation. The initial value of K for applying K-Means clustering was set to be 10 (the value was selected based on empirical evaluation to get the optimum value of K). After this step, we obtained 10 clusters for each of the Gleason grades i.e. 2, 3 and 4-5. The final dictionary is the combination of all cluster centroids for all three classes, resulting the size of final dictionary to be equal to 30. These 30 cluster centroids are then referred to as codewords of the dictionary.

2.3 Voting-Based Encoding

For the model evaluation, we used the cases that are different from those used in Section 2.2. A Voting-based encoding method was used [9] to encode the image patch data. A soft-assignment strategy was used, where each image patch is transformed to a normalized weight vector. The SA-k (Soft-Assignment k) strategy was used based on the weights of the k-nearest codewords [3], while encoding the image patch. The normalized weight w_i for raw image patch P corresponding to codeword C_i is then defined as:

$$w_i = \frac{exp(-\beta \left\| P - C_i \right\|_2^2)}{\sum_{j=1}^{k} exp(-\beta \left\| P - C_j \right\|_2^2)} \tag{1}$$

where β is set to be equal to 1 based on empirical evaluation. The normalized weights are computed for all the codewords from the dictionary for each of the raw image data corresponding to image patch P and then only the k-closest weights are used as encoded feature. The final encoded feature $\hat{P}$ is then showed as:

$$\hat{P} = [w_1, w_2, w_3, ...w_k] \tag{2}$$

where k is the total number of closest neighbours that were used for encoding the image patch P. The value of k is set to be equal to 25 based on experimental

evaluation. After this encoding, each image data corresponding to the image patch was represented as a feature vector of length 25.

3 Model Evaluation

In this section, we evaluated the model developed in Section 2.2 using k-length encoded features (Section. 2.3). As the classification results of the classifier were influenced by how balanced the data was for each class, we made the data balanced before evaluating the performance of the developed model. In order to balance the data, we investigated the least number of samples present in any class. For class 2 the total number of generates samples was 410, therefore, we select 500 samples from the other two classes, in order to make data balance in a reasonable way. For experimental evaluation, we used Weka [6], version 3.7.2. We executed five machine learning classifiers available in Weka namely Neural Network (NN), Vector Quantization Neural Network (VQNN), K Nearest Neighbour (KNN), Random Forest (RF) and Support Vector Machine (SVM). For NN, VQNN (In Weka both NN and VQNN are implementated as Fuzzy-rough K-nearest neighbours classifier that uses the nearest neighbours to construct lower and upper approximations of decision classes [7]) and KNN classifiers value of k is set to be equal to 5, whereas for SVM the filter type is set to be as standardize the training data. The rest of the settings for RF and for other four classifiers are set to default values in Weka. 10 Fold Cross Validation (FCV) scheme is used for results evaluation. The classification results using these classifiers can be found in Table 1. The results are given as Maximum (Max) Classification Accuracy (CA) ±, Average (Avg) CA, True Positive (TP) rate, False Positive (FP) rate, Precision, Recall, F-Measure (or F1 score) and Area Under ROC Curve (AUC).

Classifier	Max. CA	Avg. CA	TP Rate	FP Rate	Precision	Recall	F-Measure	AUC
NN	88.65 %	84.68 %	0.847	0.083	84.7 %	84.7 %	0.847	0.932
VQNN	88.65 %	84.68 %	0.847	0.083	84.7 %	84.7 %	0.847	0.93
KNN	89.36 %	84.32 %	0.843	0.085	84.2 %	84.3 %	0.843	0.933
RF	87.94 %	83.40 %	0.834	0.09	83.4 %	83.4 %	0.834	0.936
SVM	82.97 %	80.35 %	0.804	0.105	80.5 %	80.4 %	0.804	0.867

Table 1: Classification results for five different classifiers.

The confusion matrices for all the five classifiers (for the best fold) can be found in Table 2.

3.1 Results for 2-class Classification

In order to check the validity of the developed model for 2-class classification, we redesign the defined classes by combining Gleason score 2 and 3 as one class

Class	**2**	**3**	**4-5**
2	402	2	6
3	0	408	92
4-5	9	107	384

(a)

Class	**2**	**3**	**4-5**
2	402	2	6
3	0	408	92
4-5	9	107	384

(b)

Class	**2**	**3**	**4-5**
2	403	1	6
3	1	406	93
4-5	14	106	380

(c)

Class	**2**	**3**	**4-5**
2	401	1	8
3	2	390	108
4-5	8	107	385

(d)

Class	**2**	**3**	**4-5**
2	384	8	18
3	3	369	128
4-5	23	97	380

(e)

Table 2: Confusion matrices corresponding to all fine classifiers: (a) NN, (b) VQNN, (c) KNN, (d) RF and (e) SVM.

(Gleason score 2-3) and the second class is same (i.e. 4-5) as we defined for 3-class classification. The classification results for this 2-class classification is even improved then those achieved with 3-class classification. Th training is performed as has been done for 3-class Gleason score, whereas at the model evaluation step Gleason scores 2 and 3 are merged as one class. An average classification accuracies of 88%, 87%, 86%, 87% and 82% has been achieved by using NN, VQNN, KNN, RF and SVM classifiers. The settings of these classifiers are set to same as that of 3-class Gleason score classification. These initial classification results shows the validity of the developed model for both 2-class and 3-class Gleason score classification.

4 Discussion

For the proposed Gleason score classification approach , the model was build on the training data and composed of cluster centroids for each individual class. The final features used for the classification were the voting-based encoding for each of the patches (used as basic input throughout the approach). The initial classification reveals satisfactory results (best accuracy achieved 85% with AUC equal to 0.93) for the 3-class classification problem. Some correctly classified encoded features for the three Gleason score classes can be seen from Figure 3.

As can be seem from Figure 3, the encoded features from each of the classes 2,3 and 4-5 appear to be different from each other in terms of encoding values for each of the codeword.

4.1 Comparison with Existing Techniques

In this section, we compare the proposed method with the existing approaches for Gleason score classification for prostate cancer. Most of the existing approaches for Gleason score classification are based on two class classification.

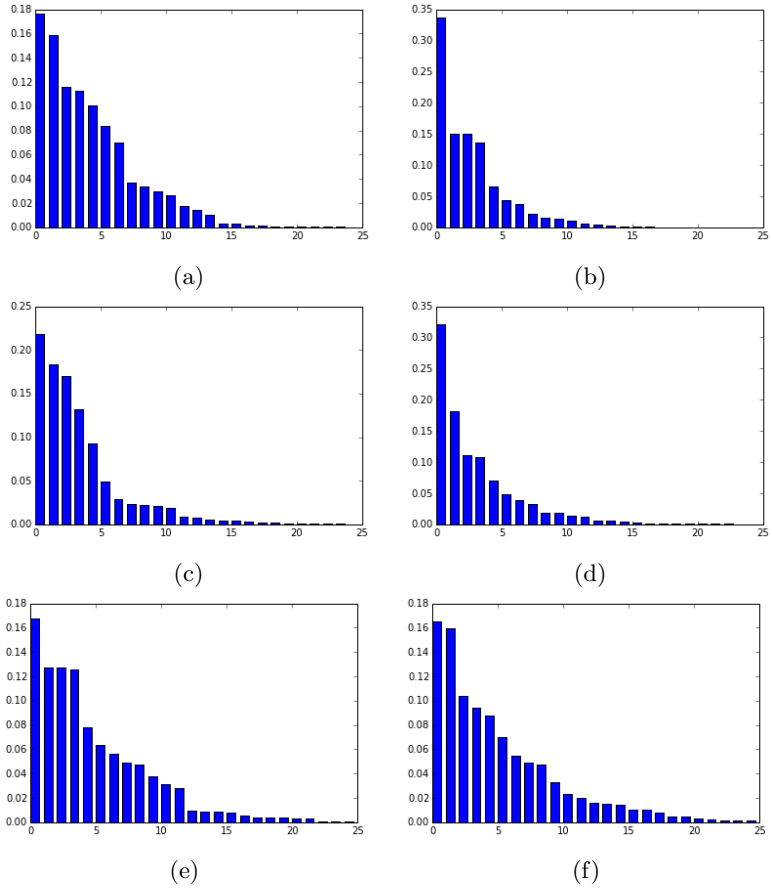

Fig. 3: Randomly selected features for all the three Gleason score classes: x-axis is showing the count of nearest codewords (i.e. 25) and y-axis is showing their normalized weights. (a) and (b) are showing the encoded features for Gleason score 2, (c) and (d) representing encoded features for Gleason score 3, whereas (e) and (f) shows encoded features for Gleason score 4-5. As can be seen from the encoded features of all the three Gleason score classes, the features are representing distinct information for each class.

For example, Tiwari *et al.* [16] proposed a method of two class classification of Gelason score for prostate cancer, where they reported AUC=0.84. Similarly, Tabesh *et al.* [14] provided a solution for classification of Gleason score into low and high grade, and reported classification accuracy equal to 77.6%. In contrast to these works, our proposed method provides three class Gleason-score classification. Our method have exhibited an improved classification accuracy of 84.68 % and achieved AUC= 0.932. Thus our proposed method have shown excellent performance compared to the existing approaches.

4.2 Investigating the Misclassified Instances

From the confusion matrices that have been shown in Table 2, it would be interested to look at the misclassified instances across all the classifiers. In total there were 216 misclassified instances for classifiers NN and VQNN, for KNN misclassified instances were 221, 234 instances were misclassified for the RF classifier and 277 instances were misclassified by the SVM classifier. We selected only those misclassified instances that appeared within all five classifiers, which turns out to be 107. This also shows the different behaviours of the classifiers to build the decision boundaries that resulted is different misclassified instances for all classifiers. The distribution of the 107 overlapped misclassified instances can be found in Table 3.

Class	**2**	**3**	**4-5**
2	-	0	1
3	0	-	53
4-5	6	47	-

Table 3: Overlapped misclassified instances across all five classifiers

In total there are 4 misclassified categories: instances originally belonging to Gleason score 2 but classified as Gleason score 4-5, instances originally belonging to Gleason score 3 but classified as Gleason score 4-5, instances that belonging to Gleason score 4-5 but classified as Gleason Score 2 and the last category where the instances belonging to Gleason score 4-5 but classified as Gleason score 3. It should be noted that most of the confusion is between classes 3 and 4-5 (and vice-versa). Figure 4 (a) is showing a encoded features of the misclassified instance that originally belongs to Gleason score 2, but because of the similar distribution of this encoded feature with the features from Gleason score 4-5 (Figure 3 (e) and (f)), this instance is classified as Gleason score 4-5. From Figure 3 (e) and (f), it can be seen that the overall frequency of initial bins are high, a similar pattern can be found in Figure 4 (a). The misclassified instance shown in Figure 4 (c) (Gleason score 4-5 is classified as Gleason score 2), an instant decrease in the frequencies of the initial bins is a similar pattern to the encoded features belonging to Gleason score 2 (Figure 3 (a) and (b)). As already

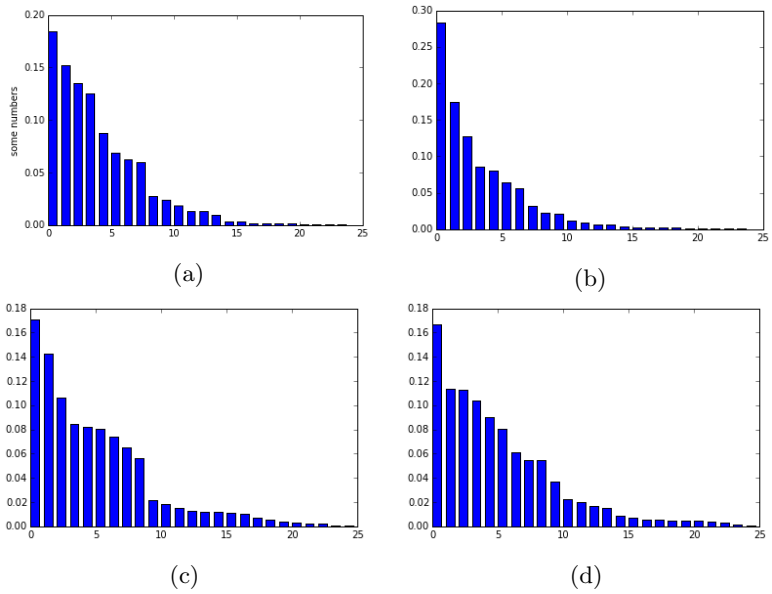

Fig. 4: Misclassified instances: along x-axis the count of nearest codewords (i.e. 25) have been shown and y-axis is representing their normalized weights.(a) Gleason Score 2 in classified as Gleason Score 4-5, (b) Gleason Score 3 in classified as Gleason Score 4-5, (c) Gleason Score 4-5 in classified as Gleason Score 2, (d) Gleason Score 4-5 in classified as Gleason Score 3

explained, the most of the confusion exists between Gleason scores 3 and 4-5 (see Table 3), misclassified instances shown in Figure 4 (b) and (d) (Gleason score 3 in classified as Gleason score 4-5 and Gleason score 4-5 is classified as Gleason score 3) is due to the similar appearance of the encoded features of both classes.

5 Conclusion

In this work, we proposed a novel voting-based encoding approach for Gleason scores classification of prostate tumours. The classes of Gleason score made under considerations are 2, 3 and Gleason scores 4 and 5 are taken as single class i.e. 4-5. After generating a dictionary from all the training data, soft assignment was used to assign each of image patch to the closest code word from the dictionary. The study proposed a novel way to transform the data from the image domain to a more discriminative representation. Initial classification reveals a satisfactory classification accuracy (85%) and AUC (0.932).

6 Future work

The current work is based of the T2-w MRI data for the classification of prostate cancer tumour aggressiveness. In future we will extend this work to combine the features from other modalities of MRI images (Dynamic Contrast Enhanced (DCE) MRI and Apparent Diffusion Coefficient (ADC) maps) further to improve the classification accuracy. In addition, due to limited size of the data we combined the Gleason scores 4 and 5 as one class (i.e. 4-5), in future based on the availability of more data we will perform four class classification (Gleason scores 2,3,4 and 5) using the same approach we did for three Gleason score classification.

References

1. Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., Tomaszeweski, J.: Automated grading of prostate cancer using architectural and textural image features. In: Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on. pp. 1284–1287. IEEE (2007)
2. Fehr, D., Veeraraghavan, H., Wibmer, A., Gondo, T., Matsumoto, K., Vargas, H., Sala, E., Hricak, H., Deasy, J.: Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. Proceedings of the National Academy of Sciences 112(46), E6265–E6273 (2015)
3. Gemert, J.V., Veenman, C., Smeulders, A., Geusebroek, J.: Visual word ambiguity. IEEE transactions on pattern analysis and machine intelligence 32(7), 1271–1283 (2010)
4. Grönber, H.: Prostate cancer epidemiology. The Lancet 361(9360), 859–864 (2003)
5. Grossfeld, G., Carroll, P.: Prostate cancer early detection: a clinical perspective. Epidemiologic reviews 23(1), 173–180 (2001)
6. Holmes, G., Donkin, A., Witten, I.: Weka: A machine learning workbench. In: Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on. pp. 357–361. IEEE (1994)
7. Jensen, R., Cornelis, C.: A new approach to fuzzy-rough nearest neighbour classification. In: 6th International Conference on Rough Sets and Current Trends in Computing, LNAI. pp. 310–319 (2008)
8. Khurd, P., Bahlmann, C., Maday, P., Kamen, A., Gibbs-Strauss, S., Genega, E., Frangioni, J.: Computer-aided gleason grading of prostate cancer histopathological images using texton forests. In: Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on. pp. 636–639. IEEE (2010)
9. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding 150, 109–125 (2016)
10. Rampun, A., Tiddeman, B., Zwiggelaar, R., Malcolm, P.: Computer aided diagnosis of prostate cancer: A texton based approach. Medical Physics 43(10), 5412–5425 (2016)
11. Ransohoff, D., Collins, M., Fowler, F.: Why is prostate cancer screening so common when the evidence is so uncertain? a system without negative feedback. The American Journal of Medicine 113(8), 663–667 (2002)
12. Siegel, R., Miller, K., Jemal, A.: Cancer statistics, 2015. CA: a Cancer Journal for Clinicians 65(1), 5–29 (2015)
13. Suhail, Z., Hamidinekoo, A., Denton, E., Zwiggelaar, R.: A texton-based approach for the classification of benign and malignant masses in mammograms. In: Annual Conference on Medical Image Understanding and Analysis. pp. 355–364. Springer (2017)
14. Tabesh, A., Kumar, V., Pang, H., Verbel, D., Kotsianti, A., Teverovskiy, M., Saidi, O.: Automated prostate cancer diagnosis and Gleason grading of tissue microarrays. In: Medical Imaging 2005: Image Processing. vol. 5747, pp. 58–71. International Society for Optics and Photonics (2005)
15. Tabesh, A., Teverovskiy, M., Pang, H., Kumar, V., Verbel, D., Kotsianti, A., Saidi, O.: Multifeature prostate cancer diagnosis and Gleason grading of histological images. IEEE Transactions on Medical Imaging 26(10), 1366–1378 (2007)
16. Tiwari, P., Kurhanewicz, J., Madabhushi, A.: Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS. Medical Image Analysis 17(2), 219–235 (2013)