

Automated High-Grade Prostate Cancer Detection and Ranking on Whole Slide Images

Chao-Hui HUANG^{a,b} and Daniel RACOCEANU^{c,d}

^aMSD International GmbH (Singapore Branch)*

^bAgency for Science, Technology and Research, Singapore

^cPontifical Catholic University of Peru, San Miguel, Lima 32, Peru

^dSorbonne Universités, UPMC Univ Paris 06, CNRS, INSERM, 75013, Paris, France

ABSTRACT

Recently, digital pathology (DP) has been largely improved due to the development of computer vision and machine learning. Automated detection of high-grade prostate carcinoma (HG-PCa) is an impactful medical use-case showing the paradigm of collaboration between DP and computer science: given a field of view (FOV) from a whole slide image (WSI), the computer-aided system is able to determine the grade by classifying the FOV. Various approaches have been reported based on this approach. However, there are two reasons supporting us to conduct this work: first, there is still room for improvement in terms of detection accuracy of HG-PCa; second, a clinical practice is more complex than the operation of simple image classification. FOV ranking is also an essential step. *E.g.*, in clinical practice, a pathologist usually evaluates a case based on a few FOVs from the given WSI. Then, makes decision based on the most severe FOV. This important ranking scenario is not yet being well discussed. In this work, we introduce an automated detection and ranking system for PCa based on Gleason pattern discrimination. Our experiments suggested that the proposed system is able to perform high-accuracy detection ($\sim 95.57\% \pm 2.1\%$) and excellent performance of ranking. Hence, the proposed system has a great potential to support the daily tasks in the medical routine of clinical pathology.

1. INTRODUCTION

Digital pathology has been benefiting from the technologies of computer vision and machine learning. Integrating pathology informatics into clinical practices is important in the future of pathology. In a consistent series of cancers, accurate cancer diagnosis is based on the fine needle biopsy. Often, due to inter- and intra- observer variability, even with experienced pathologists, agreement in prostate cancer Gleason grading could be as low as 70%.¹ As a result, a reliable and consistent computer-aided diagnosis (CAD) method for interpretation of prostate histopathology can become an essential tool in clinical practice. As long as proper training patterns are provided, the CAD method can generate critical information for rapid routine clinical reporting. Further, a range of research applications, medical training and clinical studies also can be positively impacted.

Since the capabilities of computation and data storage have been boosted, many cutting-edge technologies of CAD methods were recently proposed. Most of them are based on haematoxylin and eosin (H&E) stained whole slide images (WSIs), *e.g.*, the method proposed by Weingant *et al.* in 2015.² They suggested high- and low-grade prostate cancer (HG-PCa & LG-PCa) classification based on H&E WSIs via stain normalization and cell density estimation. As they have reported, based on their own private image set, the performance in the form of AUC [the area under the receiver operating characteristic (ROC) curve], is about 0.703 – 0.705; DiFranco *et al.*³ proposed a whole-slide prostate cancer probability map using color texture features; Huang *et al.* mentioned a combination of support vector machines (SVMs) and texture fractal analysis. The accuracy of tumor prediction on their own private image set can be as good as $93.7\% \pm 3.3\%$.⁴

Most of the existing algorithms are focusing on discriminating high-grade H&E image patterns from the given input images. Many of them achieved 90% in terms of accuracy, especially, when involving their own private image sets. However, a clinical practice is more complicate than an operation of image classification. That is,

* Current affiliation.

in the clinical practice, a pathologist selects a few regions of interest (ROIs), named fields of view (FOVs) and evaluate them in high-magnification (*e.g.*, $10\times$ as suggested in Gleason grading protocol) and the final decision will be made based on the most severe FOVs.⁵ In other words, the pathologist ranks the chosen FOVs.

It is interesting to see the rank of each ROI on the given WSIs, *e.g.*, a Google-like search engine of ROIs, which is able to give the ranks of high-grade images. Fortunately, it is possible to realize such kind of system by combining a series of learners, *e.g.*, ranking boosting (RankBoost).⁶ Based on this concept, we propose a detection and ranking method which aims to mimic the practical scenario of pathological diagnosis.

The proposed method combines 3 major components, including: 1) H&E image sampling based on nuclei density method, 2) H&E image feature extraction and 3) HG-PCa/LG-PCa ranking algorithm. The nuclei density based H&E image sampling method is a strategy of image sampling on WSIs. The sampling method identifies high nuclei density regions, as the high nuclei density regions are more relevant to cancerous areas due to the nature of cancer. The H&E image feature extraction converts a H&E image patch into a numerical feature vector which is considered as the feature descriptor (FD) for the H&E image patch. HG-PCa/LG-PCa ranking algorithm is a method that can rank a set of given image patches based on their FDs, *e.g.*, a set of ROIs. The ranking algorithm is able to assign the order of ranks for all given ROIs. Hence, a pathologist can evaluate the locations of these ROIs accordingly.

Instead of using private image sets, we used a public dataset provided by The Cancer Genome Atlas (TCGA) due to the availability of large amount of cancerous cases. TCGA is a comprehensive and coordinated effort to accelerate our understanding of the cancer through the application of medical/biomedical data analysis technologies, including large-scale genomes sequencing, medical reports and WSI analyzing. In the experiments, we used the WSIs of prostate adenocarcinoma cases in TCGA and their medical reports as the training patterns.

In this paper, Section 2 introduces the proposed method, including the proposed approach using nuclei density based H&E image sampling method, H&E image feature extraction and HG-PCa/LG-PCa ranking algorithm. Their performances are discussed in Section 3, including the presentation of the dataset used in this study. Finally, in Section 4, we conclude our study by detailing some possible future research directions.

2. METHODS

2.1 Hematoxylin & Eosin (H&E) Image Sampling based on Nuclei Density

Almost all computer vision applications are facing on a challenge called the curse of dimensionality, since a image usually contains more than 10^3 to 10^6 pixels. In the case of whole slide images, the situation is even worse: the number of pixels can easily exceed 10^9 . As a result, analyzing such kinds of images directly is often practically infeasible.

This is where the concept of image local patch sampling becomes useful. Instead of analyzing a global image, we prefer to look into a set of smaller local patches cropped from the global image, since the local patches are often experiencing less distortion than the global image. As a result, it is easier to define the similarity between two sets of local patches. In the last decade, this idea has been widely investigated.^{7,8}

Due to the “curse of dimensionality”, analyzing full-size images directly is often practically infeasible. Local image patch sampling methods thus become a famous topic in the field of image processing.⁹

Tile-based local image patch sampling and random local image patch sampling are two well-known methods for sampling local image patches. Tile-based local image patch sampling is often used in the field of image compression, as the data redundancy is minimized. However, for the fields of object recognition, this is not necessary to be a good idea, since a targeting object is not always aligned on the grid of tile-based sampling. Random local image patch sampling is often used when there is no specific targeting object, *e.g.*, texture analysis. However, it can generate a certain amount of redundant data.⁹

Another method for sampling local image patches is content-based local image patch sampling, which is more suitable for our needs since the certain pathological information can be used. The proposed method is not based on tile-based image method. Instead, it uses pathological information provided by the image itself. First, we perform stain deconvolution in order to obtain the staining intensity of hematoxylin. Hematoxylin colors nuclei of cells (as well as other objects, such as keratohyalin granules and calcified material) blue. In most of cases,

Algorithm 1 Nuclei Density based Image Sampling

```
1: procedure
2: Input:
3:   Input image:  $I = \{[I]_{x,y} | [I]_{x,y} = (r, g, b), 1 \leq x \leq M, 1 \leq y \leq N\}$ , where  $M$  is image width,  $N$  is height
   and  $(r, g, b)$  represents a point in the color space of red, green and blue.
4:    $S$ :  $\frac{1}{2}$  of image width (and height).
5: Initial:
6:    $J = \text{nuclei\_detection}(I)$ , where
       
$$J = \{[J]_{x,y} | [J]_{x,y} \in \{1, 0\}, 1 \leq x \leq M, 1 \leq y \leq N\}, \text{ and}$$

       
$$[J]_{x,y} = \begin{cases} 1, & \text{if } (x, y) \text{ belongs to a nucleus,} \\ 0, & \text{otherwise.} \end{cases}$$

7: Begin:
8:   for  $t = 1$  to  $T$  do
9:      $K = \text{kernel\_density\_estimation}(J)$ , where  $K$  is the probability density function of  $J$ ,  $K = \{[K]_{x,y} | [K]_{x,y} \in \mathbb{R}_{\geq 0}, 1 \leq x \leq N, 1 \leq y \leq M\}$ .
10:    Remove padding of  $K$  (for preventing sampling beyond image boundaries):  $[K]_{x,y} \leftarrow 0$ , if  $x \leq S$  or  $x \geq (M - S)$  or  $y \leq S$  or  $y \geq (N - S)$ .
11:    Find the global maximum point of  $[K]_{x,y}$ :  $(x'_t, y'_t) = \arg \max_{(x,y)} ([K]_{x,y})$ .
12:    Break the for-loop from here if necessary, e.g.,  $(x'_t, y'_t) = (x'_{t-1}, y'_{t-1})$ .
13:    Fetch an image:  $P_t \leftarrow \{[I]_{x,y} | (x'_t - S) \leq x \leq (x'_t + S), (y'_t - S) \leq y \leq (y'_t + S)\}$ .
14:    Remove the neighboring nuclei location points:  $[J]_{x,y} \leftarrow 0$ , if  $(x'_t - S) \leq x \leq (x'_t + S)$  and  $(y'_t - S) \leq y \leq (y'_t + S)$ .
15:   end for
16: Output: image list  $[P_1, \dots]$ .
17: end procedure
```

pathologists are more interested on the locations which contain more number of nuclei. As a result, sampling images from these locations is more meaningful for a CAD system. Many methods are ready for doing this as Veta *et al.* mentioned.¹⁰ The scenario of the nuclei density based H&E image sampling method is: first, for the given WSI, we use an existing method to detect the nuclei location points. Fortunately, many methods are ready for doing this as Veta *et al.* mentioned.¹⁰

Next, we use kernel density estimation (KDE) to estimate the nuclei density map on the given WSI. We can identify the location of the global maximum point on the nuclei density map. The location is used to define a region for sampling an image. Then, we remove the neighboring nuclei location points of the global maximum point. This procedure is repeated until all stopping criteria are satisfied, *e.g.*, the number of images is satisfied. For the further details, please refer to the algorithm shown in Alg. 1.

2.2 Hematoxylin & Eosin (H&E) Image Feature Extraction

The feature extraction method for H&E images includes two steps: stain deconvolution and feature description. The purpose of stain deconvolution is to extend the data dimensions in color space as there are significant benefits for the feature extraction.¹¹ Other than the original red, green and blue channels, the channels such as luminance, haematoxylin staining and eosin staining can also be used. There are many efficient methods can be considered for stain deconvolution, *e.g.*, Macenko's method.¹¹

The next step is to define the feature descriptor (FD) for each given image. FD should be correlated with essential indicators, such as Gleason patterns. In this sense, various methods have been proposed in the last decade, *e.g.*, Gabor Filtering,¹² Features of Fractal Dimension by Differential Box-Counting (DBC)⁴

- **Gabor Filtering:**

Gabor filtering bank set is a set of linear filters used for texture detection. They have been found to be particularly appropriate for texture representation and discrimination.¹²

Algorithm 2 HG-PCa/LG-PCa Ranking Algorithm

```
1: procedure
2:   Input:
3:     Training patterns:  $\mathbf{X} = \{(\mathbf{x}_i, l_i) | 1 \leq i \leq n, \mathbf{x}_i \in \mathbf{R}^{M \times N}, l_i \in \{0, 1\}\}$ .
4:     Pattern weights:  $w_{1,1}, \dots, w_{n,n} \in \mathbf{R}_{\geq 0}$ .
5:     Features:  $d_1(\cdot), \dots, d_m(\cdot)$ , each  $d_i(\cdot) \in \mathbf{R}^{D_i}$ .
6:     Weak learners:  $h_1(\cdot), \dots, h_p(\cdot) \in \{1, 0\}$ .
7:   Initial:
8:     Each  $w_{q,r,1} = 1/n^2$ .
9:     Weak learner pool:  $\mathcal{H} = \{h_1(d_1(\cdot)), \dots, h_p(d_m(\cdot))\}$ .
10:  Begin:
11:    for  $t = 1$  to  $T$  do
12:      Divide training patterns into  $K$  parts:  $\mathbf{X} = \{X_1, \dots, X_K\}$ .
13:      for  $k = 1$  to  $K$  do  $K$ -fold cross validation
14:        Train  $h_i(d_j(\mathbf{x}_q))$  and  $h_i(d_j(\mathbf{x}_r))$ ,  $\forall i, j$  and each  $\mathbf{x}_q$  and  $\mathbf{x}_r \notin X_k$ 
15:        Compute  $e_{i,j,k} = \sum_{q,r} w_{q,r,t} (h_i(d_j(\mathbf{x}_q)) - h_i(d_j(\mathbf{x}_r))) \forall i, j$  and each of  $\mathbf{x}_q$  and  $\mathbf{x}_r \in X_k$ .
16:      end for
17:       $E_{i,j} = \sum_{k=1}^K e_{i,j,k}, \forall i, j, \quad i'_t, j'_t = \arg \max_{i,j} E_{i,j} \quad \text{and} \quad \alpha_t = \frac{1}{2} \log \frac{1+E_{i'_t, j'_t}}{1-E_{i'_t, j'_t}}$ .
18:      Update  $w_{q,r,t+1} \leftarrow \frac{1}{Z_t} w_{q,r,t} \exp(\alpha_t (h_{i'_t}(d_{j'_t}(\mathbf{x}_q)) - h_{i'_t}(d_{j'_t}(\mathbf{x}_r))))$ , where  $Z_t$  is a normalizer for all
         $w_{q,r,t+1}$  such that all  $w_{q,r,t+1}$  will be a distribution.
19:      Remove  $h_{i'_t}(d_{j'_t}(\cdot))$  from  $\mathcal{H}$ 
20:    end for
21:  Output:  $s(\mathbf{v}) = \sum_{t=1}^T \alpha_t h_{i'_t}(d_{j'_t}(\mathbf{v}))$  for input  $\mathbf{v}$ .
22:    The larger output of  $s(\mathbf{v})$ , the higher rank.
23:    The sign of  $s(\mathbf{v})$  is the classification result of  $\mathbf{v}$ .
24: end procedure
```

- **Features of Fractal Dimension by Differential Box-Counting (DBC):**

DBC has been widely used in the fields of image classification since Sarkar *et al.* first introduced it and then Huang *et al.*⁴ proposed an automatic classification for H&E prostate cancer images.

- **Entropy-based Fractal Dimension Estimation (EBFDE):**

EBFDE is an alternative method for computing fractal dimension based on entropy. This has first been proposed by Huang *et al.*⁴

There are also some other feature extraction methods,^{3,4} *e.g.*, multi-wavelet methods, histogram statistics methods, and gray-level co-occurrence based methods, *etc.* Although they have not been implemented in our system. However, for the sake of expanding the diversity of weak learners, it will be important to include more feature description methods in the future.

2.3 High Grade / Low Grade Prostate Cancer (HG-/LG-PCa) Ranking Algorithm

The proposed ranking algorithm is an enhanced version of the RankBoost proposed by Freund *et al.*⁶ The RankBoost method is an ensemble method which uses multiple learning algorithms (called weak learners) to obtain a better performance of the final strong ranker. The more diverse weak learners, the better performance of the final strong ranker. A weak learner includes a feature descriptor and a binary classifier. The binary classifier is trained to fit the preference of the input data. In our case, HG-PCa is more preferable than LG-PCa.

The pool of weak learners is composed by the combination of the FD mentioned in the previous section and various types of existing classifiers, including: k-nearest neighbors, support vector classifiers, decision trees, adaptive boosting, Gaussian Naïve Bayes, linear discriminant analysis, quadratic discriminant analysis, as well as these classifiers with different parameters.

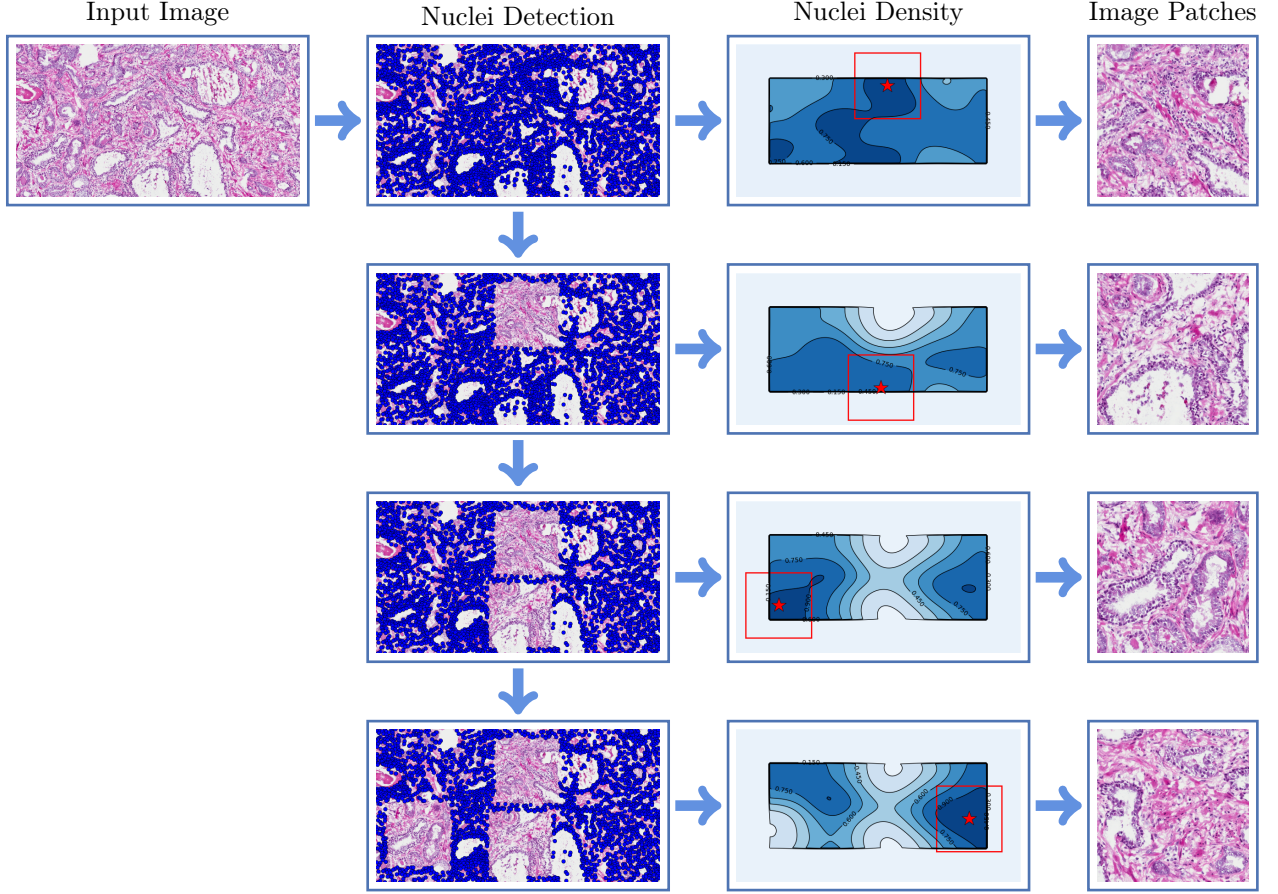


Figure 1. The procedure of nuclei density based H&E image sampling algorithm: the 1st column is an input image. The 2nd column is the nuclei detection representing in blue dots. The 3rd column is the nuclei density map using KDE (see text), in which, darker areas represent higher nuclei density areas and vice versa. The surrounding low density area is the result of removing the padding area of the density map in order to prevent sampling beyond the image boundaries. On the density map, a range of image sampling area (shown as a red rectangle) is selected around the global maximum point (shown as a red star). The 4th column is the image patch cropped from the image sampling area. After each step of image sampling, the neighboring nuclei location points are removed. Then the procedure is repeated until all stopping criteria are satisfied.

In the training phase, during each step of iteration, a weak learner is selected with a weight based on its performance. Over a certain number of iterations, a group of weak learners and their weights are determined. In the testing phase, given a testing pattern, the summation of the output of these weak learners (with the corresponding weights) will be able to give an output for ranking. The actual value of the output is not important. Instead, we are more interested in the comparison between the output values of two testing patterns. In other words, given a set of testing patterns, one can find the order of them by sorting their output values from the strong ranker. The full scenario of the proposed ranking algorithm can be found in Alg. 2.

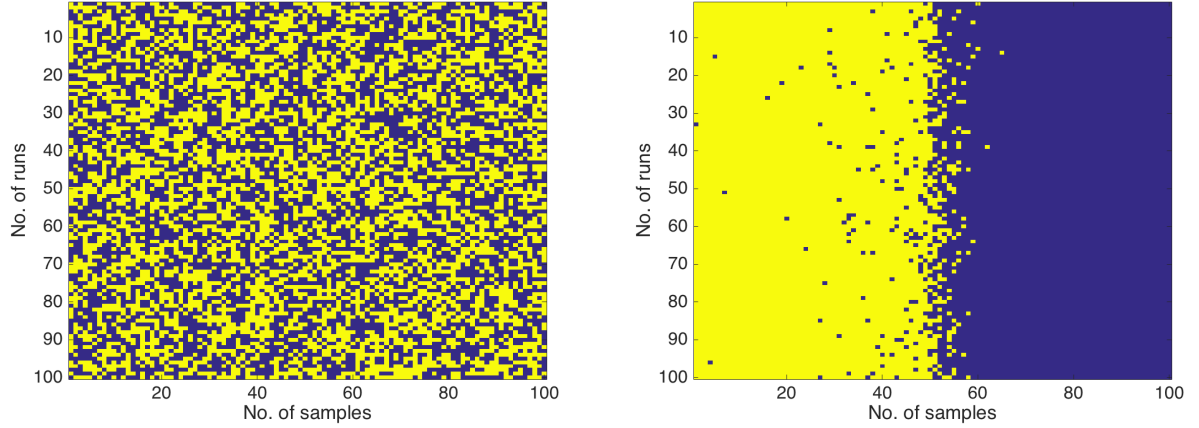
3. DATASET AND RESULTS

Due to the availability of large amount of cancerous cases, we used a public dataset provided by The Cancer Genome Atlas (TCGA). Most of the Gleason pattern types are between 3-5 and the available Gleason scores are from 6 to 10. The cases were selected based on image quality and the balance between HG-PCa (Gleason score ≥ 7) and LG-PCa (Gleason score ≤ 6).

BCR Patient Barcode	Gleason Score	BCR Patient Barcode	Gleason Score
TCGA-G9-7523	10	TCGA-2A-A8VL	6
TCGA-XQ-A8TA	10	TCGA-2A-A8VO	6
TCGA-EJ-5507	9	TCGA-2A-A8VV	6
TCGA-HI-7171	9	TCGA-2A-AAYO	6
TCGA-EJ-A46G	8	TCGA-2A-AAYU	6
TCGA-G9-7510	8	TCGA-EJ-5517	6
TCGA-EJ-7314	7	TCGA-EJ-7321	6
TCGA-EJ-7315	7	TCGA-G9-6342	6

(a) HG-PCa cases. (b) LG-PCa cases.

Figure 2. The dataset obtained from TCGA.



(a) The distribution of HG-/LG-PCa samples over 100 runs before ranking. (b) After ranking, most of HG-PCa samples are “raised” to higher ranks.

Figure 3. The change of distribution before and after ranking. In the plots, the x-axes represent the samples and the y-axes are the runs of experiments. Yellow represents HG-PCa and blue is LG-PCa.

3.1 Experiments and Breakthrough Results

We tested the system 100 runs and obtained the average performance. In each run, we randomly selected 100 images (50 high-grade and 50 low-grade) as the testing patterns from all 682 images. The rest of 582 images were used as the training patterns. In the testing phase, we shuffled these 100 testing images such that the information of the label of each individual image was excluded. Then, these 100 images were classified and ranked by the proposed method. The results of detection and ranking are shown as follows:

- **Detection:** Over all 100 ranks of all 100 runs, the mean AUC was 0.9486 ± 0.005 and the mean accuracy achieved $95.57\% \pm 2.1\%$. The average of classification accuracy of each rank is shown in Fig. 4(a). One can see that due to the ranking operation, most of the mis-classified images is located about the region of 50-51 of all 100 ranks as this region is the transition between high-grade and low-grade after ranking. We compared our algorithm with 2 cutting-edge methods of detection: first, the method proposed by Weingant *et al.*,² the AUC is about 0.703 – 0.705. Second, the method proposed by Huang *et al.*,⁴ the accuracy is $93.7\% \pm 3.3\%$. One can find significant improvement in our approach. Note that the methods that we compared were evaluated using private datasets provided by their corresponding institutes. They were not available to the public for more precise comparison. On the other hand, our method is evaluated based on a public dataset. Thus, the results are objectively repeatable and the method is objectively open for further adaptations and improvements.
- **Ranking:** The probability that the testing image patch of a rank is high-grade is shown in Fig. 4(b). We computed the probability that the image which is assigned to a specific rank is high-grade. It is clearly

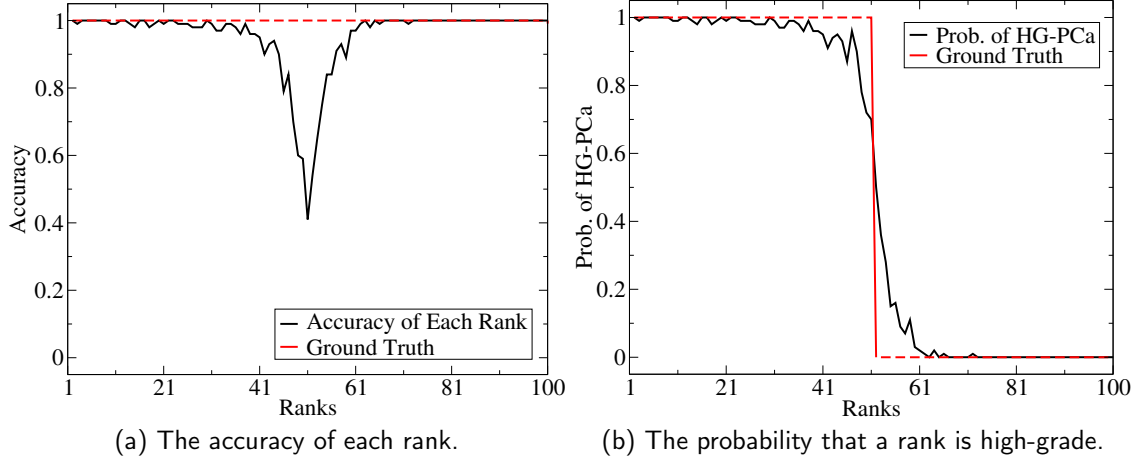


Figure 4. The results of detection and ranking 100 (50 HG- and 50 LG-) randomly selected PCa images. The x-axes are the ranks after sorting. Higher rank means higher probability to be high-grade.

that the testing patterns are ranked properly as the testing image in a higher rank has higher probability to be high-grade. Some examples are shown in Fig. 5. For the comparison with cutting-edge methods of ranking, to the authors' knowledge, our work is the first one involving ranking WSI images for high-grade PCa as we were not able to find any relevant existing work.

4. CONCLUSIONS

An automated prostate cancer computer-aided diagnosis system is reported. The system is able to evaluate high-grade prostate cancer (HG-PCa) on whole slide images (WSIs), based on the ability of mimicking the operation of the pathologist in a clinical practice, including detection and ranking fields of view (FOVs). The system has been tested using a public dataset, The Cancer Genome Atlas (TCGA). By comparing the system with existing methods, we highlighted significant improvements in terms of high accuracy of detection and excellent performance of ranking. Hence, the proposed method has a great potential to support pathologists in their daily practice.

REFERENCES

- [1] W. C. Allsbrook Jr and Others, "Interobserver reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists," *Hum Pathol.*, vol. 32, no. 1, pp. 74–80, 2001.
- [2] M. Weingant, H. M. Reynolds, A. Haworth, C. Mitchell, S. Williams, and M. D. DiFranco, "Ensemble Prostate Tumor Classification in H&E Whole Slide Imaging via Stain Normalization and Cell Density Estimation," in *6th Intl. Workshop, Machine Learning in Medical Imaging (Held in Conj. with MICCAI)*, 2015, pp. 280–287.
- [3] M. D. DiFranco and Others, "Ensemble based system for whole-slide prostate cancer probability mapping using color texture features," *Comput Med Imaging Graph.*, vol. 35, no. 7–8, pp. 629–645, 2011.
- [4] P. W. Huang and C. H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Trans. Med. Im.*, vol. 28, no. 7, pp. 1037–1050, 2009.
- [5] J. I. Epstein, W. C. A. Jr, M. B. Amin, L. L. Egevad, and I-G-Committee, "The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma," *Am J Surg Pathol.*, vol. 29, no. 9, pp. 1228–1242, 2005.
- [6] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, pp. 933–969, 2003.
- [7] M. N. Gurcan and Others, "Histopathological image analysis: a review," *IEEE Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

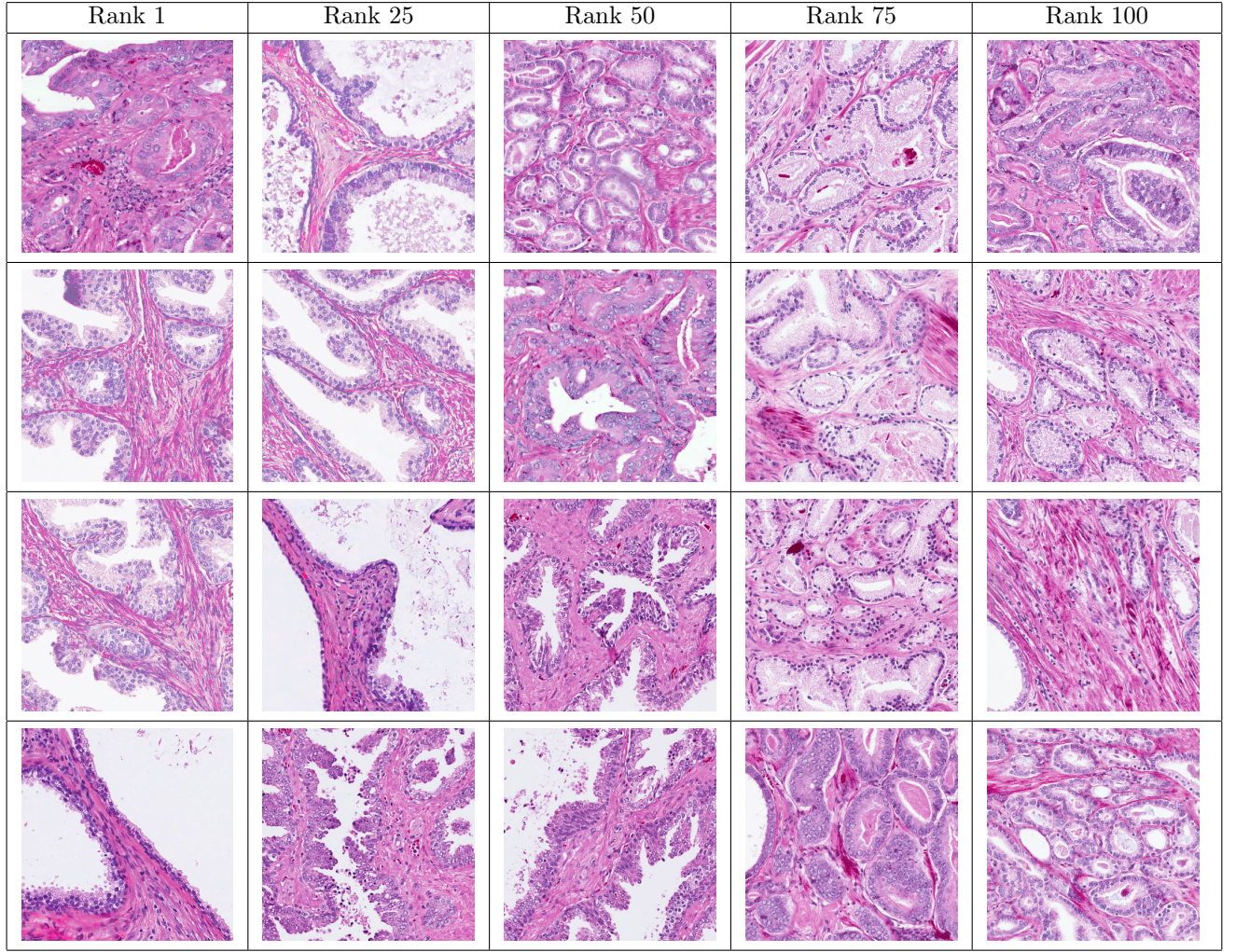


Figure 5. Some examples of various ranks, including ranks 1, 25, 50, 75 and 100.

- [8] H. Irshad, A. Veillard, L. Roux, and D. Racocceanu, "Methods for Nuclei Detection, Segmentation, and Classification in Digital Histopathology: A Review — Current Status and Future Potential," *IEEE Reviews on Biomedical Engineering*, vol. 7, pp. 97–114, 2014.
- [9] R. Lukac, *Perceptual Digital Imaging: Methods and Applications*, 1st ed. CRC Press, 2012.
- [10] M. Veta, P. J. V. Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. W. Pluim, "Automatic Nuclei Segmentation in H&E Stained Breast Cancer Histopathology Images," *PLoS ONE*, vol. 8, no. 7, p. e70221, 2013.
- [11] M. Macenko, M. Niethammer, J. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009, pp. 1107–1110.
- [12] A. K. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Trans Pattern Anal Mach Intell.*, vol. 19, no. 2, pp. 153–158, 1997.