Automatic Diagnosis of Prostate cancer using Random Forest Classifier

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper presents an automatic pathology (AutoPath) approach to prostate cancer detection based on the morphological features of the whole mount histopathology images of the prostate. To extract the features, the gland and nuclei regions of the images have been automatically segmented exploiting the color information and linear discriminant classifier. The extracted features include the size of the glands, epithelial layer density and nuclei density. We have proposed random forest classifier for the classification of malignant and benign regions in the histopathology images. Our algorithm has been tested on eight images and achieved average accuracy, specificity and sensitivity of 0.95 ± 0.03, 0.97 ± 0.02, and 0.65 ± 0.2, respectively with a leave-one-out cross validation. A comparative performance evaluation of the proposed technique with other benchmark classifiers such as Support Vector Machine and Linear Discriminant Analysis has also been presented in this paper. The experimental result corroborates that the Random Forest classifier is the most effective technique in classifying benign and malignant glands. The effectiveness of the proposed algorithm has also been demonstrated qualitatively in this paper.

1 Introduction

Prostate cancer is one of the most frequently diagnosed cancer and ranks second among the cancer related deaths of men worldwide [1]. Analysis of the histopathology specimens of prostate is an important step for prostate cancer diagnosis and treatment planning.

The tissue features of these histopathology images are the key indicators of prostate cancer. Among the different types of prostate cancer, the most common one is the prostatic adenocarcinoma, cancer pertaining to the gland units of the prostate. Pathologists determine the extent of this cancer by carefully evaluating the changes in the gland morphology. The gland is the main histopathological structural unit in prostate. Fig. 1 shows the structure of a normal gland unit. It mainly comprises a lumina of irregular shape, a layer of epithelial cells, and nuclei surrounding the lumina. The unit is supported by a surrounding fibro-muscular stroma. When the slides are stained using a Hematoxylin and Eosin (H&E) solution, the nuclei turn dark blue and the epithelial layer and stroma turn into different shades of purple to pink.

In the last few years there have been quite a number of AutoPath reports, that focus on the works are to computationally analyzing the pathology features and predicting the diagnostic decision based on these features. A method to distinguish the intermediate and high grade cancerous lesions of prostate tissues was presented in [2]. The decision was based on a number of features obtained from the shape and texture of the glands. The nuclear roundness factor analysis (NRF) was proposed in [3] to predict the behavior of low grade samples. Since this technique requires manual nuclear tracing, it is time consuming and tedious. Jafari-Khoujani *et. al.* [4] proposed a method for

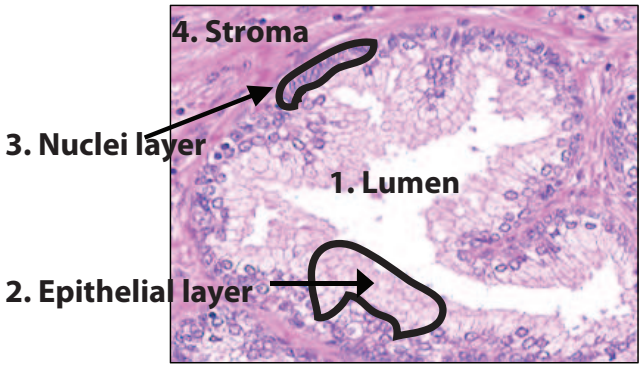

Figure 1: Graphical description of all the histopathology components associated with a complete gland unit: 1. Lumen, 2. Epithelial layer, 3. Nuclei, and 4. Stroma.

Table 1: **Literature review**

Authors	Dataset size	Classes	Accuracy
Doyle *et.al.* 2006 [4]	22 (40x)	cancer/non-cancer	88%
Tabesh *et.al.* 2007 [6]	268 (20x)	Low/High grade	81%
Naik *et.al.* 2008 [7]	44 (40x)	Benign, Grade-3, Grade-4, Grade-5	90%
Tai *et.al.* [5]2010	1000 (40x)	Benign, Grade-3, Grade-4, Grade-5	86%
Nguyen *et.al.*[8] 2011	82 ROI (10x)	Benign, Grade-3, Grade-4	85%

grading the pathological images of prostate biopsy samples by using energy and entropy features calculated from multiwavelet coefficients of an image. These multiwavelet features were used by k-nearest neighborhood classifier for classification and a leave-one-out procedure was applied to estimate the error rate. Again, there have been some works on prostate cancer grading using fractal dimension analysis [5]. In [5], the authors proposed fractal dimension (FD)-based texture features. These features were extracted by using a differential box counting method and an entropy-based fractal dimension estimation method. The feature were then combined them together as a FD-based feature set to analyze pathological images of prostate carcinoma. However this work focuses only on the separation of the different grades on manually detected cancerous regions. Tabesh *et. al.* [6] proposed an automatic two stage system for prostate cancer diagnosis and Gleason grading. The color, morphometric and texture features were extracted from the tissue images. Then, linear and quadratic Gaussian classifiers were used to classify images into cancer/noncancer classes and then further into low and high grade classes. Naik *et. al.* recently proposed an automatic gland segmentation algorithm recently [7]. A Bayesian classifier is used to detect candidate gland regions by utilizing low-level image features to find the lumen, epithelial cell cytoplasm, and epithelial nuclei of the tissue. Then, the features calculated from the boundaries of the gland that characterize the morphology of the lumen and gland region have been used to grade the cancer tissue. Another work based on gland segmentation has been proposed by Nguyen *et. al.* [8], which provides a competitive performance indices compared to other contemporary algorithms on the same topic but at a much lower magnification. These recent articles on biopsy specimen have been summarized in Table I. As can be observed from the table, among the recently published results Naik *et.al.*[7] gives the best accuracy.

By contrast, there have been much fewer reports of analysis of whole mount (WM) pathology images. Monaco *et.al.* [9] proposed an algorithm for detecting cancerous regions from whole mount slides using gland features. The information on gland proximity is modeled using a Markov Random field. The reported algorithm was applied to 40 images, among which 13 were from the same dataset that we analyze and report here. The authors report a sensitivity of 0.87 and a specificity of 0.90. Compared to these reported techniques, our proposed algorithm has been able to achieve a much higher accuracy of 0.95 ± 0.03.

The proposed algorithm performs automatic cancer classification on WM prostate slides based on gland features. The technique works in three steps: I) automatic segmentation of gland units, II) extraction of gland features, and III) detection of cancerous regions based on the features. The segmentation of gland units involve labeling of pixels in different histological objects using linear discriminant analysis. It will be discussed in detail in section II. In order to differentiate between cancerous and non-cancerous tissue the algorithm uses Random Foret classifier technique. This paper is organized as follows. Materials and methods of the complete cancer detection and grading algorithm are presented in Section II under three subsections: segmentation of gland units, feature extraction, and detection of cancerous region. In Section III, the AutoPath algorithm performance is evaluated on eight WM images. Finally, Section IV presents concluding remarks and suggestion for future work.

2 Materials and Methods

The whole mount histopathology sections were Hematoxylin and Eosin (H & E) stained and scanned into the computer at high resolution with a whole slide scanner. The original images are acquired at 20x magnification. In the proposed algorithm only the 5x magnification level has been used. Since the features distinguishing cancerous regions are quite clear at this lower resolution level, the analysis at the highest magnification is redundant here. For higher level analysis such as grading or staging, the highest magnification might be necessary. The lower resolution makes the image size much smaller and helps in achieving a faster implementation of the algorithm. At this resolution, the actual image scale is 8μm per pixel.

To extract the image features, the entire image is first divided into smaller subregions. A sample subregion is shown in Fig.2(a) The size of each sub-region is chosen to be $4mm \times 4mm$. In each sub-region, the gland units have been segmented and corresponding gland features has been extracted. Then based on the features, these sub-regions have been labeled as either cancerous or non-cancerous.

2.1 Segmentation of gland unit

The segmentation algorithm has been partially adopted from the work of Nguyen et.al. [8]. In the first step, labeling of pixels in each subregion has been performed. Each pixel has been labeled into one of these 5 categories, i.e., i) Gland lumen/lumina, ii) Epithelial layer, iii) Nuclei, iv) Stroma, and v) Annotation mark. We denote the class index by k where, $k \in \{1, 2, 3, 4, 5\}$ representing the 5 classes respectively. The first four classes are the histological objects that comprise a WM image. The fifth class is the cancer annotation that was performed on the WM slides before digitization. Some training patches of each class have been selected to train the classifier for pixel labeling The classification is based on the color information of these histological objects in *Lab* color space. In the RGB color space the lighting information and the color information is blended together. By converting to *Lab* color space the lighting information is confined into only one channel, *'L'*. The *Lab* space consists of a luminosity layer *'L'*, chromaticity-layer *'a'* indicating where color falls along the red-green axis, and chromaticity-layer *'b'* indicating where the color falls along the blue-yellow axis. Hence, each pixel to be labeled in the sub-region is now represented by three coordinates in *Lab* color space. Training pixels have also been converted to the *Lab* color space in similar way. The nth pixel in either the test data or the training data, is represented as $D_{n,j}$ where $n \in \{1, 2, ..., N\}$. N is the number of data points and $j = \{1, 2, 3\}$, for the three channel variables in the *Lab* color space.

The classification algorithm uses a linear discriminant analysis to label the testing pixels [10]. In the first step, for each class k the mean $\bar{D}_{k,j}$ is computed as,

$$\bar{D}_{k,j} = \frac{1}{N_k} \sum_{n \in \underline{N}_k} D_{n,j}; \tag{1}$$

where, $n \in \underline{N}_k$, N_k is the number of elements in the group k, and $\underline{N}_k$ denotes $\{1, 2, ..., N_k\}$.

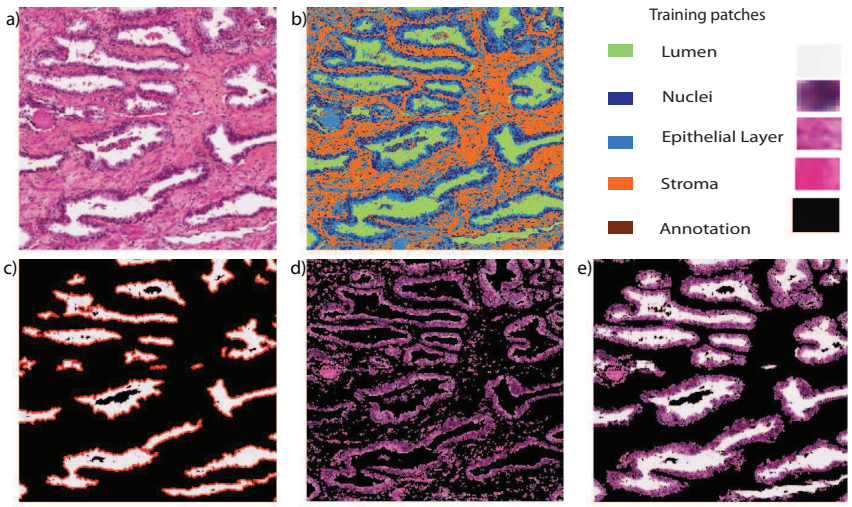

Figure 2: Gland segmentation. a) A sample subregion of WM slide, b) Labeled image of the subregion, c) Lumen objects, d) Epithelial layer-nuclei object, and e) segmented gland unit after consolidating surrounding epithelial layer-nuclei object with gland lumen.

Then the covariance matrix S for each class has been calculated. Here, S is considered to be equal for each class and estimated as single pooled estimate with entries

$$S_{i,j} = \frac{1}{N-K}\sum_{n=1}^{N}\left(x_{n;i} - \bar{x}_{k_n;i}\right)\left(x_{n;j} - \bar{x}_{k_n;j}\right), \tag{2}$$

where $\bar{x}_{k_n;i}$ means the i^{th} component of the mean vector for whichever class the data point n belongs to, k_n. N is the total number of data points and K is the total number of classes.

Then the squared Mahalanobis distance from a test data vector x to the mean of group of k is given by

$$z_k^2 = (x - \bar{x}_k)^{'} S^{-1}(x - \bar{x}_k). \tag{3}$$

Now the Bayes' formula for estimating posterior probability of data vector x to class k is,

$$P_k(x) = \frac{q_k|S_k|^{-0.5}exp[-0.5z_k^2]}{\sum_{l=1}^{K} q_l|S_l|^{-.5}exp[-0.5z_l^2]}. \tag{4}$$

As a result of single pooled estimate of covariance matrix, all the determinants of covariance estimate are equal, i.e., $|S_k|$ for all class $\{k|k \in 1, 2, ..., K\}$ is equal and hence the Bayes' formula reduces to a much simpler form,

$$P_k(x) = \frac{q_k exp[-0.5z_k^2]}{\sum_{l=1}^{K} q_l exp[-0.5z_l^2]}. \tag{5}$$

Then the data vector x is assigned to the class with which it has maximum posterior probability. Lets assume the data vector x corresponds to the n^{th} point in the subregion of interest, then the corresponding pixel label, k_n will be the $k_n = \arg\max_k(P_k(x))$.

Fig. 2(b) shows the labeled image generated after applying the pixel classification algorithm.

2.1.1 Consolidation of labeled pixels into gland unit

After having the labeled image, first we group together the lumen pixels using a connected-components algorithm which uses the eight-connectivity property. Around each lumen object, a lumen boundary is extracted. This is considered as the primary gland boundary (see Fig. 2(c)). As stated earlier in the introduction section, a complete gland unit consists of the lumina and its surrounding layer of epithelial cells and nuclei. Therefore, to segment out a complete gland unit

we consolidate the surrounding epithelial layer and nuclei with the lumina. Fig. 2(e) illustrates the resultant segmented gland units.

Several modifications to adopting the approach of Ngyuen et. al. [8] is necessitated because of the different nature of the data sets. The classification approach employed here is completely different from the reported algorithm [8]. The reported work used Voronoi tessellation based nearest neighborhood approach to classify each pixel. The main drawback of this approach is, when the number of training samples is large, the classification time for each testing data point is very high compared to linear discriminant analysis [10]. Therefore, when the number of testing samples are very large the reported nearest neighborhood approach will be very expensive in terms of computational time. In the work of Nyguen et.al. [8], they used their approach on biopsy specimens which are much smaller in size than the whole mount slides used in this project. Therefore, taking consideration of the huge size of images in this case, linear discriminant analysis has been adopted as classification approach instead of the nearest neighborhood approach.

2.2 Feature extraction

The main characteristic features of cancerous regions in the WM slides of prostate include high nuclear density, thick epithelial layers surrounding glands and smaller gland lumina. The proposed algorithm extracts these three features for each of the subregions and then classifies the subregions as either cancerous or noncancerous. The first feature is the nuclear density, ND, which is evaluated as ratio of the area of nuclei in the sub-region to the total area of sub-region. In the same way, the second feature the epithelial layer density, ED is evaluated. The third feature is the area of gland lumen, LA. It is computed as the average area of all the lumens in the sub-region.

2.3 Detection of cancerous region

Here we have employed random forest classifier for labeling each subregion as cancerous or non-cancerous . Each tree of random forest ensemble has been trained by bootstrapping two thirds of the features each time with replacement. In this experiment we have 100 trees in the ensemble. The factors affecting the parameter number of trees is the computational complexity and out-of-the-bag error. We plotted the out-of-the-bag error against total number of trees and observed that the error gets minimized as the number of grown trees get larger (Fig. 4). As can be observed from the figure, with the 100 trees we get as low as 0.05 out-of-the-bag error.

After the classification, the sub-regions that are labeled as cancerous are grouped together to form a continuous area. Any isolated detected subregion that are not in proximity of group of subregions have been discarded as false positives. The gland boundaries in the peripheral sub-regions of have been connected together to form the boundary around the group of subregions. The detected cancerous regions are then compared by overlaying the finer annotation by a second pathologist. Fig. 4(b) demonstrates strong agreement between the pathologist's annotation and experimental result.

3 Experimental result

The proposed algorithm has been evaluated on eight whole mount images. These whole mount histopathology images are digitized at $20\times$ magnification ($0.5\ \mu m$ per pixel) with an Aperio scanner. Fig 5 shows 4 example cases for qualitative evaluation. The black annotation mark has been done by the pathologist on the glass slide before digitization and does not provide a very good ground truth for performance evaluation of the proposed work. Therefore, a much finer annotation by a second pathologist on the digitized images has been obtained to provide a better ground truth. This is marked in blue. The green mark represents the detected cancerous region from the proposed algorithm. In all of the cases, both the detected region and the finer annotation shows strong agreement.

We quantitatively evaluate the performance of the proposed technique by doing leave-one-out cross validation among the eight images. Fig. 6 illustrates the graphical representation of the performance indices, sensitivity, specificity, and accuracy obtained by the proposed algorithm. We obtained average accuracy, specificity and sensitivity are 0.95 ± 0.03, 0.97 ± 0.02, and 0.65 ± 0.2, respectively. We have also tested the performance of the proposed technique with other benchmark classification

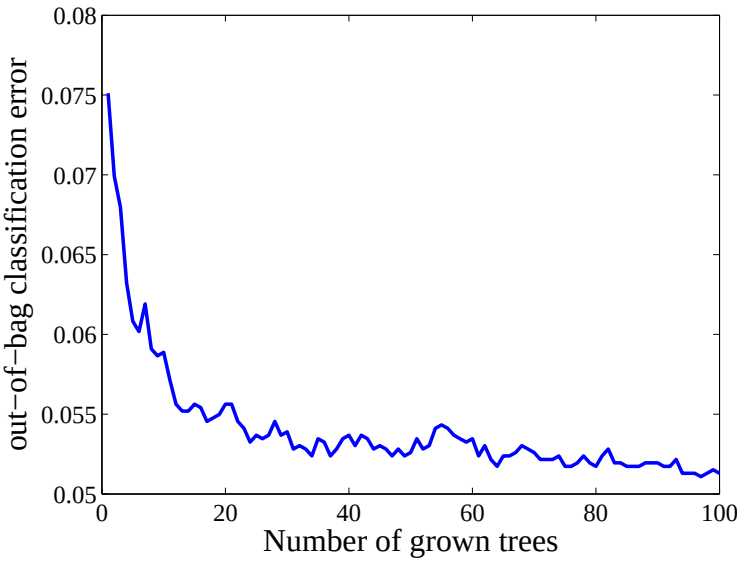

Figure 3: Out-of-bag error plot against nomber of grown trees, demonstrating that with an increase of number of trees the error gets minimized.

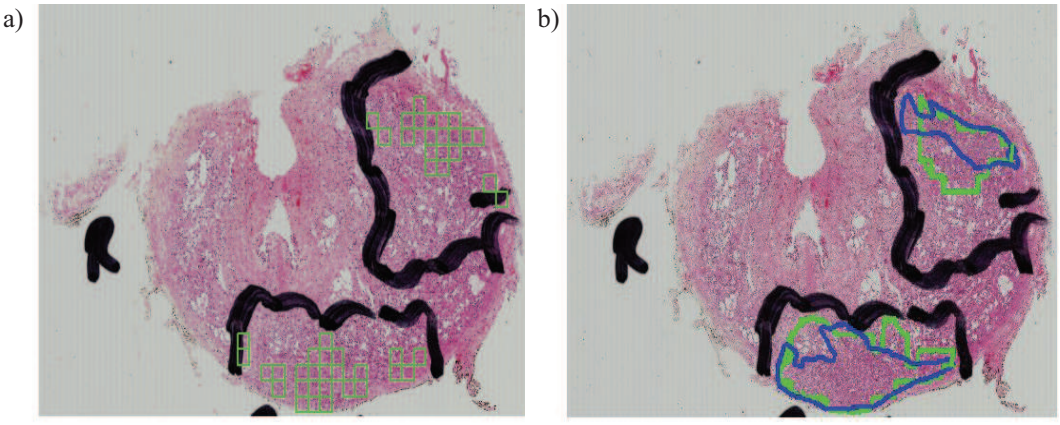

Figure 4: a) The green squares indicate the detected malignant sub-regions and b) consolidation of the subregions into a continuous region. he green annotation mark is the output of the proposed cancer detection algorithm. The blue mark is the finer annotation performed by a second pathologist.

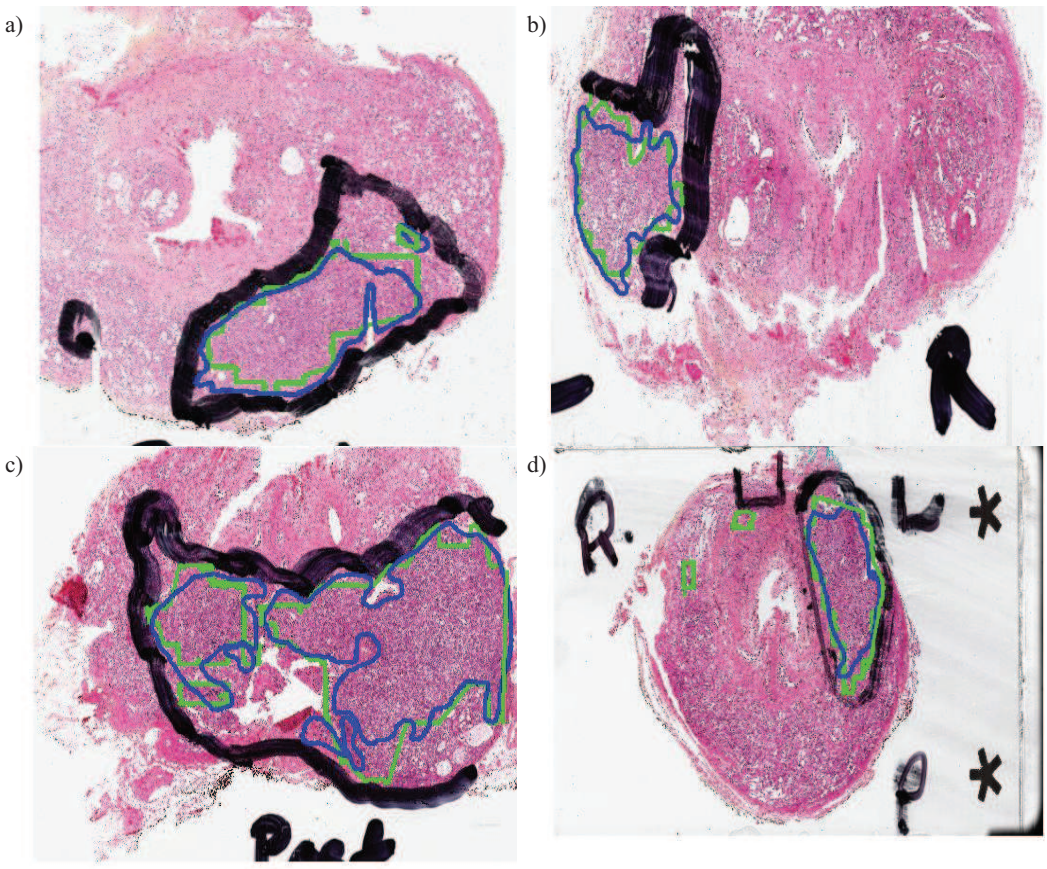

Figure 5: Performance of the proposed algorithm on 4 sample images. The green annotation mark is the output of the proposed cancer detection algorithm. The blue mark is the finer annotation performed by a second pathologist.

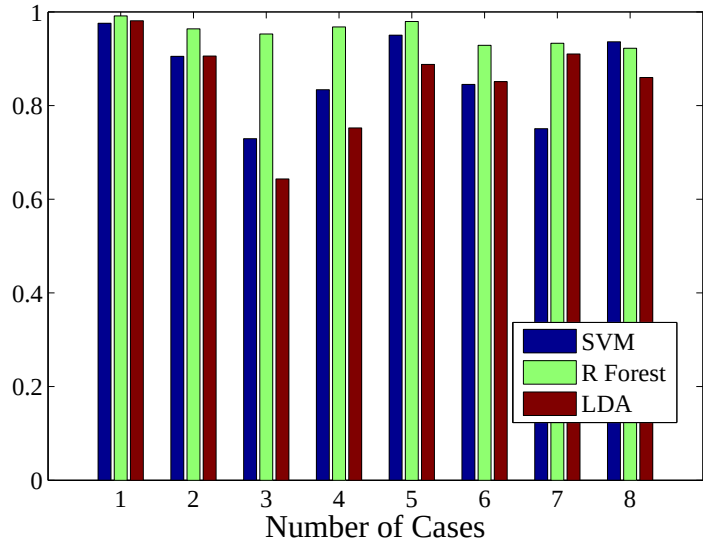

Figure 6: Comparison of accuracies of the proposed algorithm using Random Forest classifier with that of using other benchmark techniques such as Support Vector Machine and Linear Discriminant Analysis

techniques, i.e.., Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA). For both the cases the achieved accuracy is lower than the random forest classifier. Fig. 6 shows a comparative bar chart of the three classification technique.

Among the few works on cancer detection from whole mount histology of cancer, one of the most recent ones is by Monaco *et.al.* [9]. They perform the classification of benign and malignant regions based on the probabilistic pairwise markov model. They reported a sensitivity and specificity of 0.87 and 0.90, respectively on a dataset of 40 images among which 13 were of the same dataset used here. Compared to that work, our proposed technique achieves much higher sensitivity with cost of reduced specificity.

4 Conclusion

In this project, we have proposed a pathological diagnostic system AutoPath for automatic detection of cancerous region exploiting the morphological and architectural tissue features. We have used Random Forest as the automatic classifier and have shown that it performs better than the other benchmark techniques such as SVM and LDA. As part of the system, automatic gland segmentation have been performed. Apart from having application in cancer detection, the gland segmentation may have application in other fields also, as for example the segmented glands might be used as a landmark for registering between different slides of same patients. Depending on very few number of features compared to other reported techniques, the proposed system has demonstrated very high level of specificity and sensitivity which corroborates he effectiveness and robustness of the proposed algorithm.

References

[1] C. Bohring and T. Squires, "Cancer statistics," *CA Cancer J. Clin.*, vol. 43, pp. 7–26, 1993.

[2] R. Stotzka, R. Manner, P. H. Bartels, and D. Thompson, "A hybrid neural and statistical classifier system for histopathologic grading of prostate lesions," *Analytical and Quantitative Cytology and Histology*, vol. 17, no. 3, pp. 204–218, 1995.

[3] M. D. Clark, F. B. Askin, and C. R. Bagnell, "Nuclear roundness factor: a quantitative approach to grading in prostate carcinoma, reliability of needle biopsy tissue, and the effect of tumor stage fore usefulness," *The prostate*, vol. 10, pp. 199–206, 1987.

[4] J. K. Khouzani and S. H. Zadeh, "Multiwavelet grading of prostate pathological images," in *Proc. SPIE*, vol. 4628, 2002, pp. 1130–1138.

[5] P. W. Huang and C. H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Transactions on Medical Imaging*, vol. 28, no. 7, pp. 1037–1050, 2009.

[6] A. Tabesh, M. Teverovskiy, H. Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and gleason grading of histological images," *IEEE Transactions on Medical Imaging*, vol. 26, no. 4, pp. 518–523, 2007.

[7] S. Naik, S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information," in *In Proc. of 2nd Workshop on Micro. Image Anal. with Applications in Biology*, 2007.

[8] K. Nguyen, B. Sabata, and A. K. Jain, "Prostate cancer grading: Gland segmentation and structural features," vol. 33, pp. 951 – 961, 2011.

[9] J. P. Monaco, M. Feldman, J. Tomaszewski, and A. Madabhushi, "Detection of prostate cancer from whole-mount histology images using markov random fields," in *In Proc. of 2nd Workshop on Micro. Image Anal. with Applications in Biology*, 2008.

[10] W. Krzanowski and W. Krzanowski, *Principles of multivariate analysis*. Oxford University Press, 1996.