Automatic Pathology of Prostate Cancer in Whole Mount Slides incorporating Individual Gland Classification

by

Sabrina Rashid

BSc in Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Applied Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Electrical and Computer Engineering)

The University Of British Columbia
(Vancouver)

March 2014

© Sabrina Rashid, 2014

Abstract

This thesis presents an automatic pathology (AutoPath) approach to detect prostatic adenocarcinoma based on the morphological analysis of high resolution whole mount histopathology images of the prostate. We are proposing a novel technique of labeling individual glands as benign or malignant exploiting only gland specific features. Two new features, the Number of Nuclei Layers and the Epithelial Layer density are proposed here to label individual glands. To extract the features, individual gland and nuclei units are segmented automatically. The nuclei units are segmented by employing a marker-controlled watershed algorithm. The gland units are segmented by consolidating their lumina with the surrounding layers of epithelium and nuclei. The main advantage of this approach is that it can detect individual malignant gland units, irrespective of neighboring histology and/or the spatial extent of the cancer. Therefore, a more sensitive annotation of cancer can be achieved by the proposed AutoPath technique, in comparison to the current clinical protocol, where the cancer annotation is performed at the regional macro level instead of glandular level technique.

We have also combined the proposed gland-based approach with a regional approach to perform automatic cancer annotation of the whole mount images. The proposed algorithm performs the task of cancer detection in two stages: at first with pre-screening of the whole mount images in a low resolution ($5\times$), and then ii) a finer annotation of the cancerous regions by labeling individual glands at a higher magnification ($20\times$). In the first stage, the probable cancerous regions are classified using a random forest classifier that exploits the regional features of the tissue. In the second stage, gland specific features are used to label individual gland units as benign or malignant. The strong agreement between the experimental re-

sults and the pathologist's annotation corroborates the effectiveness of the proposed technique. The algorithm has been tested on 70 images. In a 10-fold cross validation experiment it achieved average sensitivity of 88%, specificity of 94% and accuracy of 93%. This surpasses the accuracy of other methods reported to date.

Preface

Material from Chapters 2 and 3 was published in the conference for Medical Image Computing and Computer Assisted Intervention (MICCAI) under the title "Separation of benign and Malignant Glands in Prostatic Adenocarcinoma". The work was co-authored by Ladan Fazli, Alexander Boag, Robert Siemnes, Purang Abolmaesumi, and Septimiu E. Salcudean[1].

The histopathology data used in this thesis was collected at Queen's University by Alexander Boag and Robert Siemens. The study was cleared by the Medical Research Ethics Board at Queen's University, research study number UROL-146-05.

[1] S. Rashid, L. Fazli, R. Siemens, A. Boag, P. Abolmaesumi, S. E. Salcudean, "Separation of benign and malignant glands in prostatic adenocarcinoma, Medical Image Computing and Computer Assisted Intervention (MICCAI), 2013. Springer Berlin Heidelberg, 2013, 461-468.

Table of Contents

Abstract . . . . . **ii**

Preface . . . . . **iv**

Table of Contents . . . . . **v**

List of Tables . . . . . **vii**

List of Figures . . . . . **viii**

List of Abbreviations . . . . . **xv**

Acknowledgments . . . . . **xvi**

Dedication . . . . . **xvii**

1 Introduction **1**

1.1 Prostate Anatomy and Pathology 3

1.2 Literature Review 5

1.3 Summary of the Proposed Technique 11

1.4 Thesis Organization 13

2 Cancer Classification Using Regional Features **14**

2.1 Introduction 14

2.2 Gland Segmentation 15

2.3 Feature Extraction 18

2.3.1 Nuclei Features 20

2.3.2 Lumen Features 20
2.3.3 Epithelial Features 20
2.4 Classification of Malignant Regions 21
2.5 Conclusion 22

3 Cancer Classification Using Glandular Features 24
3.1 Introduction 24
3.2 Segmentation of the Nuclei Unit 25
3.3 Extraction of Gland-specific Features 27
3.3.1 Number of Nuclei Layers (N_{NL}) 27
3.3.2 Ratio of Epithelial Layer area to Lumen Area (R_{EL}) 28
3.4 Consolidation of Glands 28
3.5 Conclusion 29

4 Experimental Results 32
4.1 Dataset 32
4.2 Parameter Tuning for Random Forest Classifier 32
4.3 Parameter Tuning for Individual Gland Classification 34
4.4 Qualitative Performance Evaluation 34
4.5 Discussion 35

5 Conclusions 73
5.1 Summary of Contributions 73
5.2 Future Work 75

References 76

List of Tables

Table 1.1 Literature review 10

Table 2.1 Features extracted from each image block 19

Table 4.1 AUC obtained by our algorithm for different parameter values of Rdf_{leaf}. 34

List of Figures

Figure 1.1 A typical whole mount histopathology slide of prostate. 2
Figure 1.2 Four pathology zones of prostate. 3
Figure 1.3 A normal gland structure. 5
Figure 1.4 Examples of the five grades of the Gleason grading system. a) Classic Gleason grading diagram drawn by Dr. Gleason [14]. (b-f) Evolution of Gleason grades from 1 to 5, respectively . . 6
Figure 1.5 Visual comparison between a) benign and b) cancerous prostate glands. 9
Figure 1.6 Flow-chart of the proposed algorithm. 12

Figure 2.1 Gland segmentation. a) A sample image block, b) labeled image where each histological component is represented by a different color, c) enlarged view of a small window in the labeled image, d) lumen objects (the red mark corresponds the initial gland boundary), and e) segmented gland unit after consolidating surrounding epithelial layer-nuclei object with the gland lumen. 15
Figure 2.2 Plot of Out-of-the-bag classification error against the number of grown trees. 22
Figure 2.3 Performance of the proposed algorithm on a test image. a) Cyan annotation is performed by pathologist and considered as the ground truth. b) Result of the screening phase of the proposed technique. 23

Figure 3.1 a) Input scene: inverted 'R' channel of the histopathology image, b) preprocessed image after background subtraction and thresholding, c) computed foreground and background markers overlaid on the preprocessed image, d) modified gradient image with regional minima placed at the foreground and background object markers, e) labeled image of the segmented nuclei after watershed transform, and f) segmented nuclei (marked by green boundary). 26

Figure 3.2 a) Graphical illustration of Y_{ang} calculation. b) Sample histopathology scene with a single benign (marked with blue ellipse) and two malignant gland units (marked with red ellipses). c), d), and e) illustrate the different appearances of the histograms (Y_{ang}) of benign and malignant glands. 27

Figure 3.3 a) Graphical illustration of boundary hull generation by α-shape approach. 30

Figure 3.4 a) Experimental result of the proposed technique after incorporating the glandular level classification with the regional cancer classification. 31

Figure 4.1 ROC analysis for the parameter tuning of the random forest classifier. Here the ROC curve with $Rdf_{leaf} = 12$ is shown. At the optimum operating point the sensitivity $S_n = 0.94$ and specificity $S_p = 0.83$. 33

Figure 4.2 ROC analysis of the individually labeled gland dataset. Here, the ROC curve for $R_{EL} = 0.93$ is shown. At the optimum operating point the sensitivity $S_n = 0.79$ and specificity $S_p = 0.85$. 35

Figure 4.3 Performance of the proposed technique on a sample test image (Case: 1). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5× resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units. 36
Figure 4.4 Performance of the proposed technique on a sample test image (Case: 2). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5× resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units. 37
Figure 4.5 Performance of the proposed technique on a sample test image (Case: 3). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5× resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units. 38
Figure 4.6 Performance of the proposed technique on a sample test image (Case: 4). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5× resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units. 39

Figure 4.7 Performance of the proposed technique on a sample test image (Case: 5). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5x resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units. 40
Figure 4.8 Performance of the proposed technique on a sample test image (Case: 6). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5x resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units. 41
Figure 4.9 Performance of the proposed technique on a sample test image (Case: 7). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5x resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units. 42
Figure 4.10 Performance of the proposed technique on a sample test image (Case: 8). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5x resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units. 43

Figure 4.11 Performance of the proposed technique on a sample test image (Case: 9). a) The test image with both pathologists' annotation overlayed. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5x resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units. 44
Figure 4.12 Illustration of the individual gland labeling on whole mount images. Enlarged view of four different types of detected glands, 1). Atrophy, 2) PIN, 3) Malignant glands, and 4) Seminal vesicle. 45
Figure 4.13 Illustration of the individual gland labeling on a sample whole mount image (Test image:1) 46
Figure 4.14 Illustration of the individual gland labeling on a sample whole mount image (Test image:2) 47
Figure 4.15 Illustration of the individual gland labeling on a sample whole mount image (Test image:3) 48
Figure 4.16 Illustration of the individual gland labeling on a sample whole mount image (Test image:4) 49
Figure 4.17 Illustration of the individual gland labeling on a sample whole mount image (Test image:5) 50
Figure 4.18 Illustration of the individual gland labeling on a sample whole mount image (Test image:6) 51
Figure 4.19 Illustration of the individual gland labeling on a sample whole mount image (Test image:7) 52
Figure 4.20 Illustration of the individual gland labeling on a sample whole mount image (Test image:8) 53
Figure 4.21 Illustration of the individual gland labeling on a sample whole mount image (Test image:9) 54
Figure 4.22 Illustration of the individual gland labeling on a sample whole mount image (Test image:10) 55

Figure 4.23 Illustration of the individual gland labeling on a sample whole mount image (Test image:11) 56
Figure 4.24 Illustration of the individual gland labeling on a sample whole mount image (Test image:12) 57
Figure 4.25 Illustration of the individual gland labeling on a sample whole mount image (Test image:13) 58
Figure 4.26 Illustration of the individual gland labeling on a sample whole mount image (Test image:14) 59
Figure 4.27 Illustration of the individual gland labeling on a sample whole mount image (Test image:15) 60
Figure 4.28 Illustration of the individual gland labeling on a sample whole mount image (Test image:16) 61
Figure 4.29 Illustration of the individual gland labeling on a sample whole mount image (Test image:17) 62
Figure 4.30 Illustration of the individual gland labeling on a sample whole mount image (Test image:18) 63
Figure 4.31 Illustration of the individual gland labeling on a sample whole mount image (Test image:19) 64
Figure 4.32 Illustration of the individual gland labeling on a sample whole mount image (Test image:20) 65
Figure 4.33 Illustration of the individual gland labeling on a sample whole mount image (Test image:21) 66
Figure 4.34 Illustration of the individual gland labeling on a sample whole mount image (Test image:22) 67
Figure 4.35 Illustration of the individual gland labeling on a sample whole mount image (Test image:23) 68
Figure 4.36 Illustration of the individual gland labeling on a sample whole mount image (Test image:24) 69
Figure 4.37 Illustration of the individual gland labeling on a sample whole mount image (Test image:25) 70
Figure 4.38 Illustration of the individual gland labeling on a sample whole mount image (Test image:26) 71

Figure 4.39 Illustration of the individual gland labeling on a sample whole mount image (Test image:27) 72

List of Abbreviations

WM	Whole Mount
PSA	Prostate-Specific Antigen
DRE	Digital Rectal Examination
PIN	Prostatic Intraepithelial Neoplasia
ROC	Receiver Operator Characteristics
AUC	Area Under the Curve
ROI	Region of Interest
LDA	Linear Discriminant Analysis

Acknowledgments

Firstly, I would like to express my sincere gratitude towards my supervisors Drs. Tim Salcudean and Purang Abolmaesumi for their continuous support and advice. Their guidance and enthusiasm kept me motivated through out the research project.

I would like to thank Dr. Ladan Fazli for her time and support in providing the ground truth for the project and training me about the clinical aspect of prostate pathology. I would also like to thank Drs. Alexander Boag and Robert Siemens for providing the data set of the whole mount images of the prostate.

Dr. Mehdi Moradi was extremely helpful in providing feedback on the project's machine learning aspects. I would also like to thank my colleagues at the Robotics and Control Lab for creating a very friendly and helpful environment during my stay as a masters candidate.

Lastly, I would like to thank my amazing parents back in my country Bangladesh, my father Abdur Rashid Dewan, mother Lutfa Sultana. It is difficult being so far away but their strong belief in me have always kept me motivated to achieve my goal. I would also like to take the opportunity to thank my brother Toufiqul Islam and sister in law Rubaiya Rahman for their continuous advice and support.

Dedication

I would like to dedicate this thesis to my father Abdur Rashid Dewan and mother Lutfa Sultana for always believing in me and giving me the freedom to pursue my dream.

Chapter 1

Introduction

Prostate cancer is currently the second most prevalent type of cancer for men and ranks third among the cancer related deaths of men worldwide [17]. Prostate cancer is usually suspected when a high level of prostate specific antigen (PSA) is detected in blood tests. A digital rectal examination (DRE), in which the physician palpates the prostate through the rectum is then performed to detect any abnormalities within the prostate. Anomalies in these tests lead clinicians to conduct a prostate biopsy. A prostate gland biopsy is a diagnostic procedure which involves removal and examination of small samples of tissue. Examination of the microscopic biopsy specimens by pathologists is an important step for confirming the diagnosis of malignancy and guiding the treatment [40]. In case of advanced cancers, surgeons often perform Radical Prostatectomy (RP) on patients, i.e., surgical removal of the entire prostate. The prostate specimen removed during the RP procedure is processed and analyzed by a pathologist in order to determine further treatment, such as radiation or hormone treatment, depending on the extend and location of the cancer found in the specimen. In some cases, the prostate specimen is sliced in parallel transversal slices that cross the entire organ at intervals of 4-8 mm, in order to facilitate the comparison of prostate images acquired pre-operatively tissue histopathology. The histopathology slices obtained from the cross section of these *ex vivo* prostate are termed as Whole Mount (WM) slides. Fig. 1.1 shows a typical WM slide of the prostate. The black marks present in the image are the coarse annotation by the pathologist on the glass slide before digitization.

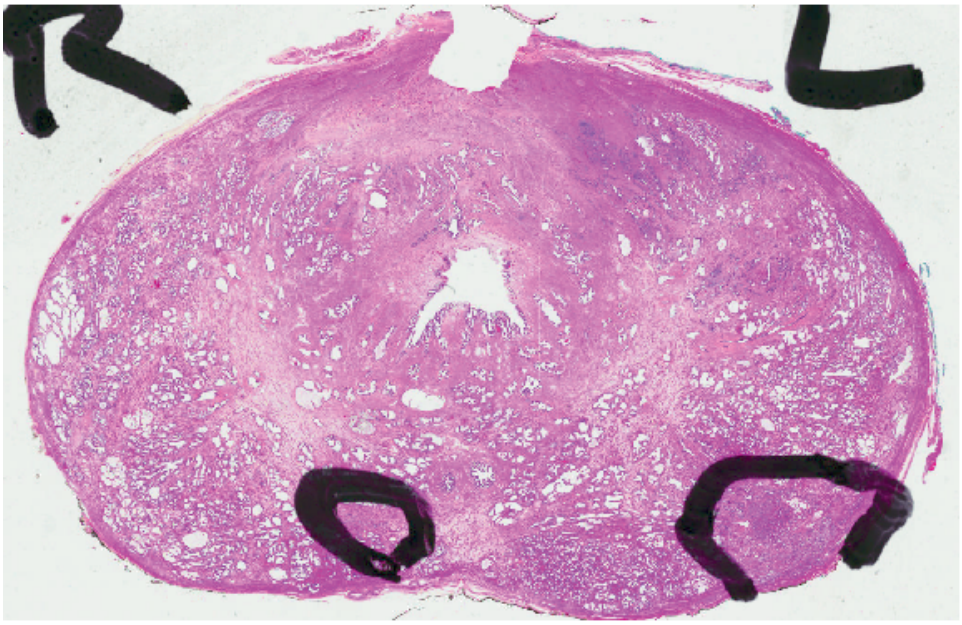


Figure 1.1: A typical whole mount histopathology slide of prostate.

The analysis of these WM slides has been an area of special interest in recent years. The analysis of the WM images help to predict the long term disease outcome of the RP patients i.e., the possibility of recurrence of cancer in adjacent organs. The whole mount pathology analysis after RP can also be used as ground truth to determine the ability of imaging, such as multi-parametric magnetic resonance imaging, to detect cancer.

In clinical practice, the analysis of the WM slides is performed by pathologists manually. Since the level of structural detail in these images is very high, the process of annotating and grading the entire image is very time consuming and also subjective to human pathologists' expertise. Therefore, for future work in prostate cancer prognosis and image-based diagnosis, it is important that an automatic pathology approach that is consistent and accurate in classifying cancer be developed. This is the objective of the AutoPath research work described in this thesis.

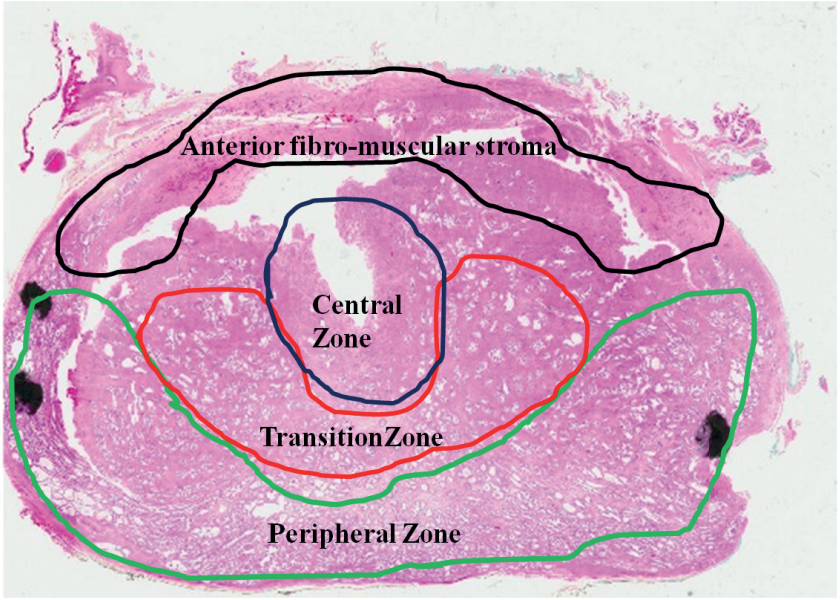


Figure 1.2: Four pathology zones of prostate.

1.1 Prostate Anatomy and Pathology

The prostate is a gland in the male reproductive system. Its function is to store and secrete a slightly alkaline fluid which usually constitutes $20-30\%$ of the volume of the semen. In healthy adult males its size is slightly larger than a walnut. The weight of a healthy prostate in adult males ranges from 7 to 16 gms with an average weight about 11 grams [21], [23]. The prostate sits above the base of the penis and below the urinary bladder and backs onto the front wall of the rectum. The apex of the prostate is pointed down to the perineum as opposed to the base which is wider and located next to the bladder. The prostatic urethra is the portion of the urethra that runs from the urinary bladder through the prostate and exits from the apex via the urinary sphincter which is a group of muscles that prevents involuntary leakage of urine. The prostate is surrounded by a membrane called the prostatic capsule. In pathology, the regions of the prostate are classified as zones Fig. 1.2. The prostate gland has four distinct glandular regions:

- Peripheral zone (PZ): This zone occupies approximately 70% of the volume of gland. $70-80\%$ of prostatic cancers originate in the peripheral zone [8].

- Central zone (CZ): This zone constitutes approximately 25% of the prostate gland. About 5% of prostate cancer cases originate in the central zone [8].

- Transition zone (TZ): The transition zone is the innermost part of the prostate gland and surrounds the urethra. It makes up about 5% of the prostate volume. About 10% of prostate cancers occur in this zone. This zone also enlarges with age and can result in benign prostatic enlargement [8].

- Anterior fibro-muscular zone (or stroma): The anterior zone is located close to the abdomen (away from the rectum). This zone constitutes 5% of the gland volume and is composed mostly of muscular tissue [8].

The main histopathological structural unit in the prostate is called a gland. Fig. 1.3 shows the structure of a normal gland unit. It mainly comprises a lumen of irregular shape, a layer of epithelial cells, and nuclei surrounding the lumina. The unit is supported by a surrounding fibro-muscular stroma. Each of these components correspond to a different color when the slides are stained using a Hematoxylin and Eosin (H&E) solution. In response to the solution, the nuclei turn into dark blue objects and the epithelial layer and stroma turn into different shades of purple to pink. The morphological and architectural features of the glands indicate whether the gland is benign or malignant. Fig. 1.5 illustrates the different appearances of cancerous and benign glands. Cancerous glands tend to have a single layer of nuclei with a higher ratio of epithelial layer area to lumen area.

By examining the glandular tissue features in the microscopic histopathology sections, the pathologist determines the histological grades. The most widespread technique for histological grading is the Gleason grading system [14].

This grading scheme was developed by a pathologist, Dr. Gleason during the 1970s. In this grading system, the prostate cancer can be classified into 5 grades representing a number ranging from 1 to 5, where 1 is the most benign and 5 is the most malignant case. A classic Gleason grading diagram containing the five basic tissue patterns associated with five cancer grades is shown in Fig. 1.4a. Gleason grading is based upon the distribution of nuclei and morphology of gland structures in the image. Fig 1.4b-f shows the evolution of glandular and nuclear regions in the different grades of prostate cancers in real pathological images. As can be

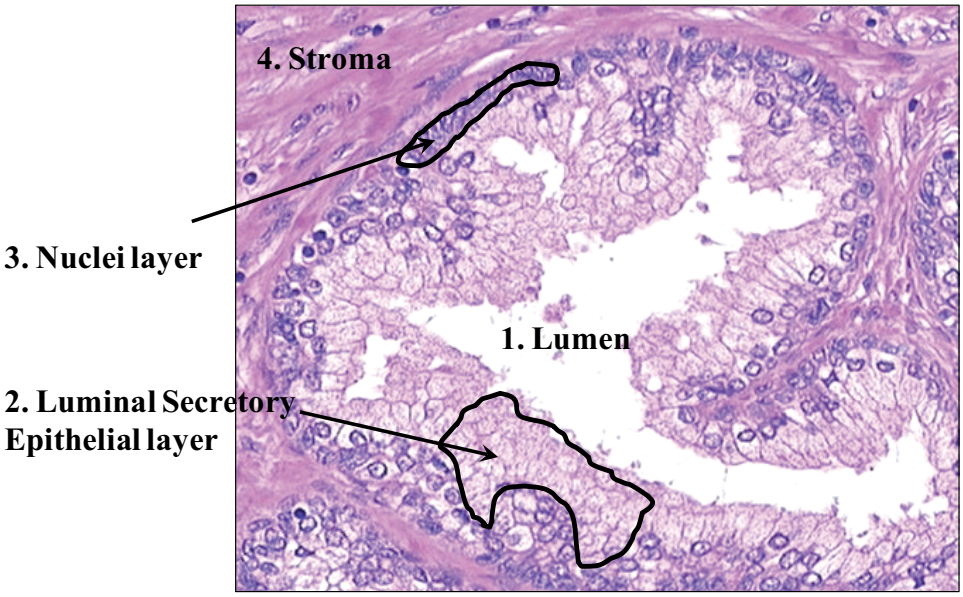


Figure 1.3: A normal gland structure.

observed from the figures, in the lower grades of cancer (Grade 1 and 2) the glands still maintain an irregular shape like normal glands but they get smaller in size and the concentration of nuclei increases slightly. In higher grades (3 and 4) the glands are smaller, and instead of having an irregular shape they tends to have a more regular circular or elliptical shape. The distance between the glands increases. In Grade 5 cancer, no distinct glands can be observed. Instead, there is a random concentration of nuclei floating randomly in the stroma. The aggressiveness of prostate cancer are scored by combining the top two grades present in a particular region. For example, if a tissue region has 50% Grade 3 cancer, 30% Grade 4 cancer, and 20% other Grade cancers then the corresponding Gleason score of the tissue region will be 3+4=7.

1.2 Literature Review

The analysis of pathological images has been an area of interest during the last few years. Normally referred as Digital Pathology, the aim of this field has been to distinguish between the normal and abnormal tissue. Research in this area can be

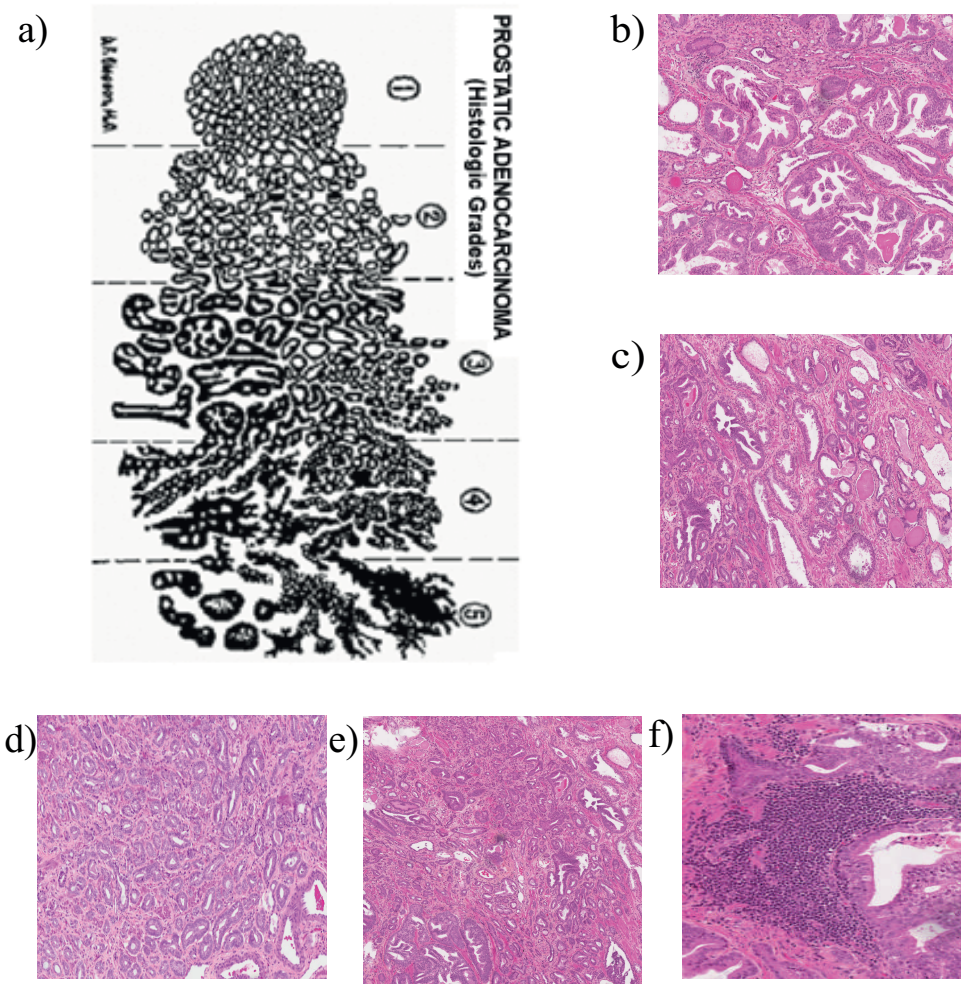


Figure 1.4: Examples of the five grades of the Gleason grading system. a) Classic Gleason grading diagram drawn by Dr. Gleason [14]. (b-f) Evolution of Gleason grades from 1 to 5, respectively

broadly divided into two categories: work on biopsy specimens and work on whole mount pathology images.

Most of the research in this field has been carried out on the biopsy specimen samples. A method to distinguish the moderately and poorly differentiated lesions of prostate tissues was presented by Stotzka *et al.* in [34]. The decision was based on a number of features obtained from the shape and texture of the glands. The nuclear roundness factor analysis (NRF) is proposed in [9] 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.* [19] proposed a method for grading the pathological images of prostate biopsy samples by using energy and entropy features calculated from multi-wavelet coefficients of an image. These multi-wavelet features were used by a k-nearest neighborhood classifier for classification and a leave one out procedure [19] was applied to estimate the error rate. In other research, prostate cancer grading was carried out using fractal dimension analysis [16]. In [16], the authors proposed fractal dimension based texture features that were extracted through a differential box counting method. These features were combined with an entropy-based fractal dimension estimation method as a fractal-dimension based feature set to analyze pathological images of prostate carcinoma. This research focuses only on the separation of the different grades on manually detected cancerous region. Tabesh *et al.* [36] proposed a two stage system for prostate cancer diagnosis and Gleason grading. The color, morphometric and texture features are 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.

Since the grading of the cancer depends on gland morphology and nuclei distribution, proper segmentation of these has been an active area of research interest for last couple of years. A good part of the research activity addressed the segmentation of nuclei, as they are clearly visible on histology. Bamford and Lovell [3] used an active contour scheme for segmenting nuclei in pap-stained cervical cell images. A fuzzy logic engine was proposed by for segmentation of prostate tissue that uses both color and shape based constraints [5]. But these studies focus only on finding the nuclei units only. Segmentation of multiple structures on prostate histology has been carried out by Gao *et al.* using a color-based histogram thresholding tech-

nique to enhance regions of cytoplasm and nuclei to aid in manual cancer diagnosis [13]. Recently, Naik *et al.* proposed an automatic gland segmentation algorithm [28]. 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. The most recent articles on cancer classification in biopsy specimen have been summarized in Table I. As can be observed from the table, among the recent literature, Naik *et al.* [28] gives the best accuracy.

Compared to the cancer classification works on biopsy specimens, there have been fewer reports on automatically annotating whole mount images. Gorelick *et al.* proposed an automatic cancer classification method for sub-images extracted from whole mount images [15] using the superpixel [12], [37] partitioning and Ada-Boost classification [33]. The authors did not report the performance of annotating complete whole mount images. Monaco *et al.* [26] proposed an algorithm for annotating cancerous regions in whole mount slides using gland features. The reported technique segmented gland lumens and classified glands into normal or cancer by (i) using gland size feature to assign initial gland labels and (ii) applying a probabilistic pairwise Markov model (PPMM) to update gland labels. See Table I for a summary of the results reported for biopsy specimen and WM images.

Most of these previous works do not evaluate the features that are determined for each individual gland unit. Since prostatic adenocarcinoma is the cancer pertaining to the gland unit and the pathological changes in malignant tissue occur at the gland level, a clinically more relevant approach would be to incorporate the gland specific features in the computational cancer detection process. In a recent article on gland classification, Nguyen *et al.* [31] achieved an accuracy of 0.79 in classification of benign and malignant glands by exploiting region-specific/contextual features such as percentage of nuclei pixels, lumen shape similarity, lumen size similarity, and neighborhood. Since the proposed technique utilizes contextual information for classifying individual glands, therefore in cases where benign gland appear in close proximity of malignant glands, the technique will most likely fail. In addition, the proposed technique does not perform nuclei segmentation but utilizes percentage of nuclei pixels as classification. Therefore,

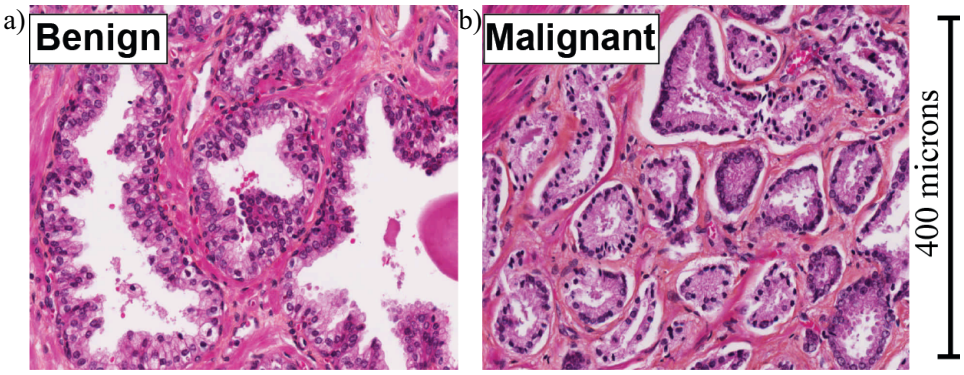


Figure 1.5: Visual comparison between a) benign and b) cancerous prostate glands.

multiple smaller nuclei (usually in benign) and single large nuclei (usually in malignant) will result in similar feature index which will not be true representative of the gland condition. In comparison to that, our proposed features are strictly gland specific and involve i) pixel labeling, ii) segmentation of each nuclei in the gland, and iii) finding the number of layers of nuclei for each gland from angle-dependent histograms. The advantage of this technique is that it can detect a malignant/suspicious gland irrespective of the region properties. In cases where malignant glands are present in close proximity to benign glands, this approach might provide a more sensitive cancer annotation compared to approaches that use region-dependent image features [31].

Table 1.1: Literature review

Authors	Dataset size	Classes	Accuracy
Doyle et al. 2006 [19]	22 (biopsy)	cancer/non-cancer	88%
Tabesh et al. 2007 [36]	268 (biopsy)	Low/High grade	81%
Naik et al. 2008 [28]	44 (biopsy)	Benign, Grade-3, Grade-4, Grade-5	90%
Tai et al. 2010 [16]	1000 (biopsy)	Benign, Grade-3, Grade-4, Grade-5	86%
Monaco et al. 2012 [26]	40 (37 quarter sections, 13 WM)	Benign/malignant	90%
Gorelick et al. 2013 [15]	991 sub-images from 50 WM sections	Cancer/non-cancer	90%
Nguyen et al. 2012 [31]	48 images	Gland labeling	79%

1.3 Summary of the Proposed Technique

The AutoPath technique proposed in this thesis performs automatic cancer classification on WM prostate in two stages, i) screening of probable cancerous regions at low magnification (5×) and ii) finer annotation of the detected cancerous regions at high magnification (20×). To extract the tissue features in the first stage, we automatically segment the individual gland units and its associated tissue components. The whole image is divided into small blocks and features extracted from each of these blocks are fed into a random forest classifier [7] to detect benign and malignant regions of the image. Further analysis of the detected regions is performed in the second stage at a higher magnification. At this step, the malignant regions are detected based on their gland-specific properties. We propose two new features for classifying glands, i) Number of Nuclei Layers (N_{NL}) and ii) Ratio of Epithelial Layer area to Lumen area (R_{EL}). To extract the first feature, nuclei units have been automatically segmented from the image using a marker-controlled watershed algorithm[25]. The introduction of these two gland-specific features allow us to detect malignant or suspicious glands without relying on surrounding histology. Therefore, a more specific and sensitive annotation of the images is possible. We have tested our technique on 70 images obtained from 30 patients. In a 10-fold cross validation we have achieved an average sensitivity of 84%, specificity of 94%, and accuracy of 93%. A flow-chart of the proposed technique is illustrated in Fig. 1.6 and the detailed explanation of each step is provided in the methodology section.

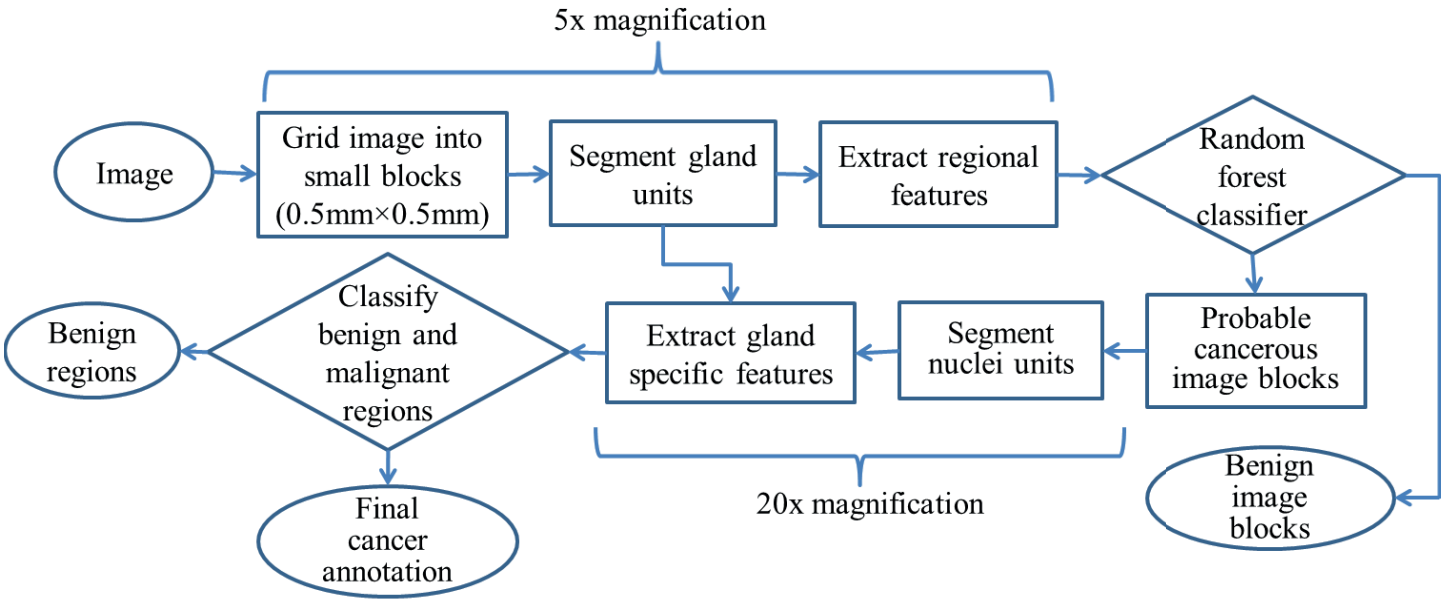


Figure 1.6: Flow-chart of the proposed algorithm.

1.4 Thesis Organization

We start with Chapter 2 explaining the screening phase of our prostate cancer detection technique exploiting regional features. This chapter discusses the gland segmentation algorithm, regional feature extraction, and the application of random forest classifier in classifying benign and malignant regions. Chapter 3 describes the glandular level classification of the proposed technique providing the details of nuclei segmentation and glandular feature extraction. We present the performance evaluation of our algorithm in Chapter 4. Finally, conclusions, achievements and future directions of our work are laid out in Chapter 5.

Chapter 2

Cancer Classification Using Regional Features

2.1 Introduction

In prostatic adenocarcinoma, a tumor is defined to be comprised of a group of malignant gland units. The arrangement and architecture of the glands in the tumor deviate from the healthy tissue type depending the cancer grade. The higher the cancer grade, the more deviation of the tissue architecture is observed. At this stage we have quantified these changes in tissue architecture and based on that we have screened the pathology slices for the possible cancerous regions. For the screening phase, only the $5\times$ magnification level have been utilized. At this magnification, the image resolution is 2μmm per pixel. In the first step, the entire image is divided into small blocks of 0.5 mm$\times$0.5 mm images. The choice of the block size corresponds to the size of the smallest annotated region present in our dataset. Each of these blocks are categorized into probable cancerous and non-cancerous region by using a random forest classifier [7]. The features exploited by the classifier are extracted from the segmented gland images.

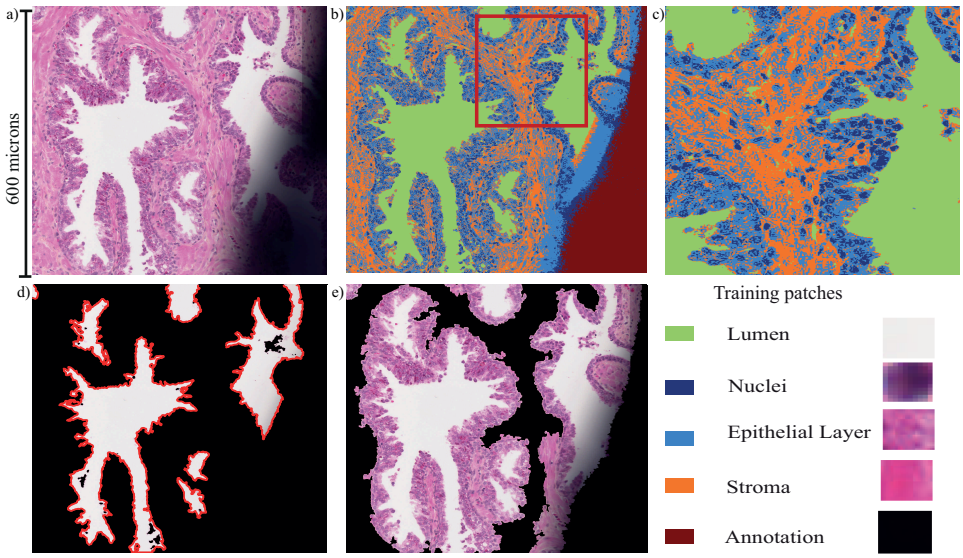


Figure 2.1: Gland segmentation. a) A sample image block, b) labeled image where each histological component is represented by a different color, c) enlarged view of a small window in the labeled image, d) lumen objects (the red mark corresponds the initial gland boundary), and e) segmented gland unit after consolidating surrounding epithelial layer-nuclei object with the gland lumen.

2.2 Gland Segmentation

The gland segmentation algorithm has been partially adopted from the work of Nguyen et al. [30]. In the first step, each image block has been segmented into five categories, i.e., i) Gland lumen, ii) Cytoplasm, iii) Nuclei, iv) Stroma, and v) Annotation mark. This segmentation uses the distinct color information of each category. Variations in illumination caused by variations in staining or changes in ambient lighting conditions at the time of digitization may dramatically affect image characteristics and then potentially affect the performance of the algorithm. In the RGB color space the lighting information and the color information is blended together. This is why each sub-region is converted from RGB color space to Lab color space. 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. By converting to Lab

color space the lighting information is confined into only one channel, L. Small training patches of each categories has been used to train the classifier, which labels each pixel of the image into one of the five different categories listed above.

For i^{th} pixel in either test data or training data, the pixel is represented as $D_{i,j}$ where $i = \{1,2,...,n\}$; n is the number of data points and $j = \{1,2\}$, for the two chromaticity layers in the Lab color space. The classification algorithm uses a linear discriminant analysis to label the testing pixels. Given a training data set with class known for each of the data point, the j^{th} component of the mean vector for class k is simply the mean for variable j over the N_k data points in group k.

$$\bar{D_{j,k}} = \frac{1}{N_k} \sum_{n \in k} D_{n,j}, \tag{2.1}$$

where $n \in k$ indicates the set of data points in group k.

The covariances matrices for each class is considered to be equal and estimated as single pooled estimate of S, with entries

$$S_{i,j} = \frac{1}{N-K} \sum_{n=1}^{N} \left(x_{n;i} - \bar{x_{k(n);i}}\right)\left(x_{n;j} - \bar{x_{k(n);j}}\right), \tag{2.2}$$

where $\bar{x_{k(n);i}}$ is the i^{th} component of the mean vector for which class the data point n belongs to, $k(n)$. Then the squared Mahalanobis distance from a data vector x to the mean of group of k is given by

$$z_k^2 = (x - \bar{x_k})' S^{-1} (x - \bar{x_k}). \tag{2.3}$$

As a result of pooled estimate of covariance matrix, all the determinants of covariance estimate is equal and the Bayes' formula for estimating posterior probability of data vector x to class k is reduces to

$$P_k(x) = \frac{q_k exp[-0.5 z_k^2]}{\sum_{l=1}^{K} q_l exp[-0.5 z_l^2]}. \tag{2.4}$$

Then the data vector x is assigned to the class with which it has maximum posterior probability. For i^{th} point, lets assume the data vector is x, the the corresponding pixel label (l_i) will be the $l_i = arg_{max} P_i(x)$.

The training patches used in the classifier are manually selected from 5 differ-

ent patient images. The number of training patches and the number of training pixels for each category are: lumen: 2 (Number of pixels: 1580), nuclei: 4 (Number of pixels: 440), epithelial layer: 4 (Number of pixels: 1564), stroma: 10 (Number of pixels: 3033), and annotation: 2 (Number of pixels: 1503). Among the histology components, the stroma units exhibit highest variation in color information. Therefore, more number of training patches for these category has been utilized. By contrast, the lumen and nuclei are the most homogenous tissue components in the image and hence lower number of training patches from these categories have been utilized.

Fig. 2.1b-c shows the result of the color based segmentation of an example image sub-region. For pixel labeling, we have used linear discrimination analysis instead of the Voronoi tesselation based approach from [30]. In the Vronoi tesselation approach, the training points create a Voronoi tessellation [2] of the Lab space. Each training point is associated with one convex polygon which includes all points closer to it than any other training point. Each test point is assigned to the same class associated with the training point of the polygon to which it belongs. The main drawback of Voronoi tesselation approach is that when the number of testing samples is large, the classification time for each testing data point is very high compared to that of linear discriminant analysis [20]. Therefore, when the number of testing samples is, very large the reported Voronoi tesselation based approach will be very expensive to compute.

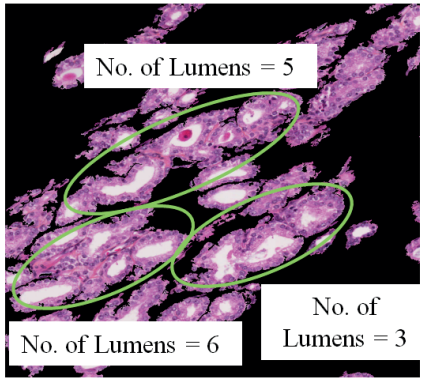
After having labeled the image into the categories listed above, first we group together the lumen pixels using a connected-components algorithm [35] which uses the eight-connectivity property. Then, the flood-fill [22] algorithm is employed to label all the pixels in the connected neighborhood. By putting a constraint on the maximum possible size of the lumen, some objects are discarded that are too big to be considered as a gland. This constraint eliminates the background object of the histology section that has almost the same color information as lumen objects. Around each lumen object, a lumen boundary is extracted. This is considered to be the primary gland boundary (see Fig. 2.1d). As stated earlier in the introduction section, a complete gland unit consists of the lumen and its surrounding layer of epithelial cells and nuclei. Therefore, to segment a complete gland unit we have to consolidate the surrounding epithelial layer and nuclei with the lumen. To accom-

plish this, an iterative search around $\delta \times \delta$ neighborhood centered at each boundary point of the lumen is carried out. Here, δ is set to 3 pixels, the minimum neighborhood window length, in order to eliminate the possibility of adding any extra non-gland pixels to the gland unit. The pixels labeled as epithelial layer or nuclei within this neighborhood are grouped together with the lumen object and the corresponding gland boundary is updated. This procedure stops when the gland boundary reaches the pixels labeled as stroma. Also, under the assumption that a true gland unit is always surrounded by epithelial layer, the lumen pixels that are not surrounded by the cytoplasm and nuclei are discarded as false lumen objects. Fig. 2.1e illustrates the resultant segmented gland units.

2.3 Feature Extraction

From each block of the image, an array of nine features related to the arrangement and morphology of the lumen, nuclei, and epithelial layer of the gland is extracted. The histological changes occurring in the malignant regions are most pronounced in these three tissue components. The stroma rarely shows any alteration in malignant regions. Malignant regions in prostate histology are usually characterized by group of closely packed glands that are similar in shape, whereas in benign regions, the glands usually have highly irregular shapes. The malignant glands are usually circular in shape and posses a thicker epithelial layer. In high grade of cancers, the malignant regions also exhibit high concentration of randomly floating nuclei. Based on these observations, we have synthesized a set of nine features to classify benign and malignant regions. Following are the detail description of the features:

Table 2.1: Features extracted from each image block

Nuclei features	Entropy of the Nuclei pixel distribution (E_n)	$E_n = \sum P_n \times logP_n$ P_n is the spatial histogram of nuclei pixels
	Nuclei Density (N_D)	$N_D = \frac{A_n}{A_b}$ A_n = Total area of nuclei units (in pixels) A_b = Total area of the block image (in pixels)
Lumen features	Mean Lumen Roundness Metric (M_{LRM})	$LRM_i = \frac{(\pi \times ED_i)}{P_i}$; P_i = Perimeter of the i^{th} lumen ED_i = Equivalent diameter of the i^{th} lumen
	Standard Deviation of the Lumen Roundness Metric (S_{LRM})	
	Maximum Number of Lumens per Cluster Lumens (MNLC)	
	Average Number of Lumens per Cluster (ANLC)	In this block image, MNLC=6; ANLC=(5+6+3)/3=4.67
Epithelial features	Average Epithelial layer Density per Gland unit (A_{EDG})	$EDG_i = \frac{A_{e(i)}}{A_{gland_i}}$ A_{e_i} = Area of epithelial layer of i^{th} gland A_{gland_i} = Area of the i^{th} gland
	Standard deviation of Epithelial layer (S_{EDG}) Density per Gland unit	
	Overall Epithelial layer Density (E_D)	$E_D = \frac{A_e}{A_b}$ A_e = Total area of the epithelial layer (in pixels)

2.3.1 Nuclei Features

Nuclei Density, N_D

This feature is calculated by taking the ratio of nuclear area to the total area of the image block. Formally, Nuclei density, $N_D = \frac{A_n}{A_b}$, where A_n = Total area of nuclei units (in pixels) and A_b = Total area of the block image (in pixels).

Nuclei Entropy, E_n

This feature is proposed here to capture the randomness of nuclei appearance in the block image. It is a newly proposed feature in cancer classification. We have quantified the Entropy of the nuclei pixels as, $E_n = \sum P_n \times logP_n$. Here, P_n is the spatial histogram of nuclei pixels.

2.3.2 Lumen Features

Lumen Roundness Metric (LRM)

The LRM is a measure of circularity of the lumen shapes. Two features, Average LRM (ALRM) and Standard deviation of the LRM (SLRM) are extracted from each of the image blocks to incorporate the shape information in the feature set. The LRM of i^{th} gland is calculated as, $LRM_i = \frac{(\pi \times ED_i)}{P_i}$. Here P_i = Perimeter of the i^{th} lumen and ED_i = Equivalent diameter of the i^{th} lumen.

Lumen Cluster

: Clustering of the lumens is a common regional property of prostatic adenocarcinoma. Two new features, Maximum Number of Lumens per Cluster (MNLC) and Average Number of Lumens per Cluster are calculated in each image block to represent the clustering of lumens in the feature set.

2.3.3 Epithelial Features

Three new features associated with the epithelial layer structure are proposed here. The malignant regions usually posses a higher epithelial area compared to the nor-

mal regions. The overall Epithelial Density (ED) is calculated by taking ratio of epithelial area to total area of the image block, i.e., $E_D = \frac{A_e}{A_b}$, where A_e = Total area of the epithelial layer (in pixels). In addition to the overall higher density of epithelial layers, the individual glands also posses a thicker epithelial layer. The Epithelial layer Density per Gland unit (EDG) for i^{th} gland in the image block is calculated as $EDG_i = \frac{A_{e(i)}}{A_{gland_i}}$, where A_{e_i} = Area of epithelial layer of i^{th} gland and A_{gland_i} = Area of the i^{th} gland.

These features are tabulated in Table 2.1.

2.4 Classification of Malignant Regions

These features are utilized by the random forest classifier for separating benign and malignant blocks. The random forest algorithm was developed by Leo Breiman and Adele Cutler [7]. Random forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. Each tree in the random forest can be considered as a 'weak' learner and in the ensemble they come together to form a 'strong' learner. Single decision trees often have high variance or high bias. Random forests attempt to mitigate the problems of high variance and high bias by averaging to find a natural balance between the two extremes. Each tree of random forest ensemble has been trained by randomly selecting two thirds of the samples each time with replacement. The remaining samples are used to test the tree and the mean squared error in classification of all the trees constitutes the out-of-the-bag classification error. Fig. 2.2 illustrates the plot of the out of the bag classification error against the number of trees. As the number of trees in the ensemble goes up, the classification error goes down. As can be observed from the figure, with our proposed features the out of the bag classification error goes down to 0.04 for 100 trees.

In the random forest classifier model, there are two parameters to be tuned, i) minimum leaf size and ii) threshold of class probability for the classifier model. We have determined the optimum value for these parameters by performing a Receiver Operating Characteristics (ROC) analysis on 50 out of the 70 images of the dataset. The parameter selection process and the detail cross-validation of the classifier are

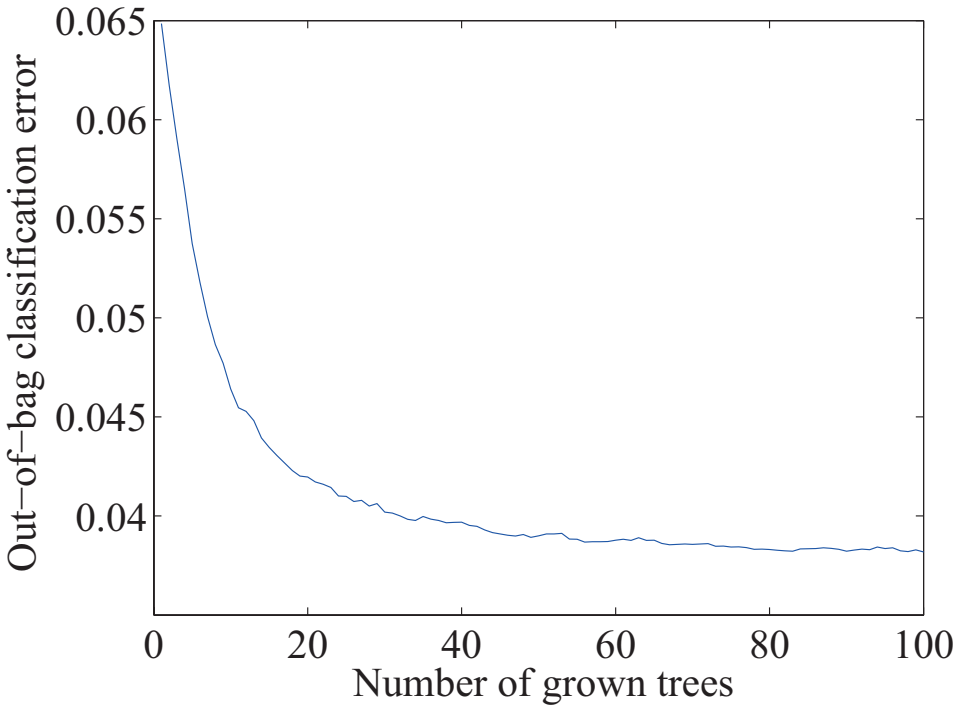


Figure 2.2: Plot of Out-of-the-bag classification error against the number of grown trees.

described in detail in the results section. The remaining 20 images constitutes test set and the performance of the proposed technique on the test is obtained by a leave-25%-out experiment with the tuned model parameters. After the classification, the detected blocks are grouped together to form a continuous area. Fig. 2.3 illustrates the result of the first stage of cancer annotation on a sample image. More results are provided in the results chapter. A finer annotation of the detected regions is performed in the next step at a higher magnification.

2.5 Conclusion

In this chapter, we presented a cancer annotation approach at the regional level. With the help of the extracted features from segmented gland units, a random forest classifier separates the probable cancerous regions from the benign regions. The array of regional features proposed here quantify the histopathology changes

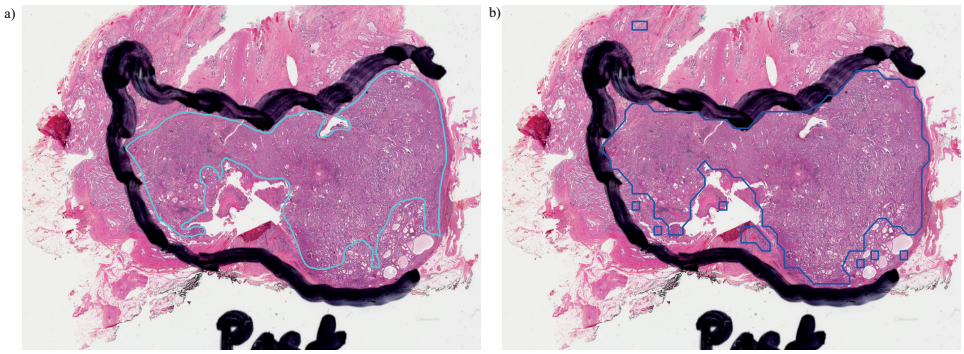


Figure 2.3: Performance of the proposed algorithm on a test image. a) Cyan annotation is performed by pathologist and considered as the ground truth. b) Result of the screening phase of the proposed technique.

occurring in a cancerous region. Changes in all the tissue components, such as lumen, epithelium, and nuclei, are accounted here to generate a unique set of features that is used for classification using a random forest classifier.

Chapter 3

Cancer Classification Using Glandular Features

3.1 Introduction

Glands are the basic building blocks of the prostate. Prostatic adenocarcinoma, the most common type of prostate cancer i.e., prostatic adenocarcinoma is originate from the epithelium layer of the glands. Subsequently, the evolution of malignancy is most evident at the glandular level. Here we focus on the quantification of the changes that occur in the individual gland units from high resolution images. In this step, the detected regions from the previous step are further magnified to perform gland-level analysis. In our experiments, we use whole-mount histology scans at magnification of $20\times$ with a resolution of 0.5 μm per pixel. The gland units in the cancer-probable regions are classified based on their individual gland specific properties. The features exploited here are, 1) Number of Nuclei Layers and 2) Ratio of Epithelial Layer Area to Lumen Area. To quantify the number of nuclei layer associated with each gland, the nuclei units are segmented automatically. Then based on these two features, the benign and malignant glands are separated and the final annotation consolidating the malignant glands is obtained.

3.2 Segmentation of the Nuclei Unit

The nucleus is the smallest visible histological component present in a pathology image. Several techniques for segmenting nuclei [1], [3], [4], [6], [18], [24], [38], [39] have already been proposed exploiting very high magnification images (40× or higher). However, automatic segmentation of nuclei from images with lower magnification is yet to be investigated. Here we employ a modified watershed algorithm for automatic nuclei segmentation exploiting the foreground and background object markers.

For the nuclei segmentation, only the 'R' channel of the image has been used since it produces maximum histogram separation between nuclei and non-nuclei objects. The nuclei objects appear as dark objects in the tissue image. The 'R' channel is inverted to make the nuclei units foreground objects for segmentation (see Fig. 3.1 a). The preprocessing steps before applying the marker controlled watershed algorithm are the background subtraction and thresholding. The background of the image is estimated by performing a morphological opening of the image with a disk shaped structural element of radius 10. The radius is chosen such that the element cannot fit inside an individual nuclei unit. Therefore, nuclei units are not affected by the morphological filtering to estimate the background. The threshold for separating the nuclei units is computed by minimizing the intra-class variance of the image [32]. The resultant preprocessed image after background subtraction and thresholding is shown in Fig. 3.1b.

The segmentation function used in this watershed algorithm is the gradient image. But before applying the watershed algorithm, foreground and background markers need to be computed to reduce over segmentation. To compute the foreground markers, morphological opening and closing by reconstruction is performed on the preprocessed image (Fig. 3.1b) to create a flat maxima inside each of the foreground objects which are used as the foreground markers. The watershed ridge line calculated from the Euclid distance transform of the thresholded binary image is taken as the background marker (see Fig. 3.1c). After computing the markers, the final watershed segmentation is performed on the gradient image of the preprocessed input scene. This gradient image is modified by placing regional minima in the marked pixels of foreground and background objects of the image (see Fig

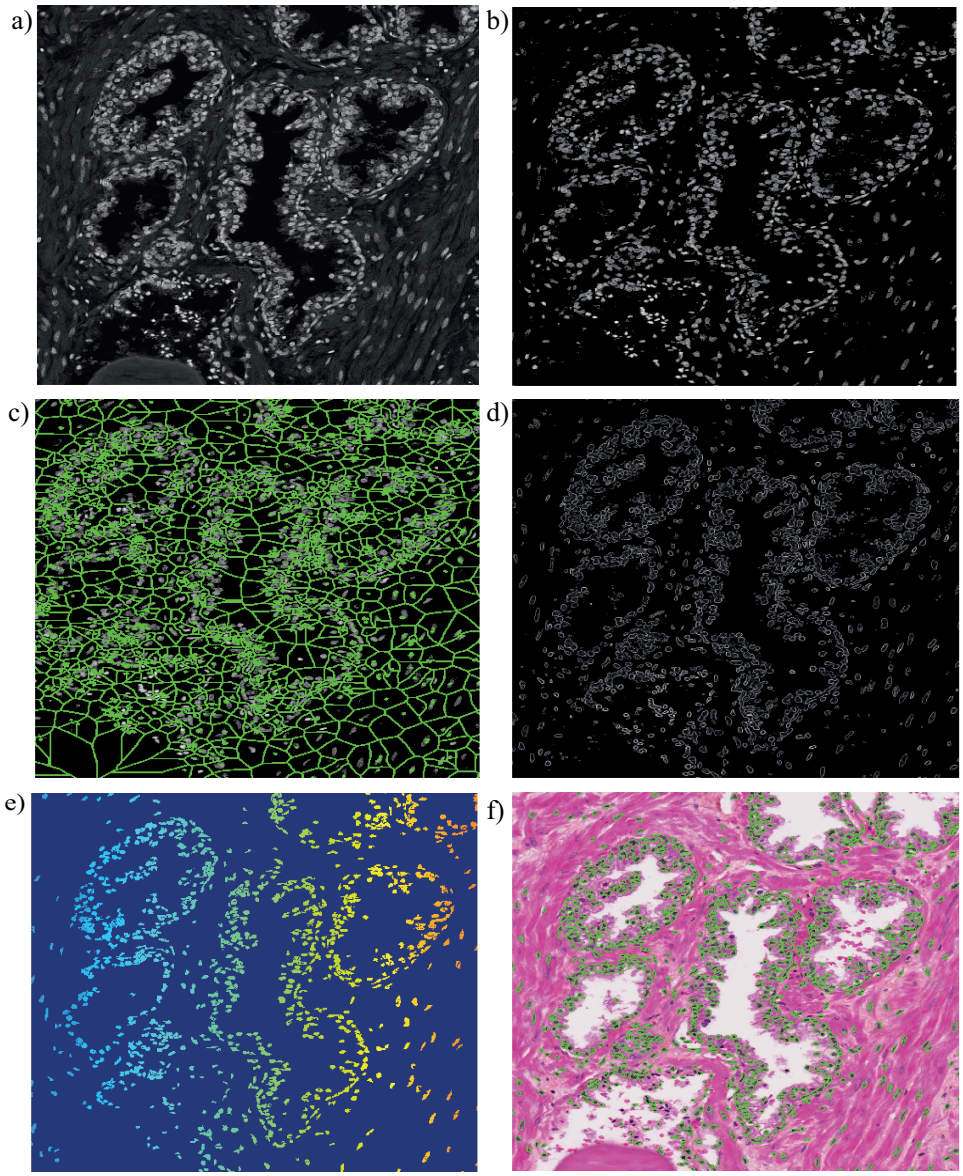


Figure 3.1: a) Input scene: inverted 'R' channel of the histopathology image, b) preprocessed image after background subtraction and thresholding, c) computed foreground and background markers overlaid on the preprocessed image, d) modified gradient image with regional minima placed at the foreground and background object markers, e) labeled image of the segmented nuclei after watershed transform, and f) segmented nuclei (marked by green boundary).

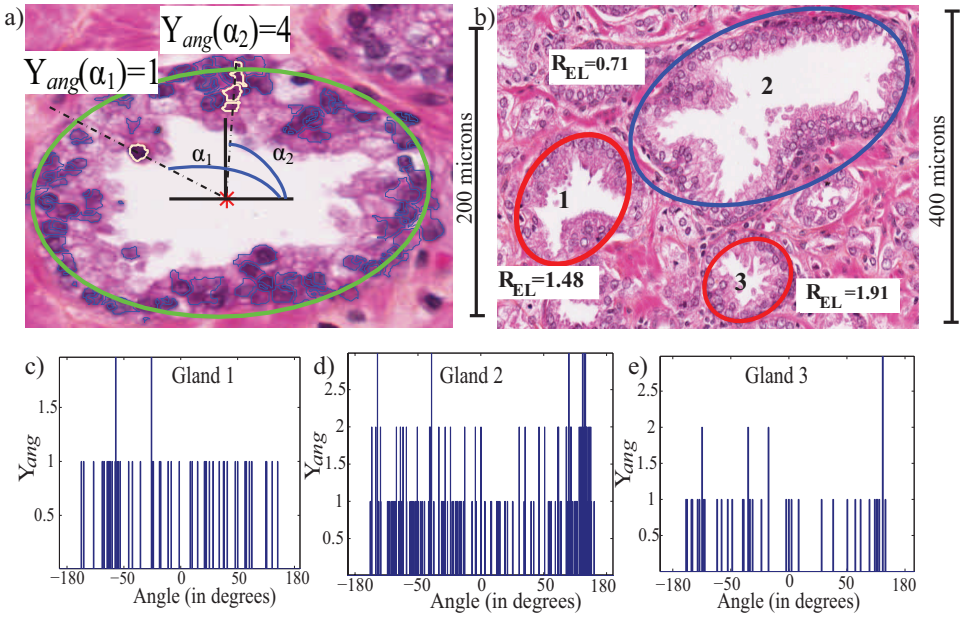


Figure 3.2: a) Graphical illustration of Y_{ang} calculation. b) Sample histopathology scene with a single benign (marked with blue ellipse) and two malignant gland units (marked with red ellipses). c), d), and e) illustrate the different appearances of the histograms (Y_{ang}) of benign and malignant glands.

3.1d). These modified regional minima limits the number of segments produced the watershed segmentation. Fig. 3.1d-e illustrates the final segmentation of the nuclei units.

3.3 Extraction of Gland-specific Features

3.3.1 Number of Nuclei Layers (N_{NL})

To determine the number of nuclei layers pertaining to each gland, at first the segmented nuclei objects are paired with the corresponding gland unit that minimizes the distance between the centroid of the nuclei and the gland lumen boundary. For each of the combined gland-nuclei object, an ellipse is fit around it. The angular location of each of the nuclei is evaluated by calculating the angle of the connect-

ing line of the gland centroid and corresponding nuclei centroid (see Fig. 3.2a). Then the feature N_{NL} is evaluated from the histogram Y_{ang} of angular locations of nuclei. Customized bin spacing has been utilized to account for glands of different sizes. The bin spacing for the histogram is evaluated as $360^{\circ}/P_g$, where P_g is the perimeter of the corresponding ellipse surrounding the gland. Then the N_{NL} is evaluated by counting the total number of instances where multiple nuclei have the same angular bin in the histogram and then normalizing it by dividing by P_g. Mathematically, $N_{NL} = \frac{1}{P_g}|\{n|Y_{ang}(n) >= 1\}|$. Fig. 3.2c-e illustrates the different nature of histogram, Y_{ang} in case of benign and malignant glands. As can be observed from the figure, the benign histogram provides more instances of multiple nuclei having same angular location.

3.3.2 Ratio of Epithelial Layer area to Lumen Area (R_{EL})

This feature is evaluated by simply taking the ratio of the epithelial layer area to lumen area of the gland. In case of malignant glands, fast multiplication of cells lead the epithelial layer to invade more in to gland lumen. As a result, the ratio gets larger in case of malignant gland units.

After the feature extraction we choose optimum thresholds on the features, $\tau_{N_{NL}}$ and $\tau_{R_{EL}}$ for the classification of benign and malignant glands. We classify a gland (G_i) as benign when the parameters fulfill the following criteria, $Label_{G_i} = \{Benign|N_{NL}(G_i) > \tau_{N_{NL}}, R_{EL}(G_i) < \tau_{R_{EL}}\}$. These threshold parameters are tuned by performing a ROC analysis that will be discussed in the following section of the thesis.

3.4 Consolidation of Glands

The final stage of the algorithm consolidates the malignant glands into continuous regions. The glands are separated into distinct groups based on their parent region from the previous step. One approach for the encapsulation of the detected glands is to generate a convex hull of the gland centroids. Unfortunately, since the true spatial extent of the prostate cancer rarely forms convex hulls, using such an algorithm will not represent the true extent of the detected regions. As a solution to this problem we have utilized α-shape approach [11] of generating a continuous

boundary from a point cloud. In this approach, a continuous area from a point cloud can be generated by point pairs that can be touched by an empty disc of radius α. A graphical illustration of the hull generation is presented in Fig. 3.3. To implement this technique, at first 2-D Delauney [10] triangulation of the points are obtained. Each edge/triangle of the Delaunay triangulation is associated with a characteristic radius, the radius of the smallest empty circle containing the edge or triangle. For a specific radius α, the α-complex of the given set of points is the complex formed by the set of edges and triangles whose radii are at most α. The union of the edges and triangles in the α-complex forms the α-shape. Here we have chosen the radius $\alpha = 0.12$ mm in the order of a typical size of a malignant gland unit. Fig. 3.4 illustrates the final annotation obtained by the proposed technique on the sample image demonstrated in the previous chapter.

3.5 Conclusion

This chapter elaborates on the detail of individual gland labeling which include nuclei segmentation, gland-specific feature extraction, and consolidation technique for detected malignant glands. Apart from cancer annotation, these gland specific features have the the potential to diagnose other prostate anomalies too, such atrophy and benign prostatic hyperplasia. One limitation of individual gland based approach is the less frequent Grade 5 cancers. At Grade 5, the individual glands get raptured and no longer becomes visible as a single gland unit. Therefore, individual gland labeling will generate false negative regions. One future improvement on this technique can be to propose a solution to this problem.

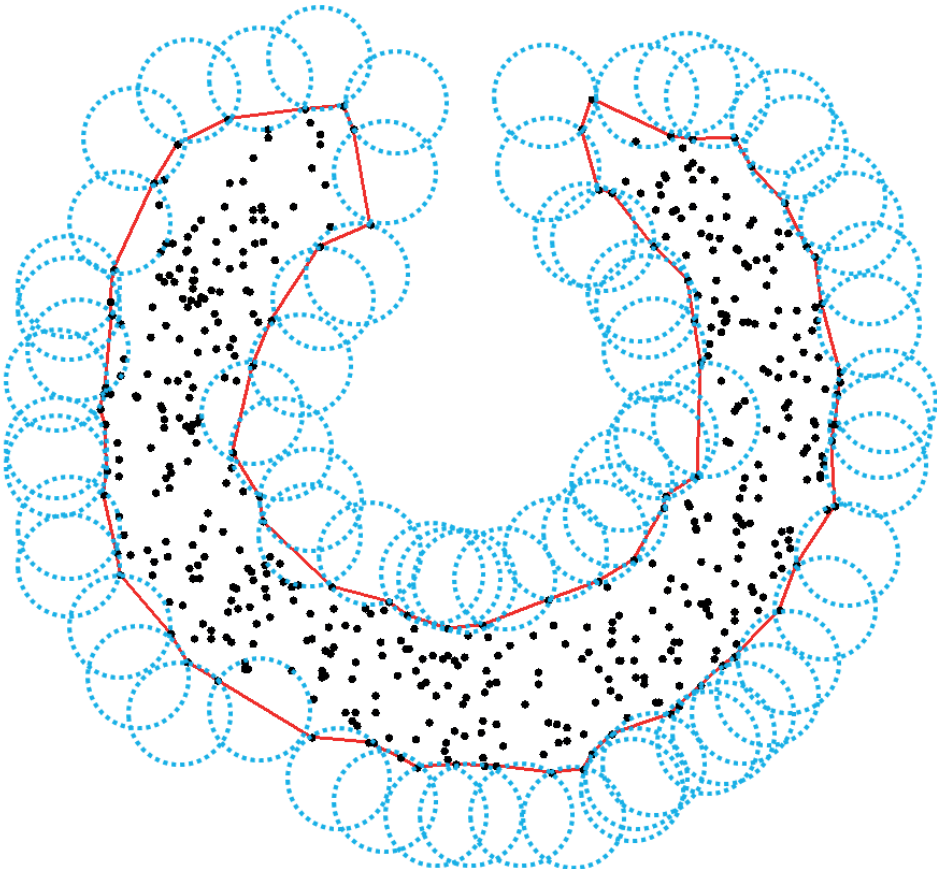


Figure 3.3: a) Graphical illustration of boundary hull generation by α-shape approach.

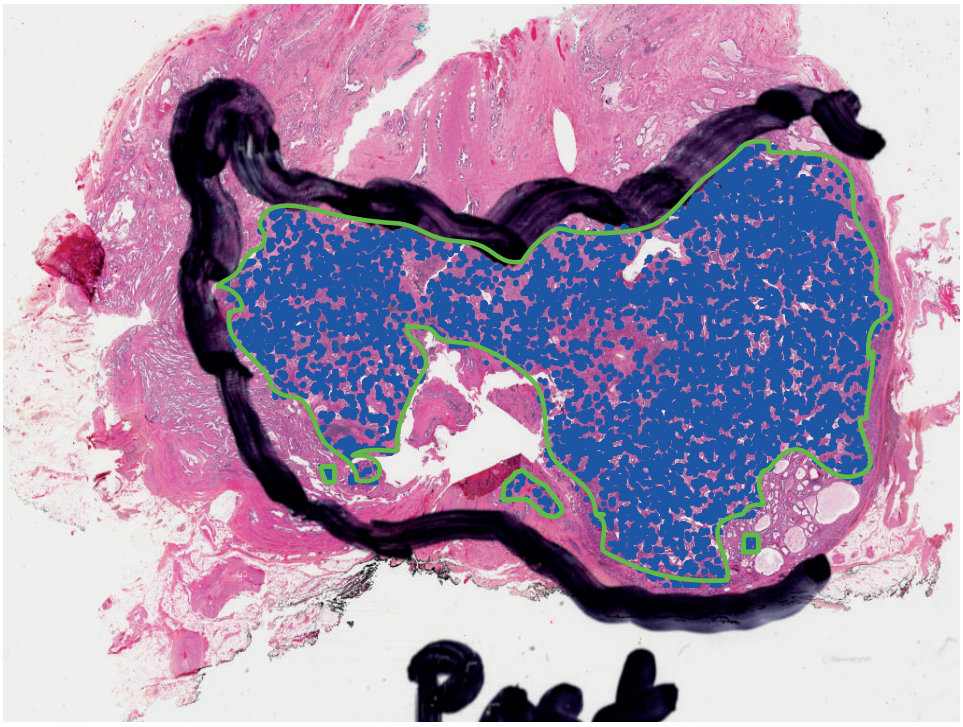


Figure 3.4: a) Experimental result of the proposed technique after incorporating the glandular level classification with the regional cancer classification.

Chapter 4

Experimental Results

4.1 Dataset

The proposed algorithm has been evaluated on 70 different histopathology images obtained from 30 radical prostatectomy patients. These whole mount histopathology images are digitized at 20$\times$ magnification (0.5 μm per pixel) with an Aperio scanner. Each image was annotated by two pathologists. At first, the images were marked by a pathologist on the glass slide before digitization. Then, a second pathologist performed a detailed annotation on the digitized images. The annotations from the second pathologist have been used here as the gold standard to evaluate the performance of the proposed algorithm.

4.2 Parameter Tuning for Random Forest Classifier

Among the 70 images, 50 have been exploited to develop the random forest classifier model. The parameters of the random forest classifier, the minimum leaf size of the tree and the classification threshold have been tuned by performing a 10-fold cross validation for each set of the parameters on the 50 images. For each leaf parameter in the set $Rdf_{leaf} = \{1, 12, 23, ..., 100\}$ a ROC curve is generated by varying the threshold $Rdf_{th} = \{0.20, 0.23, 0.26, ..., 0.80\}$. Fig. 4.1 illustrates the the ROC curve obtained for the optimum Rdf_{leaf} value. Table III provides the Area Under the Curve (AUC) for each of the ROC curves. At the optimum

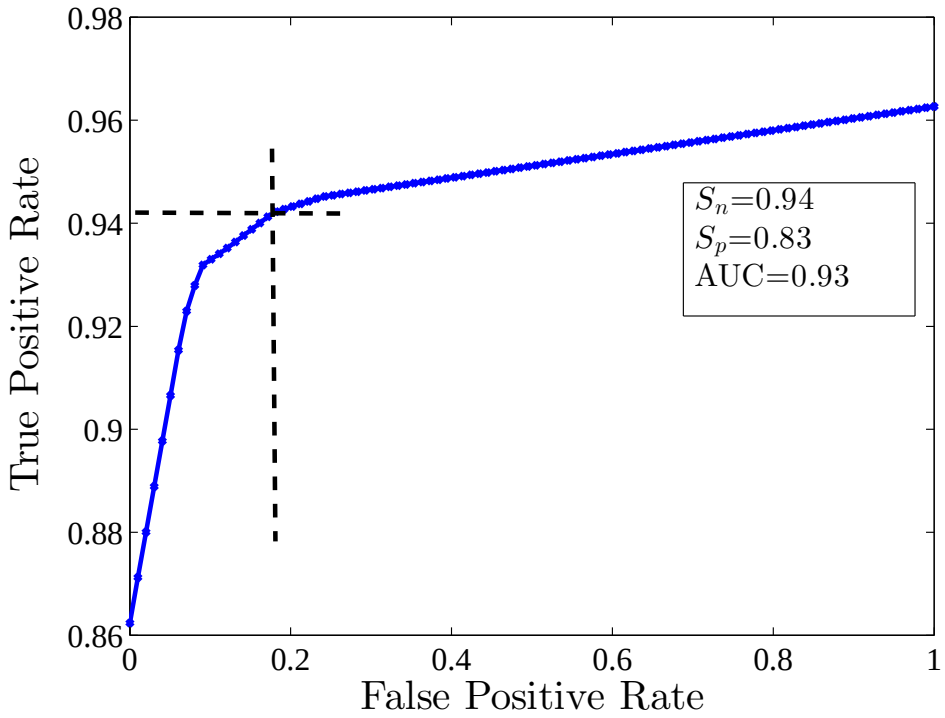


Figure 4.1: ROC analysis for the parameter tuning of the random forest classifier. Here the ROC curve with $Rdf_{leaf} = 12$ is shown. At the optimum operating point the sensitivity $S_n = 0.94$ and specificity $S_p = 0.83$.

operating point, the parameters are $Rdf_{leaf} = 12$ and $Rdf_{th} = 0.58$. At this optimum point the sensitivity is 0.94 and specificity is 0.83. With these parameters, the remaining 20 images have been tested in a leave-25%-out cross validation experiment. In 50 independent repetitions of the experiment, the algorithm achieved sensitivity, specificity, and accuracy are 0.88, 0.92, and 0.92, respectively. The probable malignant regions detected after the random forest classification are then further classified using their gland specific properties.

Table 4.1: AUC obtained by our algorithm for different parameter values of Rdf_{leaf}.

Rdf_{leaf}	1	12	23	34	45	56	67	78	89	100
AUC	0.90	0.93	0.94	0.92	0.90	0.92	0.88	0.87	0.85	0.80

4.3 Parameter Tuning for Individual Gland Classification

The performance of the algorithm at the gland level is influenced by the choice of the parameter values R_{EL} and N_{NL}. We tune the parameters by performing a similar ROC analysis on a set of individually labeled benign and malignant glands. A total of 4230 labeled glands have been used in this tuning process. The ROC curve of the classifier is generated by varying the parameter N_{NL} as $\{0, 0.08, 0.16, ..., 4\}$. To determine the effect of varying R_{EL} on the classifier performance the following operation has been performed: for each choice of R_{EL} in the set $R_{EL} = \{0, 0.25, ..., 2.5\}$ the individual ROC curve by varying N_{NL} has been generated. We choose the thresholds $\tau_{N_{NL}}$ and $\tau_{R_{EL}}$ corresponding to the optimum operating point in the ROC curve. In this experiment, we found the thresholds to be $\tau_{N_{NL}} = 2.36$ and $\tau_{R_{EL}} = 0.92$. At the optimum operating point, the sensitivity and specificity are 0.79 and 0.85, respectively. In each test image, the glands present in the pre-screened malignant regions are classified using these two parameters. With the gland level analysis, the sensitivity, specificity, and the accuracy of the proposed technique reaches 0.88, 0.94, and 0.93, respectively.

4.4 Qualitative Performance Evaluation

For qualitative performance evaluation, we have illustrated four test WM images from four different patients (see Fig. 4.3- 4.6). The first column represents the WM images with the pathologist's annotation (green). The middle column is the intermediate classification result at the $5\times$ resolution (blue) and the third column is the final annotation after the gland-level analysis (green). A strong agreement between the final annotation and the pathologist's annotation corroborates the effectiveness of the proposed algorithm.

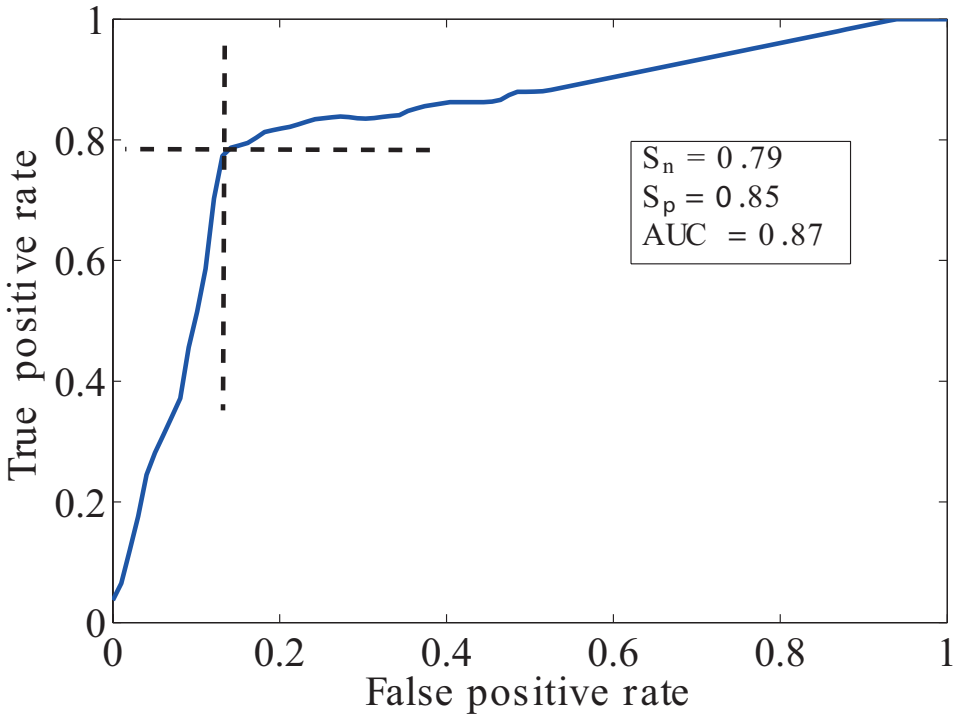


Figure 4.2: ROC analysis of the individually labeled gland dataset. Here, the ROC curve for $R_{EL} = 0.93$ is shown. At the optimum operating point the sensitivity $S_n = 0.79$ and specificity $S_p = 0.85$.

4.5 Discussion

The individual gland classification approach is the major novelty of the proposed approach. When this technique is applied independently without combining with the regional approach, it can detect other tissue anomalies in the prostate. Most of the prostate diseases are closely related to the individual gland units and directly affect gland morphology [21]. Therefore, beyond cancer annotation, this individual gland labeling can potentially generate a map of abnormality in prostate tissue. Since, these abnormalities are not reported in current clinical protocol of WM analysis, a gold standard to compare the performance of the proposed technique could not be collected. When we perform the individual gland labeling of the whole mount images, a distribution of the abnormal glands is observed. Apart from detecting malignant glands, the approach detects glands from Prostatic Intraepithelial

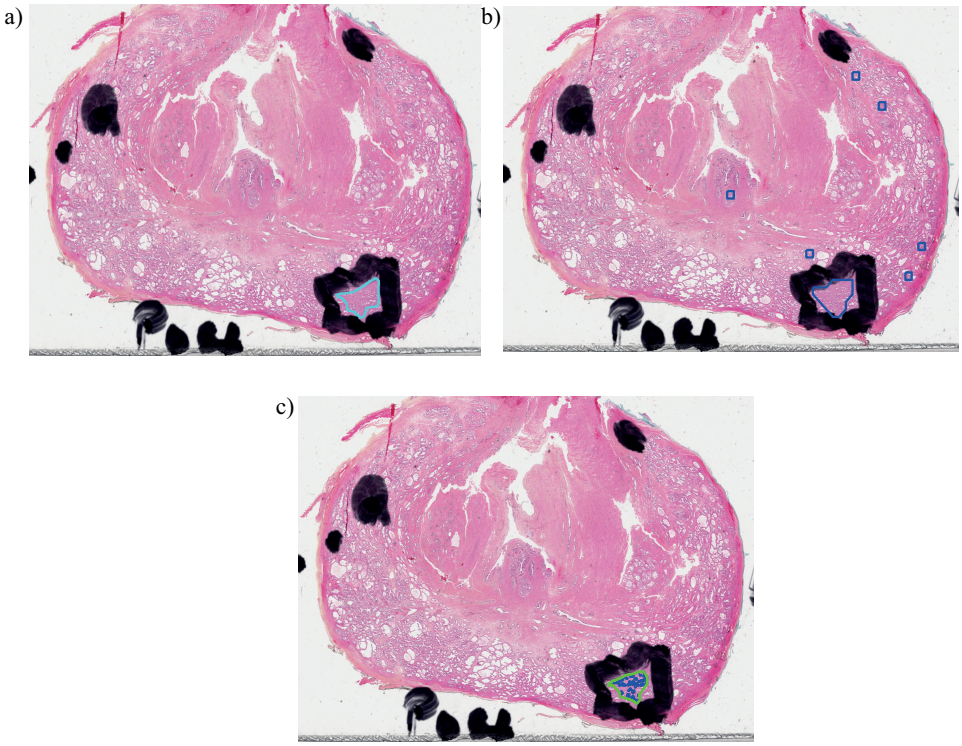


Figure 4.3: Performance of the proposed technique on a sample test image (Case: 1). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5× resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units.

Neoplasia (PIN), seminal vesicle , and prostatic atrophy. Fig. 4.12, illustrates an example image of applying individual gland classification in WM images. The enlarged figures illustrates the detected malignant glands, atrophic glands, and pre-cancerous PIN. Further improvement of this technique to detect only malignant units can be achieved by identifying basal cell layers that are present in only the benign units. To identify basal layer, a different chemical staining of the pathology images with Glutathione S-Transferase π (GST-π) is required [27]. More test images with the individual gland segmentation is illustrated in Fig. 4.13- 4.39.

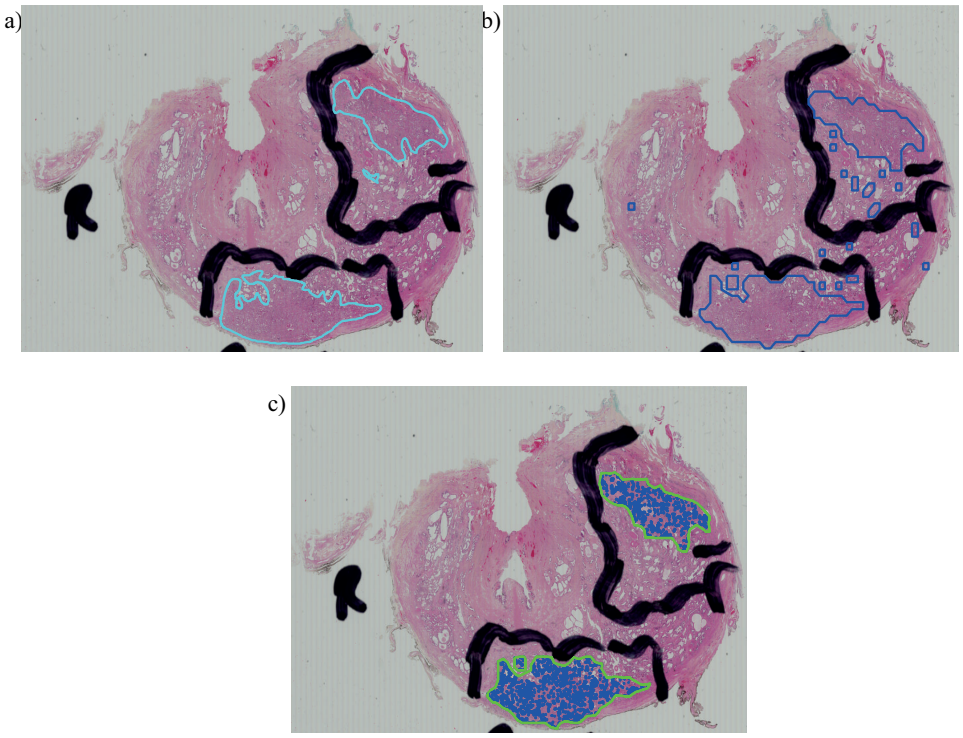


Figure 4.4: Performance of the proposed technique on a sample test image (Case: 2). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5× resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units.

The combined classification accuracy obtained by our proposed technique is the highest among the previously reported techniques on automatic cancer annotation of whole mount slides [26], [29]. Incorporating individual gland based approach with the region based approach increases the specificity by 2% and accuracy by 1%.

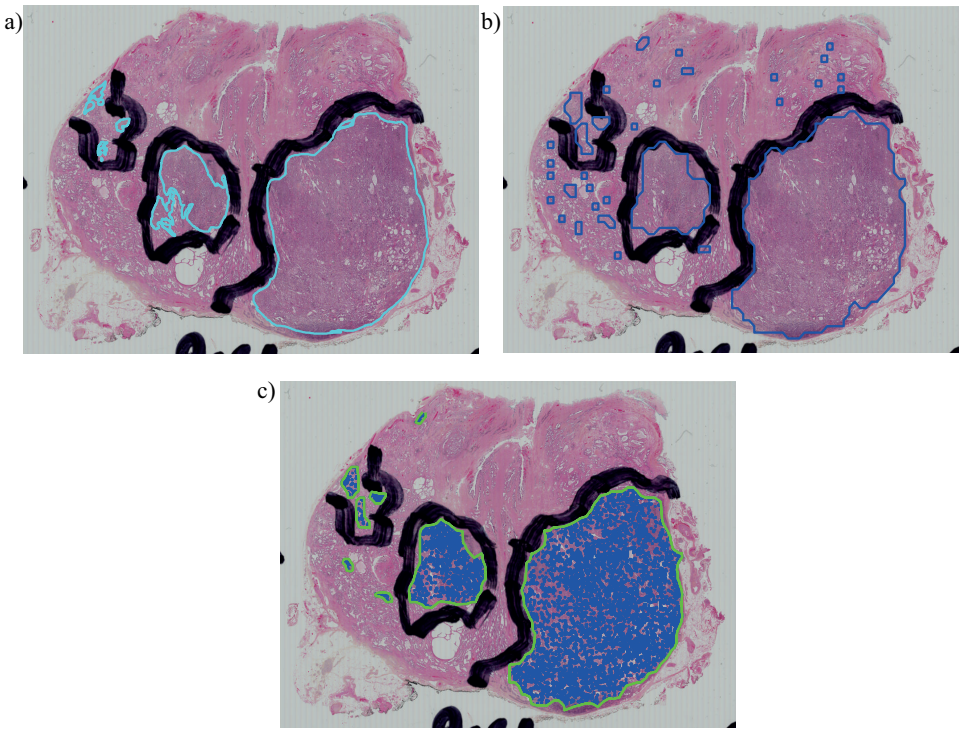


Figure 4.5: Performance of the proposed technique on a sample test image (Case: 3). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5× resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units.

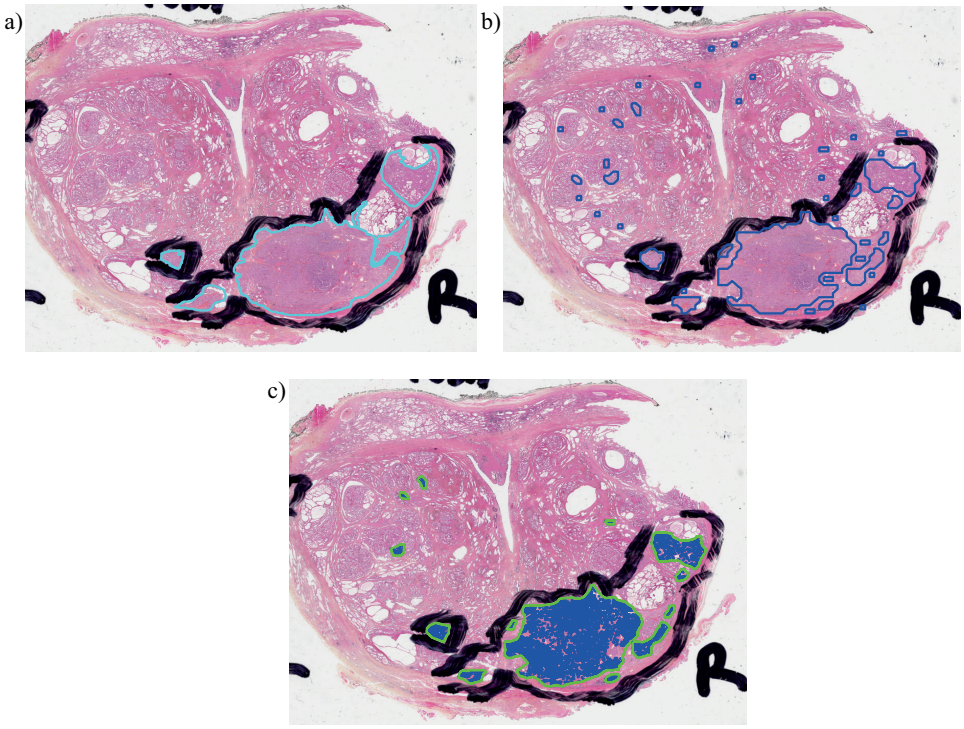


Figure 4.6: Performance of the proposed technique on a sample test image (Case: 4). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5× resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units.

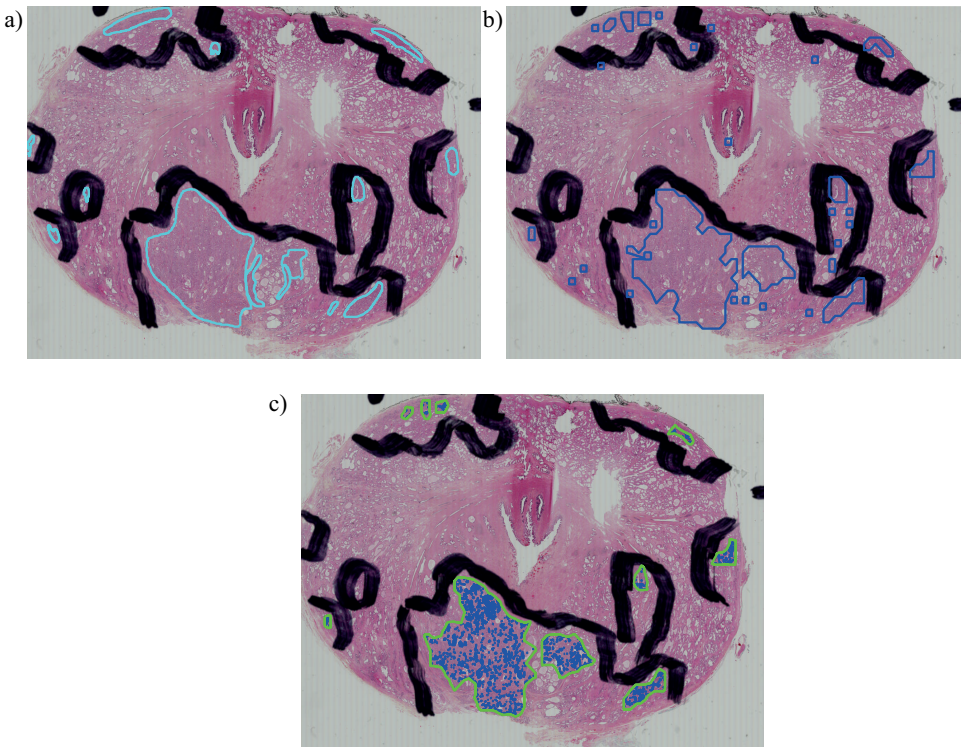


Figure 4.7: Performance of the proposed technique on a sample test image (Case: 5). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5x resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units.

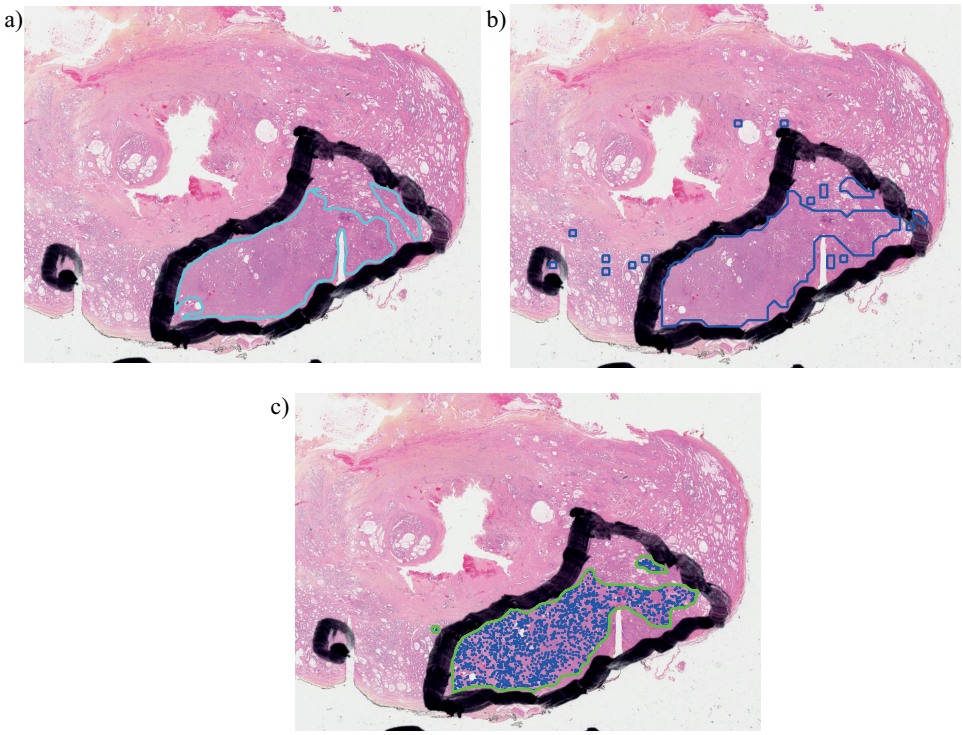


Figure 4.8: Performance of the proposed technique on a sample test image (Case: 6). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5x resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units.

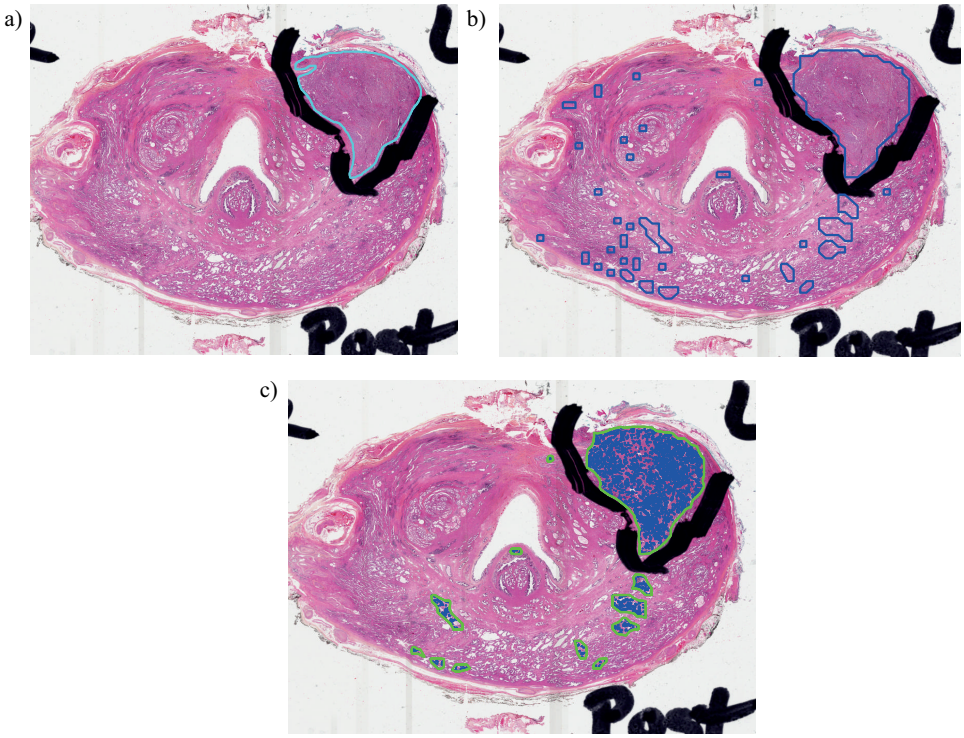


Figure 4.9: Performance of the proposed technique on a sample test image (Case: 7). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5x resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units.

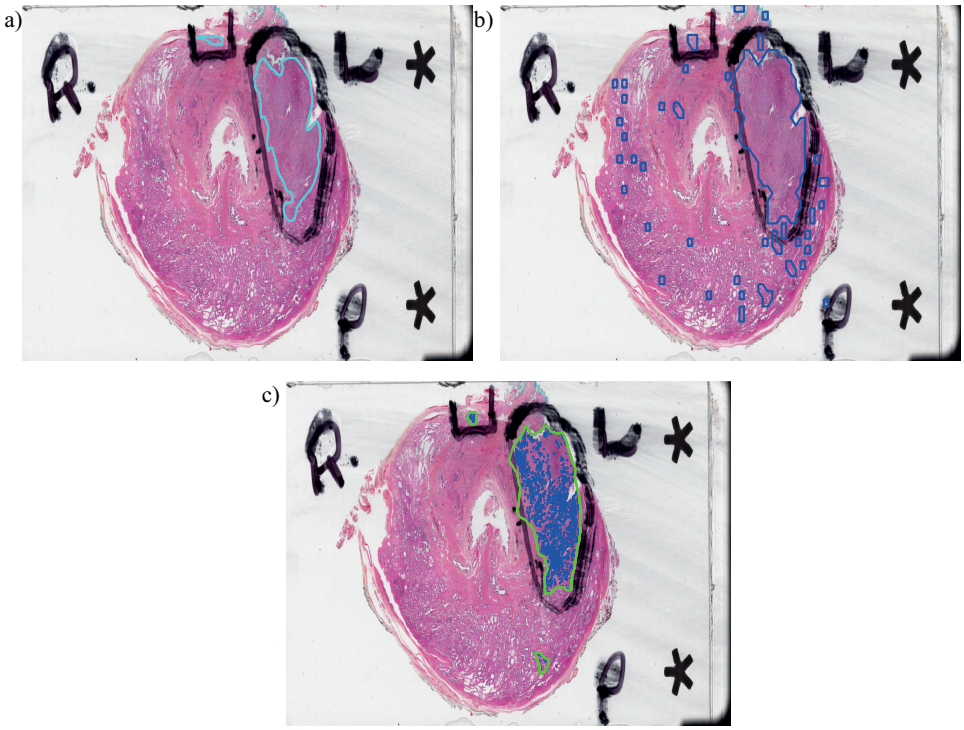


Figure 4.10: Performance of the proposed technique on a sample test image (Case: 8). a) The test image with both pathologists' annotation overlaid. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5x resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units.

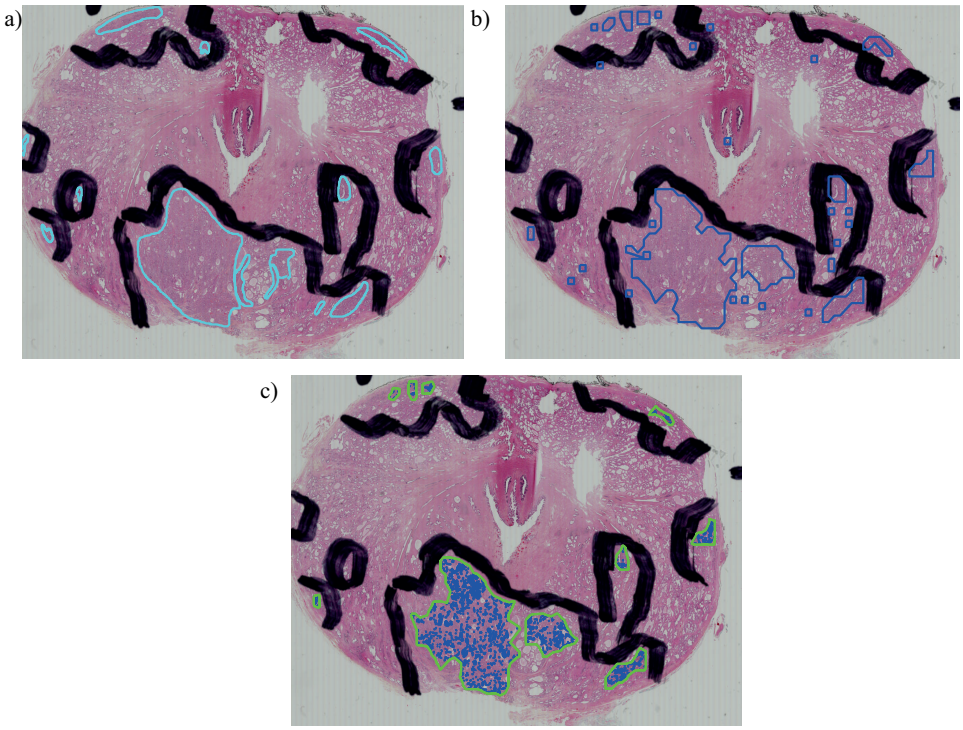


Figure 4.11: Performance of the proposed technique on a sample test image (Case: 9). a) The test image with both pathologists' annotation overlayed. The cyan annotation is from the second pathologist and considered as the ground truth. b) The blue mark is the intermediate annotation after random forest classification at 5x resolution. c) The green mark is the final cancer annotation obtained by the proposed technique.The blue dots represent the detected malignant gland units.

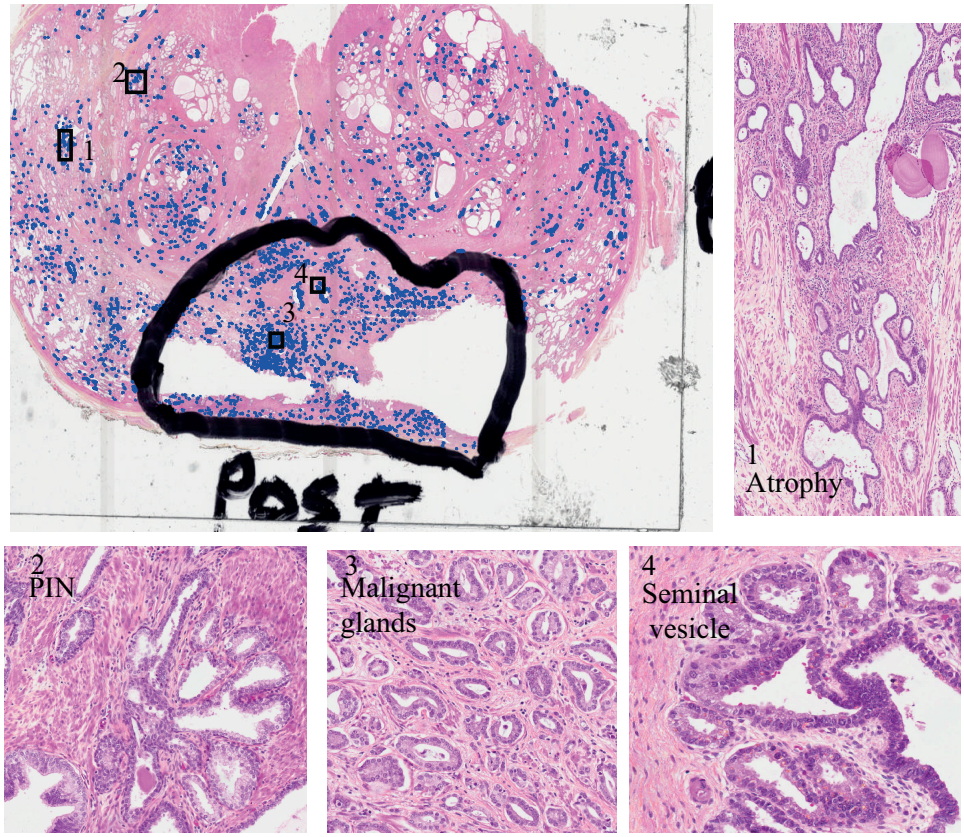


Figure 4.12: Illustration of the individual gland labeling on whole mount images. Enlarged view of four different types of detected glands, 1). Atrophy, 2) PIN, 3) Malignant glands, and 4) Seminal vesicle.

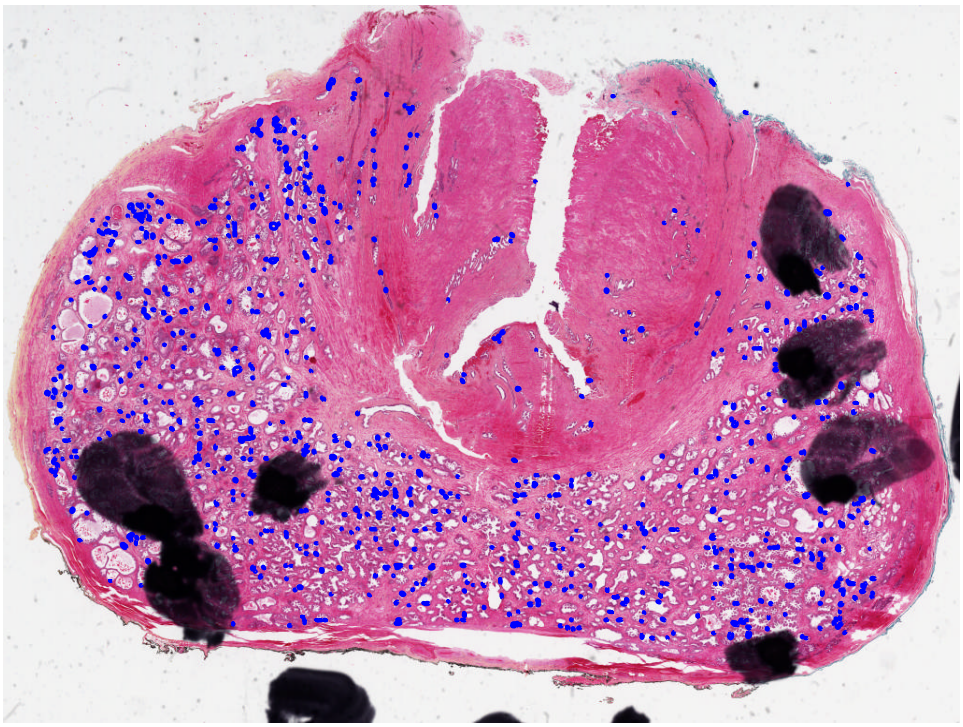


Figure 4.13: Illustration of the individual gland labeling on a sample whole mount image (Test image:1)

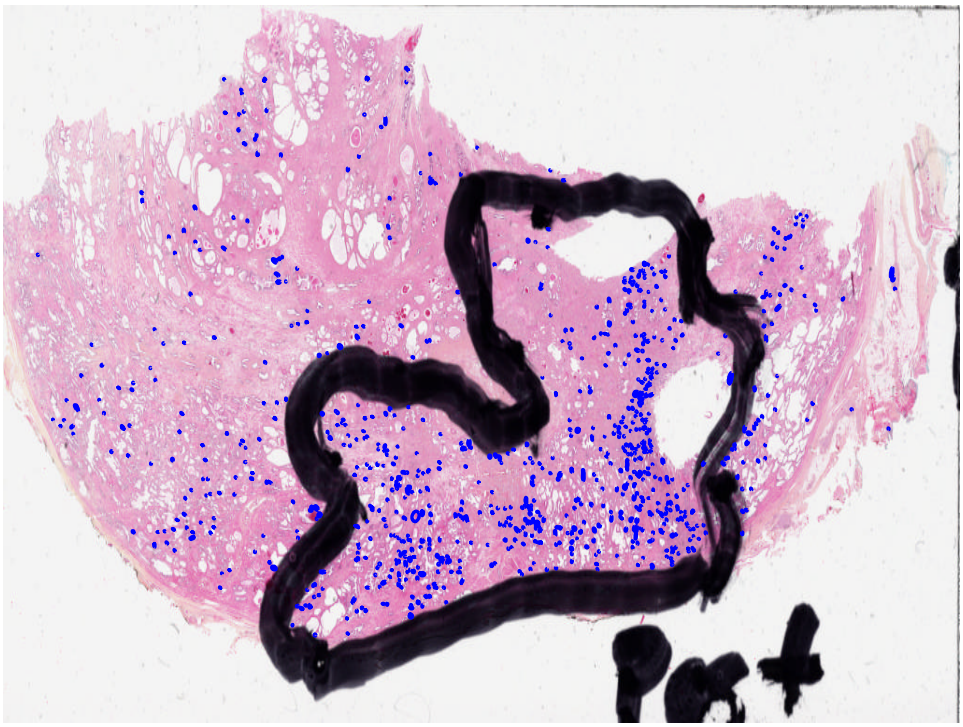


Figure 4.14: Illustration of the individual gland labeling on a sample whole mount image (Test image:2)

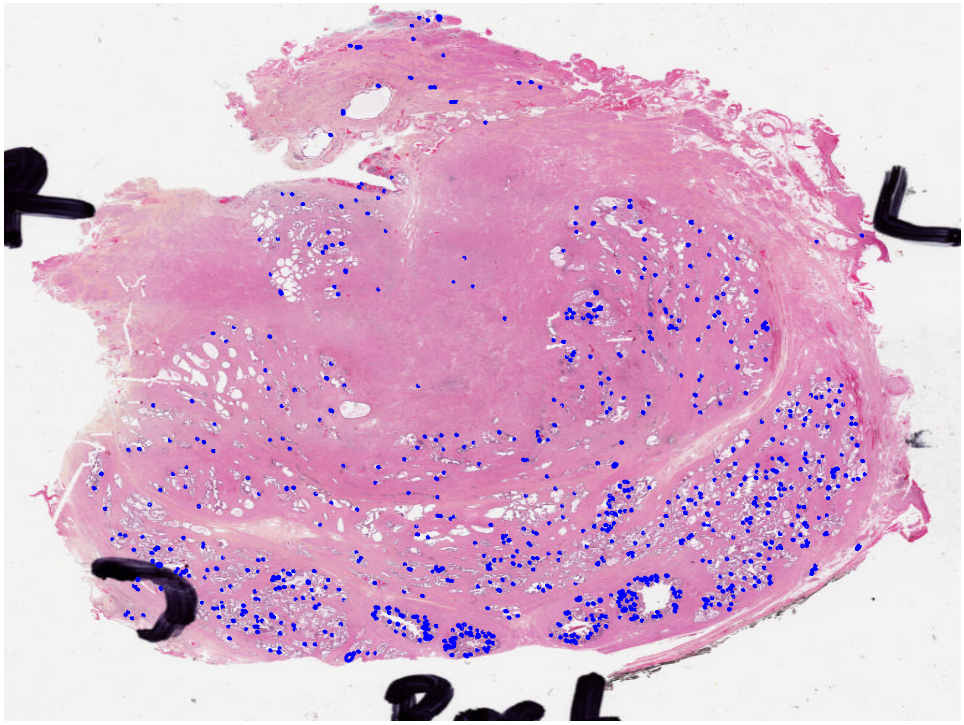


Figure 4.15: Illustration of the individual gland labeling on a sample whole mount image (Test image:3)

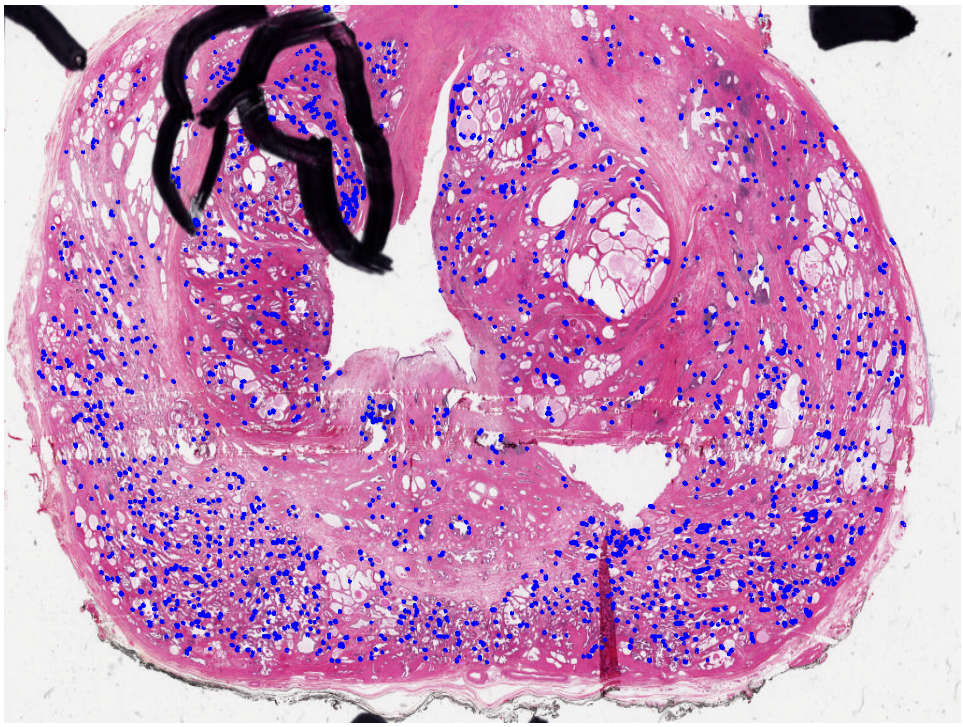


Figure 4.16: Illustration of the individual gland labeling on a sample whole mount image (Test image:4)

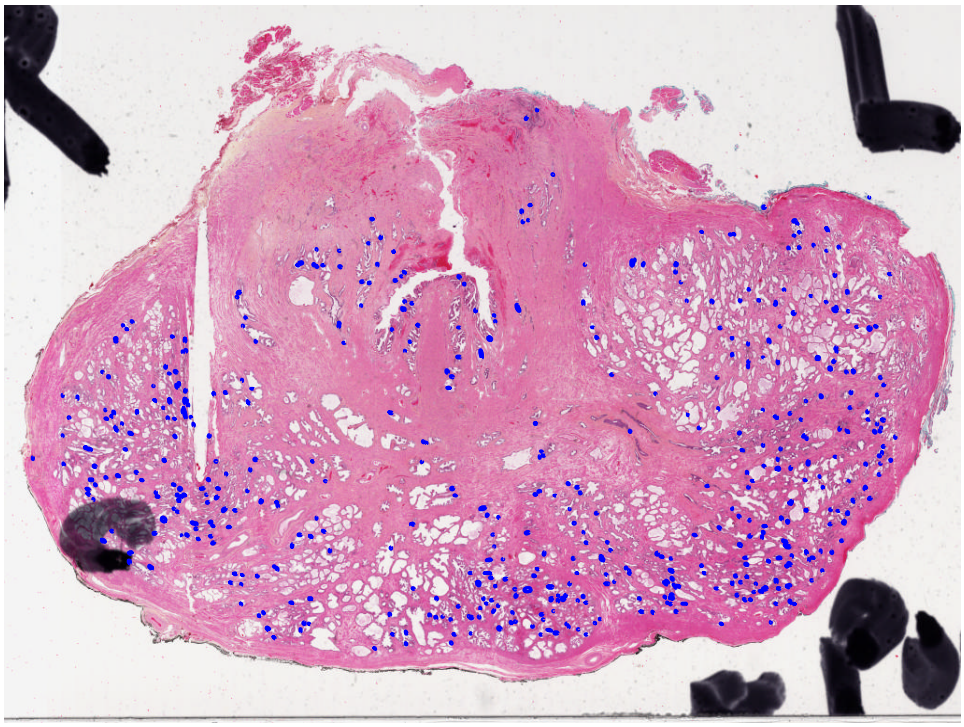


Figure 4.17: Illustration of the individual gland labeling on a sample whole mount image (Test image:5)

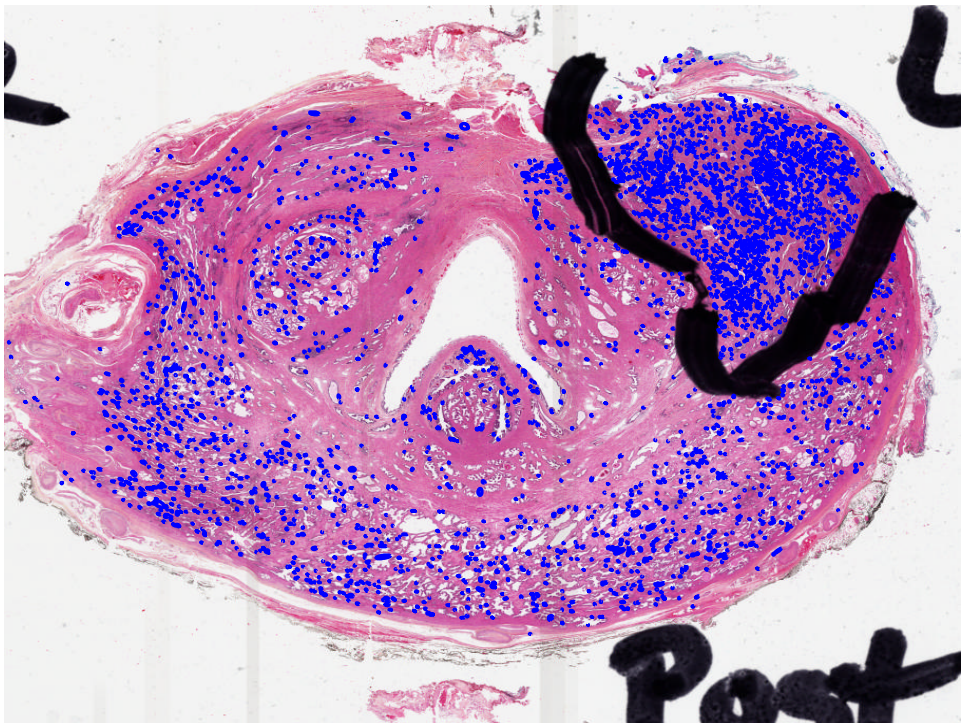


Figure 4.18: Illustration of the individual gland labeling on a sample whole mount image (Test image:6)

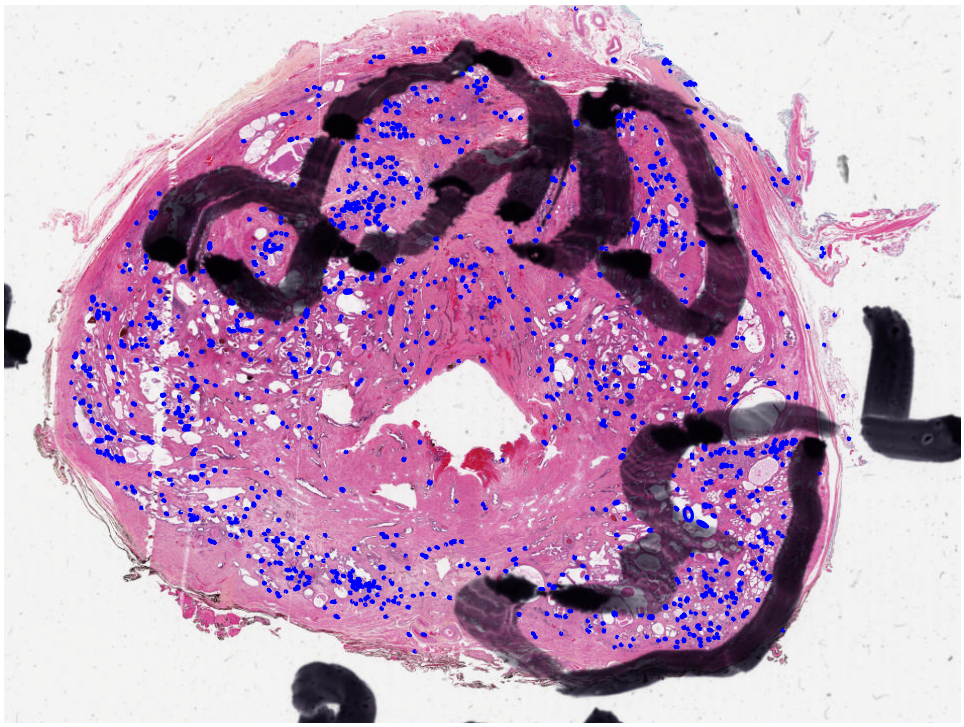


Figure 4.19: Illustration of the individual gland labeling on a sample whole mount image (Test image:7)

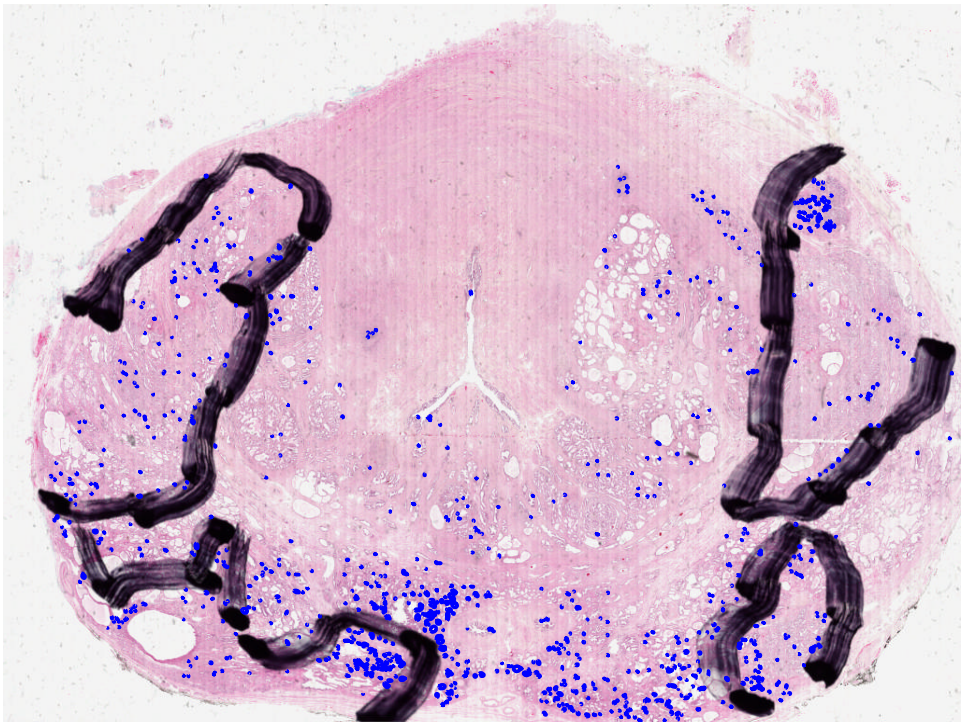


Figure 4.20: Illustration of the individual gland labeling on a sample whole mount image (Test image:8)

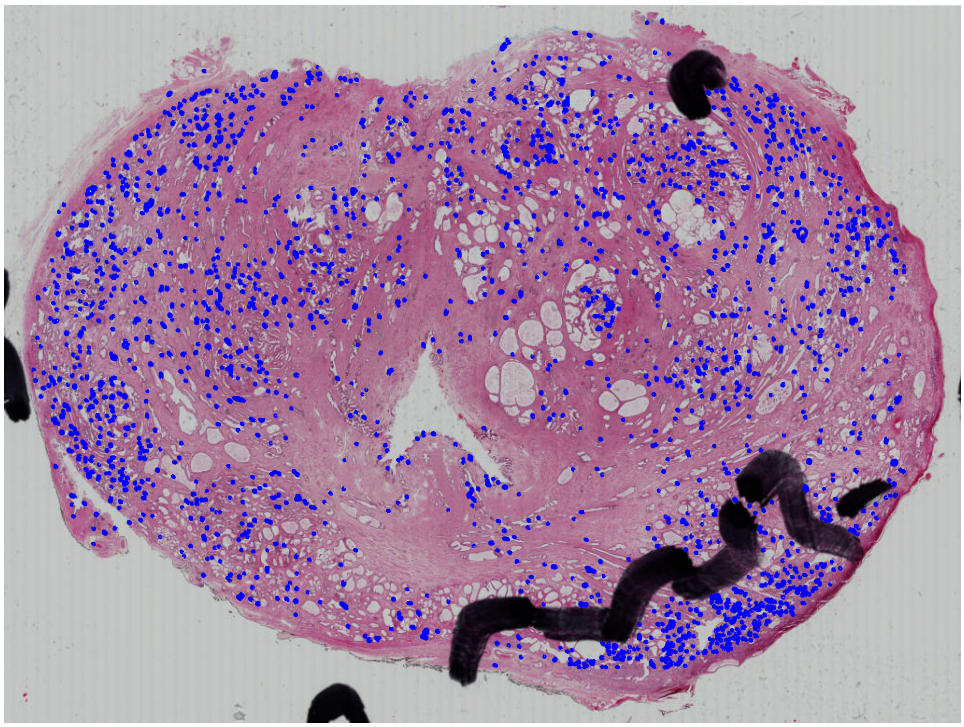


Figure 4.21: Illustration of the individual gland labeling on a sample whole mount image (Test image:9)

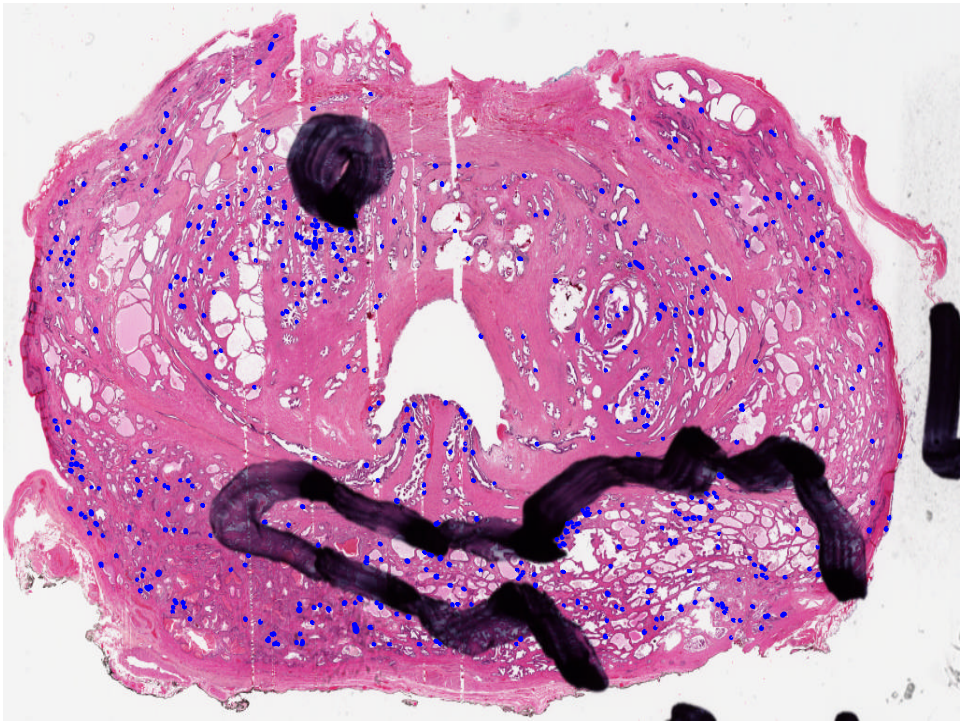


Figure 4.22: Illustration of the individual gland labeling on a sample whole mount image (Test image:10)

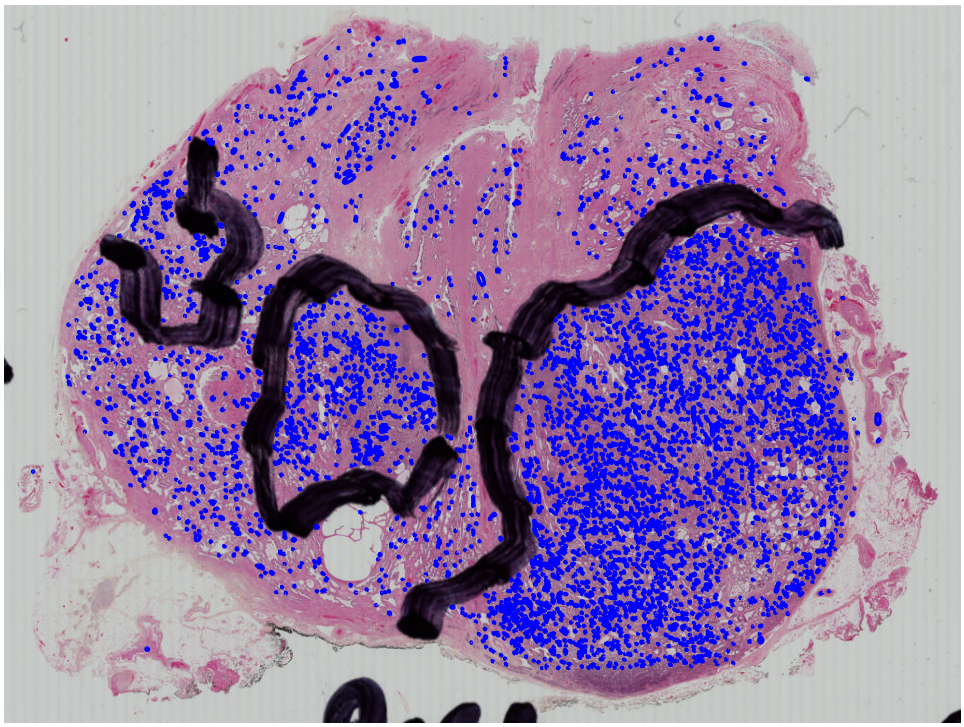


Figure 4.23: Illustration of the individual gland labeling on a sample whole mount image (Test image:11)

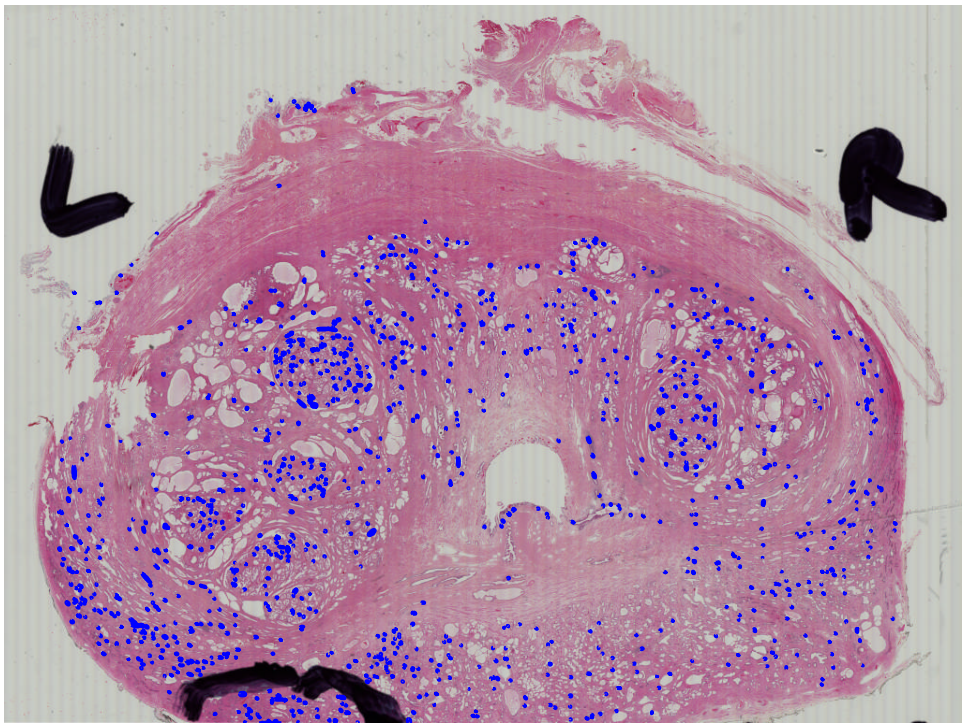


Figure 4.24: Illustration of the individual gland labeling on a sample whole mount image (Test image:12)

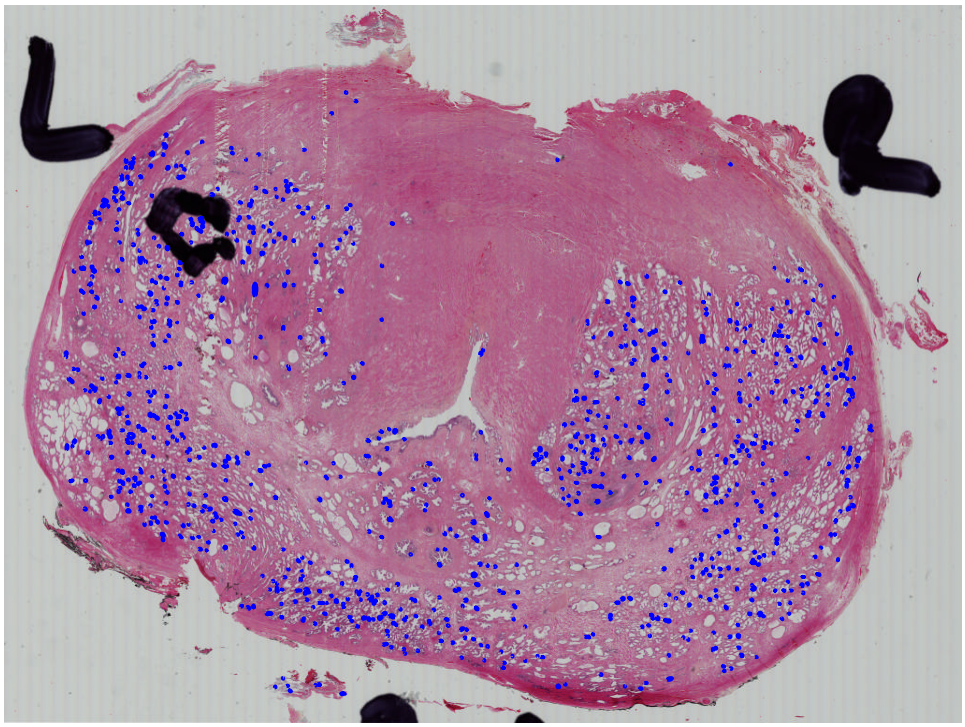


Figure 4.25: Illustration of the individual gland labeling on a sample whole mount image (Test image:13)

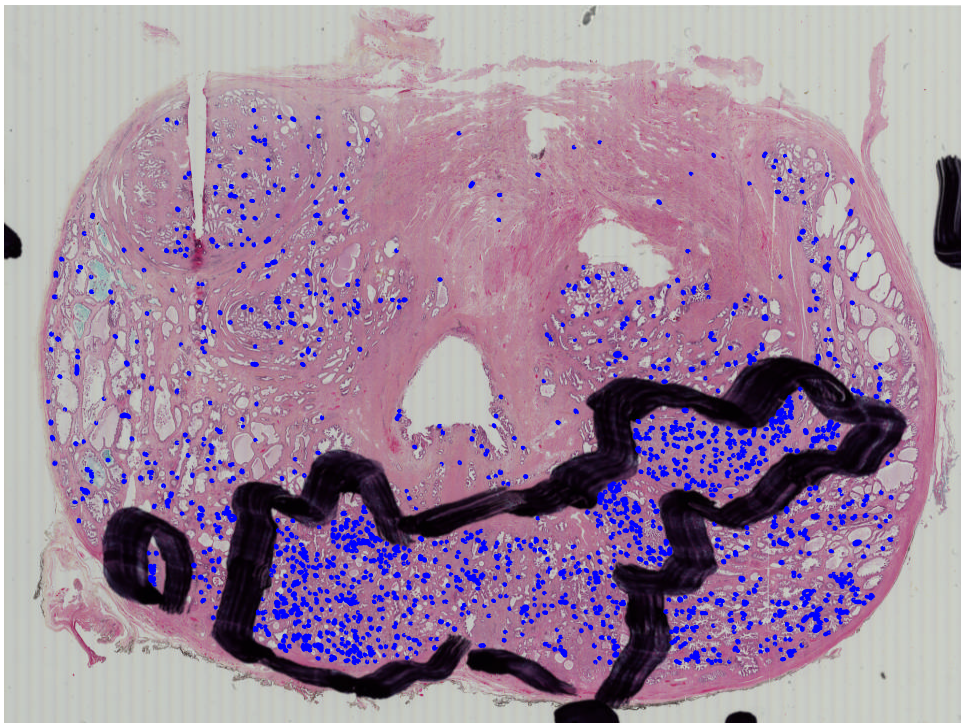


Figure 4.26: Illustration of the individual gland labeling on a sample whole mount image (Test image:14)

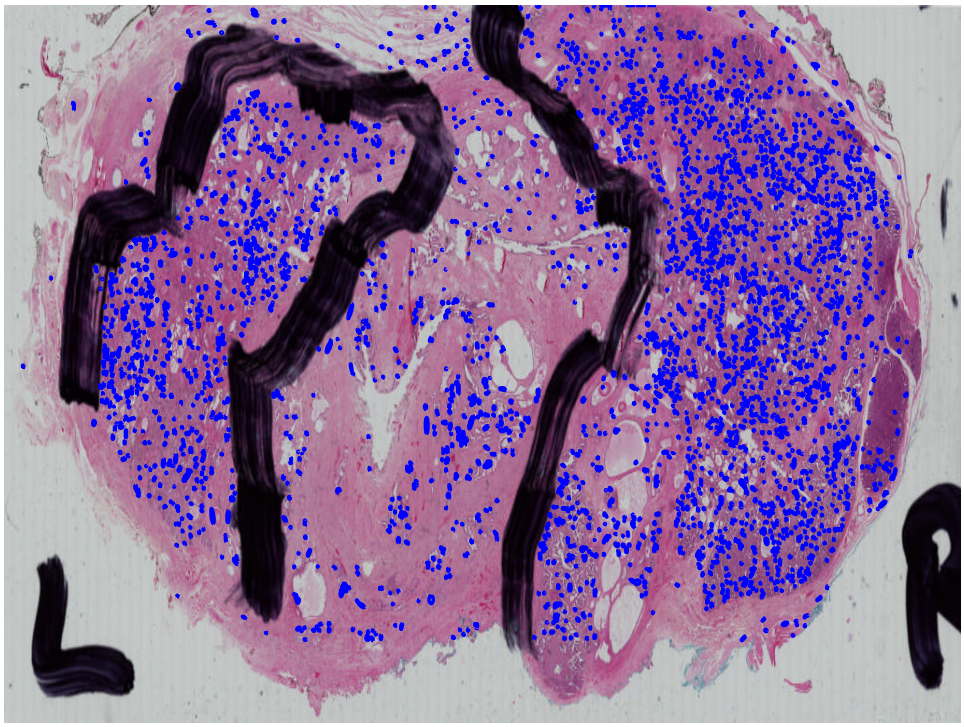


Figure 4.27: Illustration of the individual gland labeling on a sample whole mount image (Test image:15)

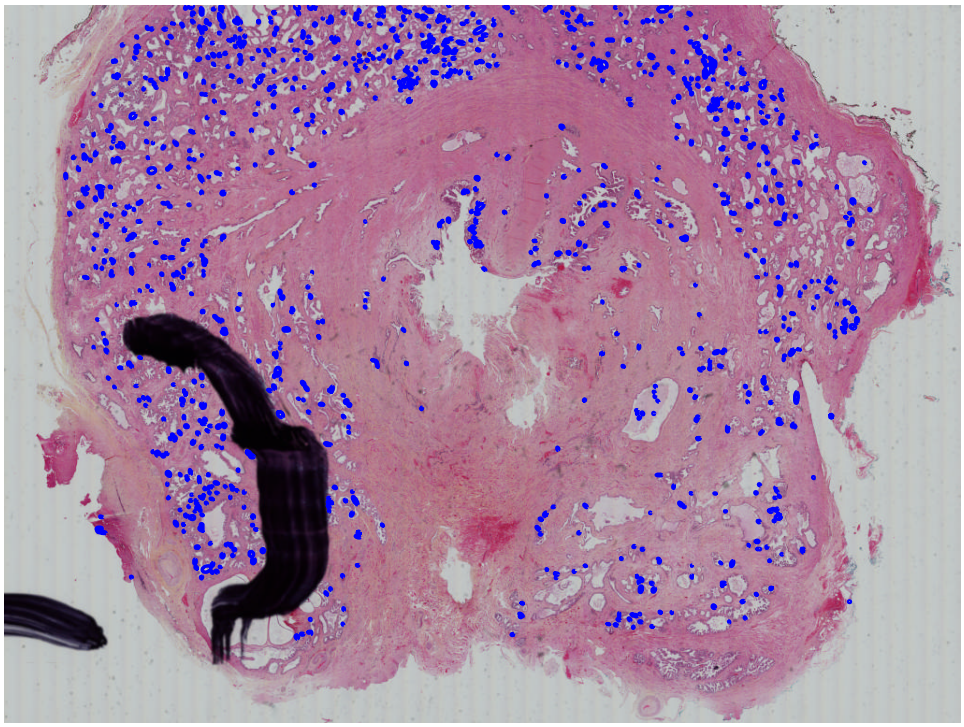


Figure 4.28: Illustration of the individual gland labeling on a sample whole mount image (Test image:16)

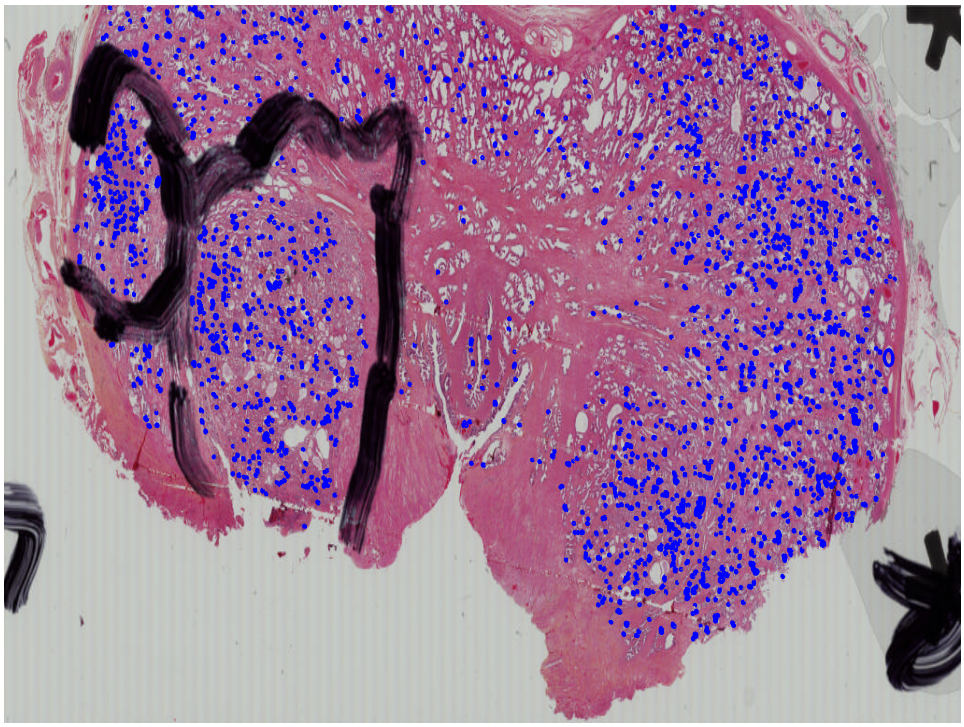


Figure 4.29: Illustration of the individual gland labeling on a sample whole mount image (Test image:17)

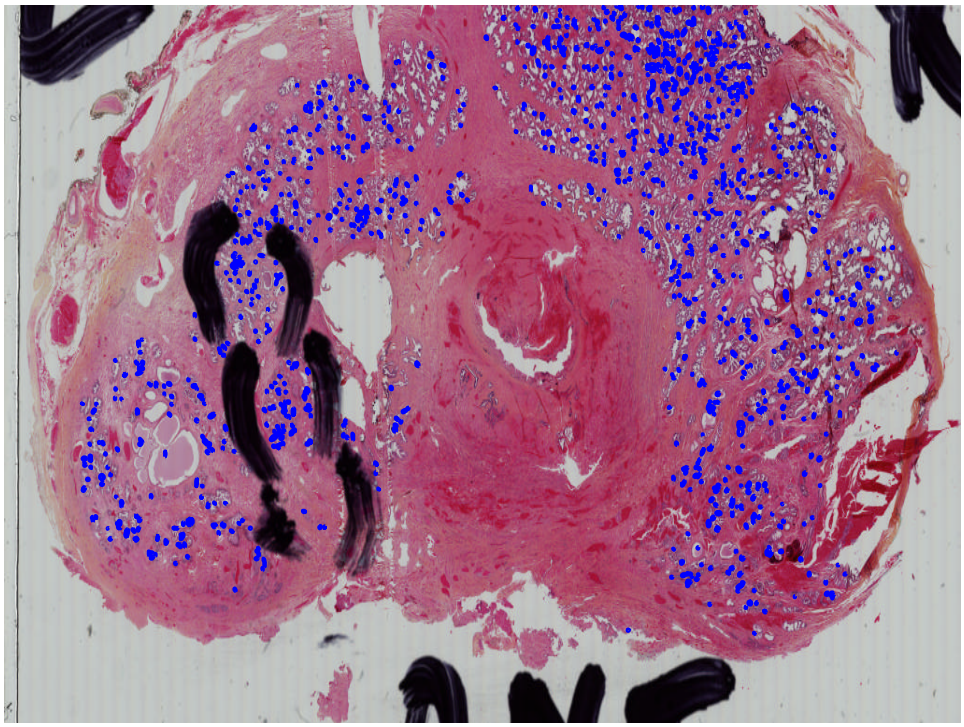


Figure 4.30: Illustration of the individual gland labeling on a sample whole mount image (Test image:18)

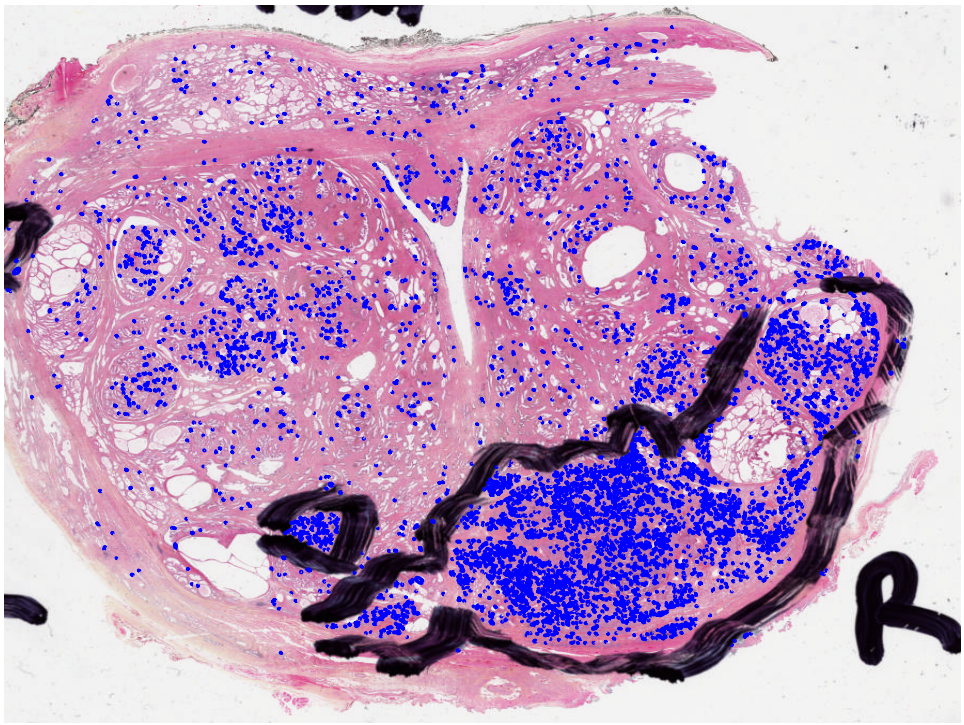


Figure 4.31: Illustration of the individual gland labeling on a sample whole mount image (Test image:19)

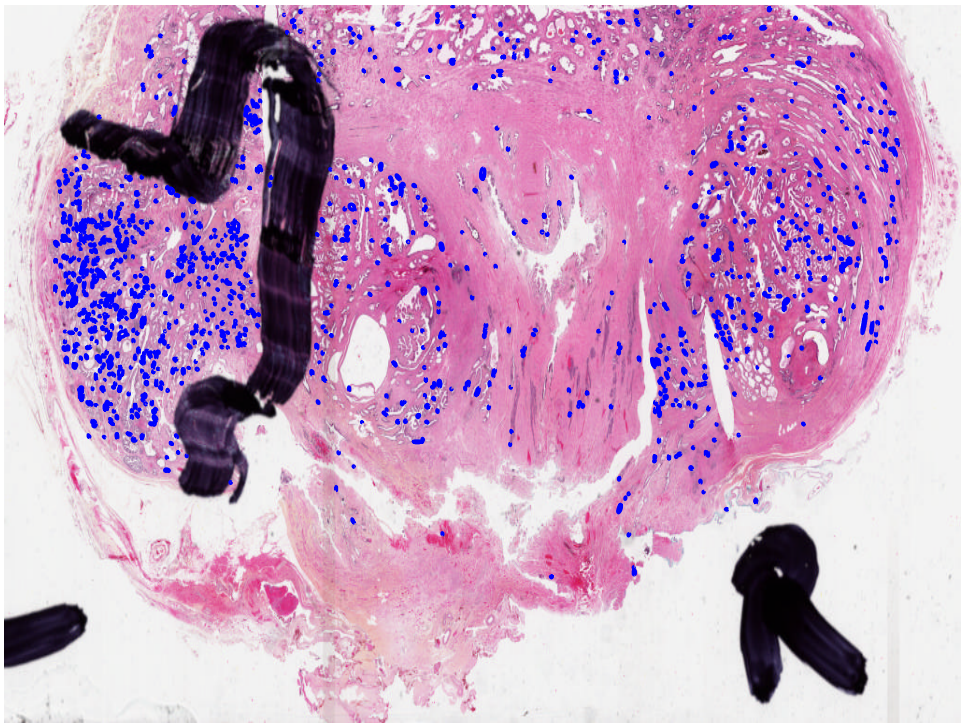


Figure 4.32: Illustration of the individual gland labeling on a sample whole mount image (Test image:20)

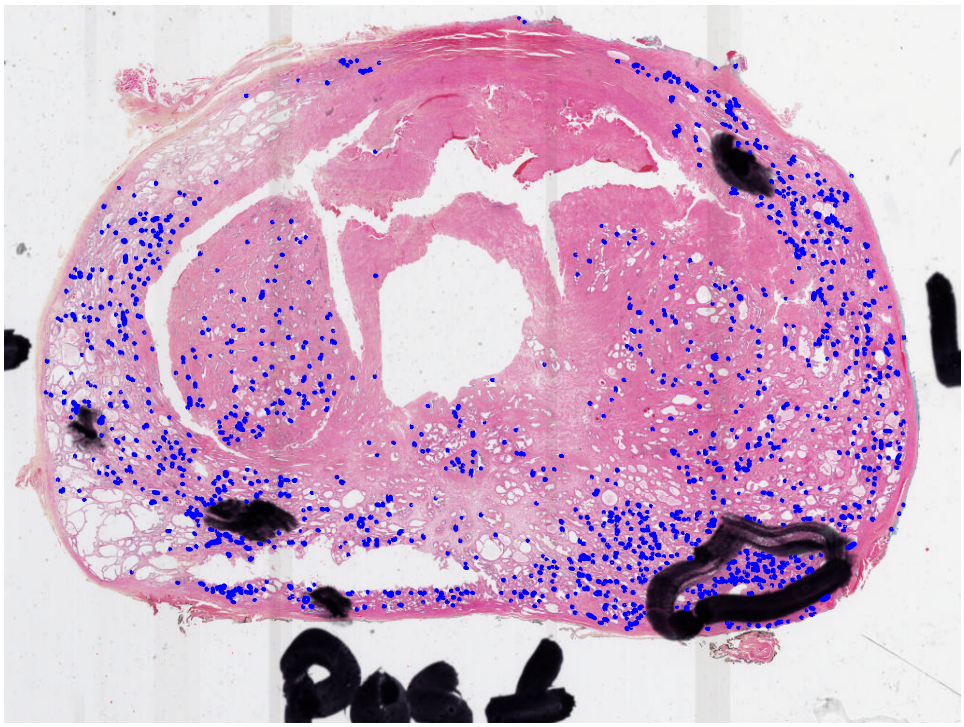


Figure 4.33: Illustration of the individual gland labeling on a sample whole mount image (Test image:21)

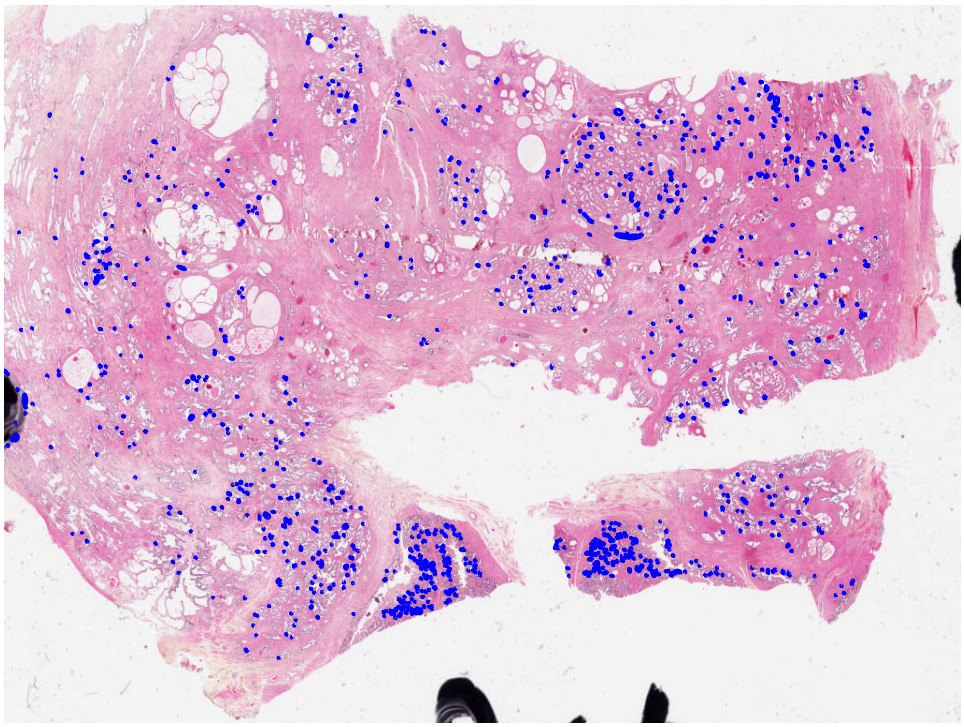


Figure 4.34: Illustration of the individual gland labeling on a sample whole mount image (Test image:22)

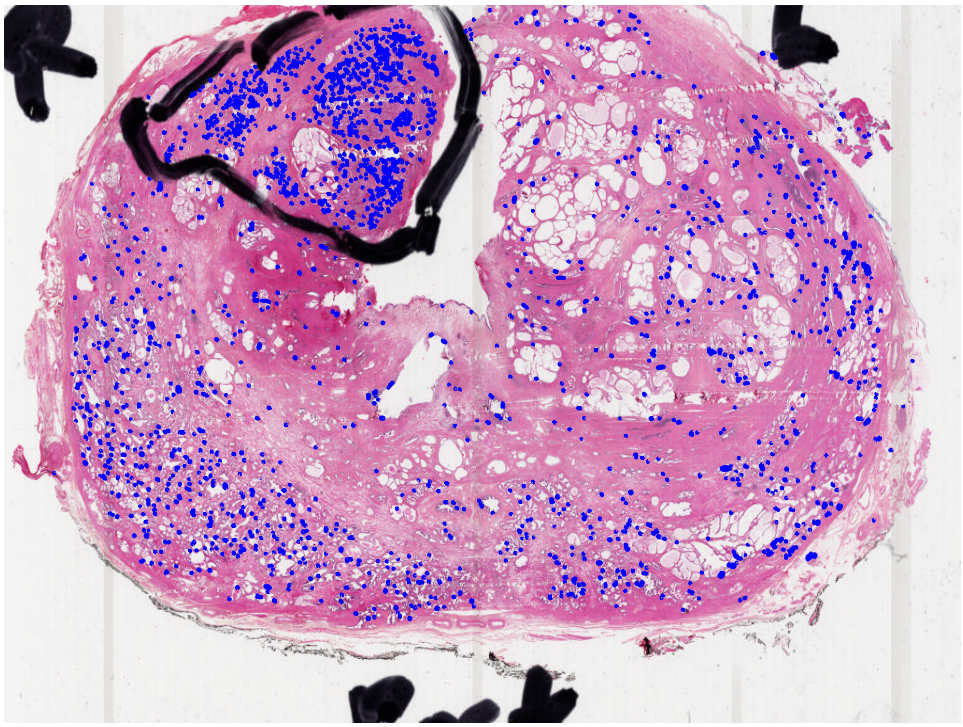


Figure 4.35: Illustration of the individual gland labeling on a sample whole mount image (Test image:23)

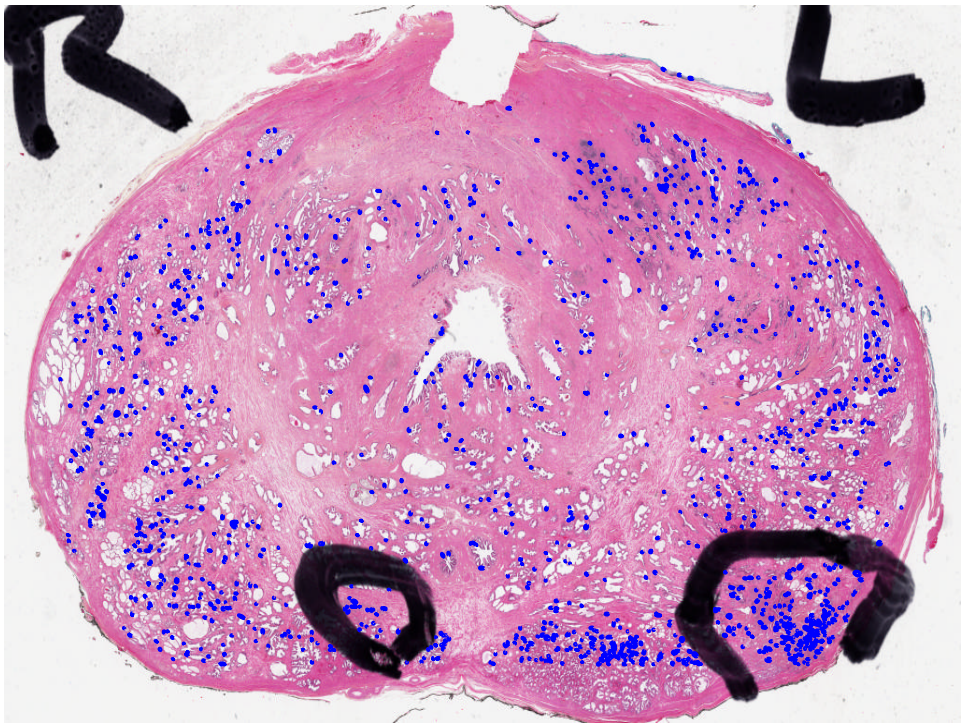


Figure 4.36: Illustration of the individual gland labeling on a sample whole mount image (Test image:24)

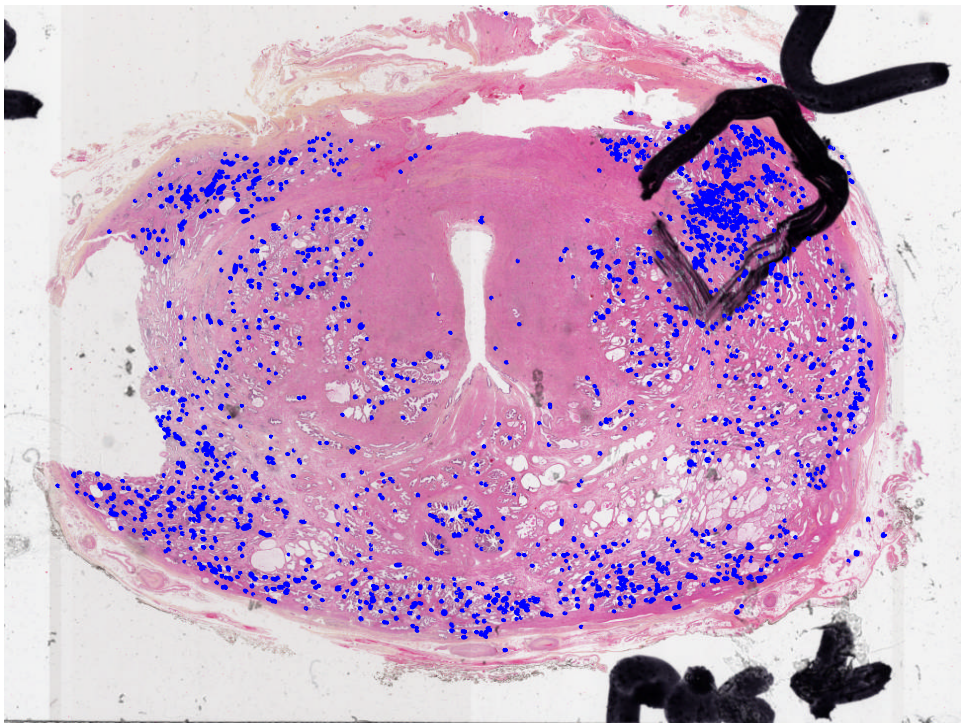


Figure 4.37: Illustration of the individual gland labeling on a sample whole mount image (Test image:25)

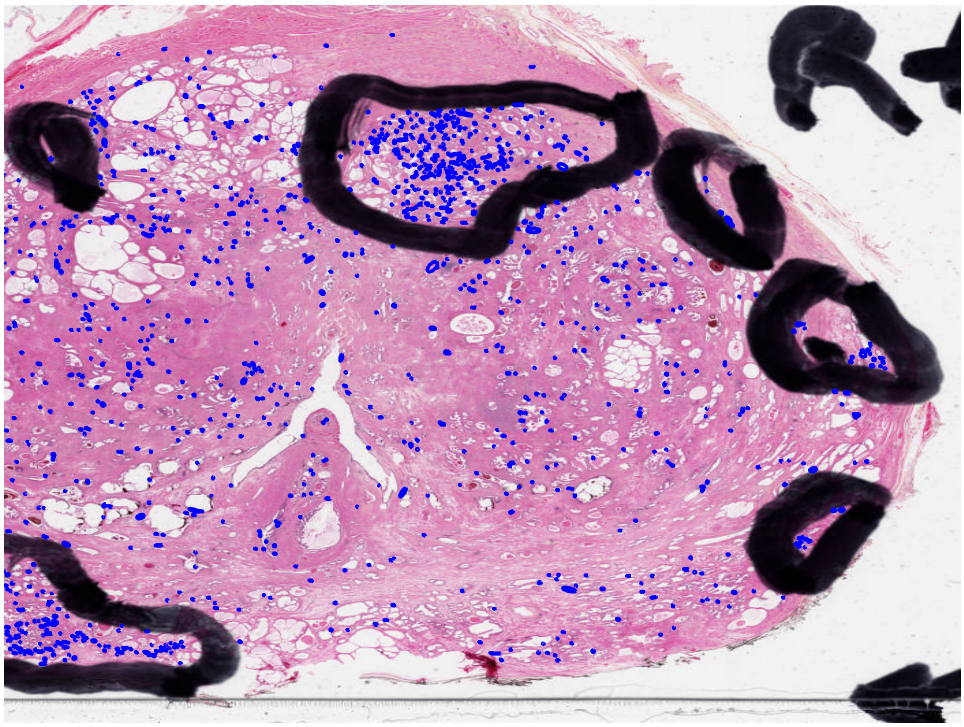


Figure 4.38: Illustration of the individual gland labeling on a sample whole mount image (Test image:26)

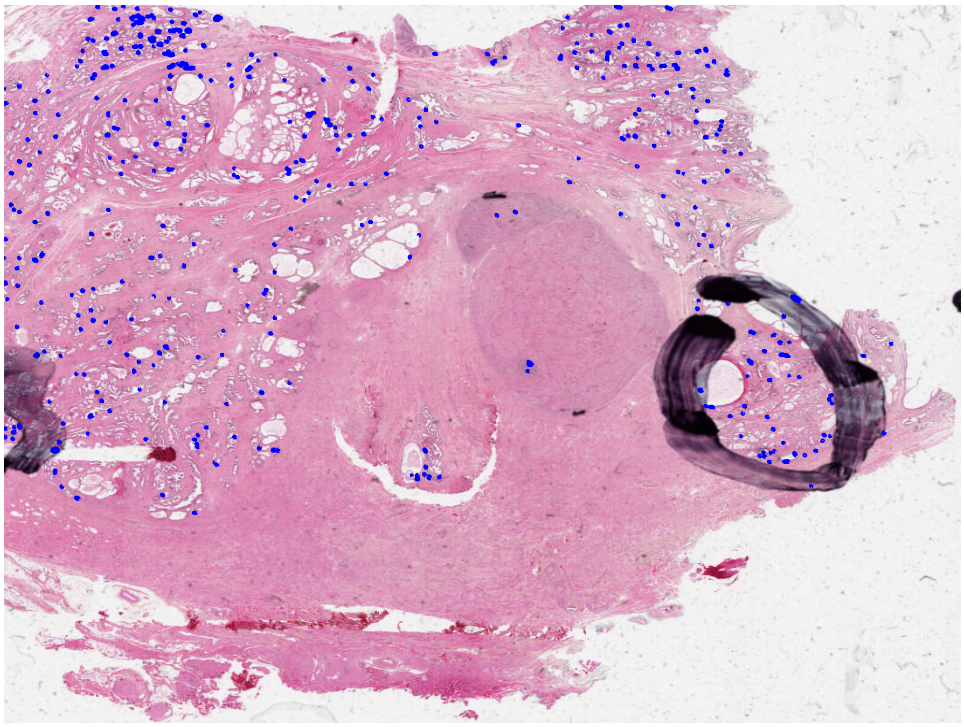


Figure 4.39: Illustration of the individual gland labeling on a sample whole mount image (Test image:27)

Chapter 5

Conclusions

5.1 Summary of Contributions

In this thesis, we have proposed a technique for automatic cancer annotation exploiting both regional and gland-specific properties. Combining these two aspects of histology, we have been able to achieve the best performance of automatic prostate cancer annotation. The major contributions of the works are:

1. Gland segmentation: Here, we have implemented an algorithm to automatically segment the gland units. We have employed linear discriminant analysis for labeling the tissue components associated with the gland units such as lumen, nuclei, epithelium, and stroma. The application of the linear discriminant analysis provides a faster classification compared to other similar technique of pixel labeling that uses voronoi tesselation based classification [30]. The segmentation of the gland units facilitates the extraction of features associated with it that are used to classify benign and malignant regions.

2. Nuclei Segmentation: We have also proposed a technique for nuclei segmentation using a marker controlled watershed algorithm. To the best of our knowledge this is the first technique to segment nuclei from images at $20\times$ magnification. The existing techniques of nuclei segmentation are performed on $40\times$ or higher magnification [1], [3], [4], [6], [18], [24], [38], [39], .

3. Quantification of Number of Nuclei Layers: We observed that the number of nuclei layers is one of the most important features for classification of individual gland unit. Here, We have proposed an innovative technique for the quantification of the number of nuclear layers. The number of layers are quantified by calculating the angular histogram of the nuclei surrounding each gland. The average of the number of bins in the histogram having multiple nuclei occurrences represents the Number of Nuclei Layers associated with the gland.

4. Individual Gland classification: Here, we have presented the first technique for labeling individual glands in prostate using gland-specific features only. We have proposed two novel features for labeling individual glands: number of nuclei layers and ratio of epithelial layer. The application of this individual gland-based technique will lead to a more sensitive cancer annotation. In addition to cancer annotation, these gland-specific properties might also be useful in identifying other prostate anomalies such as atrophy, benign prostatic hyperplasia, and prostatic intraepithelial neoplasia.

5. Regional feature set: We have also proposed a set of regional tissue features to detect cancerous regions in WM images. We observed that in the malignant tissue regions the morphology and architecture of the histology components i.e., i) lumen, ii) epithelial layer, and iii) nuclei are altered from their normal condition. Hence we have proposed a set of nine features associated with these tissue components to capture these morphology and architectural changes. Among the nine features six of them are newly proposed features. The high classification accuracy of the random forest classifier corroborates the effectiveness the of the proposed feature set.

6. Cancer annotation using regional and glandular features: Here we have presented a two stage algorithm for automatic annotation of prostate cancer from WM images. The multi-resolution technique presented here is a new approach for WM cancer annotation. The incorporation of the gland specific features with that of the regional features achieved the highest accuracy among the existing techniques for automated cancer annotation from WM

images.

5.2 Future Work

The proposed research has a number promising research directions in the field of computational pathology. Future work on this topic may include:

1. The features proposed here can be utilized in future to automatically grade the prostate cancers. The Gleason grading scheme is based on the tissue architecture of the prostate. Since the proposed features captures the architecture and morphology of the tissue, those features can be useful in developing a computational grading scheme.

2. The proposed individual gland classification technique often include gland from other prostate anomalies such as atrophy and prostatic intraepithelial neoplasia. Further improvement of the technique can be achieved to exclude non-malignant units. One possible improvement can be to identify the basal layer associated with each benign glands. Since malignant gland units do not posses a basal layer this criteria can be very effective to detect the malignant glands. This basal layer identification can be a post processing step after the proposed individual gland classification algorithm. The detected glands by the proposed technique can be further classified to detect malignant glands based on the presence of the basal layer.

3. Another exciting future research on this topic can be to investigate the correlation of the tissue features with the long term disease progression. Recurrence of cancer in prostatectomy patients often does not correlate with their gleason score. Therefore a better indicator is required to predict the patient outcome. Study of the histology features of prostate might be very useful in this area.

4. The proposed cancer annotation from whole mount images can also be used to find correlation of cancer in other imaging modalities, such as multiparametric ultrasound and multiparametric MRI.

Bibliography

[1] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, 57(4):841–852, 2010. → pages 25, 73

[2] F. Aurenhammer. Voronoi diagramsa survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991. → pages 17

[3] P. Bamford and B. Lovell. Unsupervised cell nucleus segmentation with active contours. *Signal Processing*, 71:203–213, 1998. → pages 7, 25, 73

[4] P. H. Bartels, T. Gahm, and D. Thompson. Automated microscopy in diagnostic histopathology: From image processing to automated reasoning. *International Journal of Imaging Systems and Technology*, 8(2):214–223, 1997. → pages 25, 73

[5] G. Begelrnan, E. Gur, E. Rivlin, M. Rudzsky, and Z. Zalevsky. Cell nuclei segmentation using fuzzy logic engine. In *International Conference on Image Processing (ICIP)*, volume 5, pages 2937 – 2940, Oct. 2004. → pages 7

[6] L. Boucheron, Z. Bi, N. Harvey, B. Manjunath, and D. Rimm. Utility of multispectral imaging for nuclear classification of routine clinical histopathology imagery. *BMC Cell Biology*, 8(Suppl 1):S8, 2007. → pages 25, 73

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. → pages 11, 14, 21

[8] P. cancer treatment guide. Prostate cancer information from the foundation of the prostate gland. *http://www.prostate-cancer.com*, 2013. → pages 3, 4

[9] 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*, 10:199–206, 1987. → pages 7

[10] B. Delaunay. Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, 7(793-800):1–2, 1934. → pages 29

[11] H. Edelsbrunner and E. P. Mücke. Three-dimensional alpha shapes. *ACM Transactions on Graphics (TOG)*, 13(1):43–72, 1994. → pages 28

[12] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *IEEE 12th International Conference on Computer Vision*, pages 670–677. IEEE, 2009. → pages 8

[13] M. Gao, P. Bridman, and S. Kumar. Computer aided prostate cancer diagnosis using image enhancement and jpeg2000. In *Proc. SPIE*, volume 5203, pages 323–334, 2003. → pages 8

[14] D. F. Gleason and M. Tannenbaum. The veteran's administration cooperative urologic research group: Histologic grading and clinical staging of prostatic carcinoma. *Urologic Pathology: The Prostate*, pages 171–198, 1977. → pages viii, 4, 6

[15] L. Gorelick, O. Veksler, M. Gaed, J. Gomez, M. Moussa, G. Bauman, A. Fenster, and A. Ward. Prostate histopathology: Learning tissue component histograms for cancer detection and classification. *IEEE Transactions on Medical Imaging*, 2013. → pages 8, 10

[16] P. W. Huang and C. H. Lee. Automatic classification for pathological prostate images based on fractal analysis. *IEEE Transactions on Medical Imaging*, 28(7):1037–1050, 2009. → pages 7, 10

[17] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman. Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61(2):69–90, 2011. → pages 1

[18] T. Jiang and F. Yang. An evolutionary tabu search for cell image segmentation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 32(5):675–678, 2002. → pages 25, 73

[19] J. K. Khouzani and S. H. Zadeh. Multiwavelet grading of prostate pathological images. In *Proc. SPIE*, volume 4628, pages 1130–1138, 2002. → pages 7, 10

[20] W. J. Krzanowski and W. Krzanowski. *Principles of multivariate analysis*. Clarendon, 2000. → pages 17

[21] V. Kumar, A. K. Abbas, N. Fausto, and J. C. Aster. *Robbins & Cotran Pathologic Basis of Disease*. Elsevier Health Sciences, 2009. → pages 3, 35

[22] E. Lee, Y. Pan, and P. Chu. An algorithm for region filling using two-dimensional grammars. *International Journal of Intelligent Systems*, 2 (3):255–263, 1987. → pages 17

[23] K.-H. Leissner and L.-E. Tisell. The weight of the human prostate. *Scandinavian Journal of Urology and Nephrology*, 13(2):137–142, 1979. → pages 3

[24] G. Li, T. Liu, J. Nie, L. Guo, J. Chen, J. Zhu, W. Xia, A. Mara, S. Holley, and S. Wong. Segmentation of touching cell nuclei using gradient flow tracking. *Journal of Microscopy*, 231(1):47–58, 2008. → pages 25, 73

[25] F. Meyer. Topographic distance and watershed lines. *Signal processing*, 38 (1):113–125, 1994. → pages 11

[26] J. P. Monaco, M. Feldman, J. Tomaszewski, and A. Madabhushi. Detection of prostate cancer from whole-mount histology images using markov random fields. In *Proc. of 2nd Workshop on Micro. Image Anal. with Applications in Biology*, 2008. → pages 8, 10, 37

[27] C. A. Moskaluk, P. H. Duray, K. H. Cowan, M. Linehan, and M. J. Merino. Immunohistochemical expression of π-class glutathione s-transferase is down-regulated in adenocarcinoma of the prostate. *Cancer*, 79(8): 1595–1599, 1997. → pages 36

[28] 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 *Proc. of 2nd Workshop on Micro. Image Anal. with Applications in Biology*, 2007. → pages 8, 10

[29] K. Nguyen, A. K. Jain, and B. Sabata. Prostate cancer detection: Fusion of cytological and textural features. *Journal of Pathology Informatics*, 2, 2011. → pages 37

[30] K. Nguyen, B. Sabata, and A. K. Jain. Prostate cancer grading: Gland segmentation and structural features. *Pattern Recognition Letters*, 33:951 – 961, 2011. → pages 15, 17, 73

[31] K. Nguyen, A. Sarkar, and A. K. Jain. Structure and context in prostatic gland segmentation and classification. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 115–123. Springer, 2012. → pages 8, 9, 10

[32] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975. → pages 25

[33] R. E. Schapire. A brief introduction to boosting. In *IJCAI*, volume 99, pages 1401–1406, 1999. → pages 8

[34] 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*, 17(3):204–218, 1995. → pages 7

[35] K. Suzuki, I. Horiba, and N. Sugie. Linear-time connected-component labeling based on sequential local operations. *Computer Vision and Image Understanding*, 89(1):1–23, 2003. → pages 17

[36] 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*, 26(4): 518–523, 2007. → pages 7, 10

[37] O. Veksler, Y. Boykov, and P. Mehrani. Superpixels and supervoxels in an energy optimization framework. In *Computer Vision–ECCV 2010*, pages 211–224. Springer, 2010. → pages 8

[38] S. Wienert, D. Heim, K. Saeger, A. Stenzinger, M. Beil, P. Hufnagl, M. Dietel, C. Denkert, and F. Klauschen. Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach. *Scientific Reports*, 2, 2012. → pages 25, 73

[39] H.-S. Wu, J. Barba, and J. Gil. A parametric fitting algorithm for segmentation of cell images. *IEEE Transactions on Biomedical Engineering*, 45(3):400–407, 1998. → pages 25, 73

[40] Y. Zhu, S. Williams, and R. Zwiggelaar. Computer technology in detection and staging of prostate carcinoma: A review. *Med. image Anal*, 10:178–199, 2006. → pages 1