



University of
Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study program/specialization:
Information Technology -
Automation and Signal Processing

Spring semester, 2016

Open / Confidential

Author: Jarle Urdal

.....
Jarle Urdal
.....
(signature author)

Instructor: Professor Kjersti Engan

Supervisor: Professor Kjersti Engan

Title of Master's Thesis:

Image processing and classification of urothelial carcinoma using tissue sample images

Norwegian title:

Bildebehandling og klassifisering av uroteliale karsinomer ved bruk av bilder fra vevsprøver

ECTS: 30

Subject headings:
Urothelial Carcinoma, Local Binary Pattern,
Local Variance, Chi-Squared, RUSBoost

Pages: 72
+ attachments/other: 15 + embedded file

Stavanger, 15th of June/2016
Date/year



University of
Stavanger

Image processing and classification of
urothelial carcinoma using tissue sample
images

Jarle Urdal

June 2016

MASTER'S THESIS

Faculty of Science and Technology
Department of Electrical Engineering and Computer Science
University of Stavanger

Supervisor: Professor Kjersti Engan

Abstract

Bladder cancer is the 6th most common cancer in the world, with 429.000 new incidents reported in 2012, of which urothelial carcinoma is the most common. Prognostic value of the current grading systems are low, with only a significant difference for progression between lowest and highest grade on one system. As a consequence of a recurrence rates of 50 – 70% and progression to a higher stage in 10 – 30% of patients, extensive follow-ups are given regularly over several years after first diagnosis.

The objective of this thesis is to determine if a local texture analysis can be used as an aid in the prediction of recurrence and progression on patients originally diagnosed with TaT1 urothelial carcinoma. An analysis is done using microscopic tissue samples from 42 patients. Textures are described using local binary pattern and local variance, and features are computed as the chi-squared of the descriptor histograms and predefined models for each prognoses.

Local binary pattern achieves approximately 80% correct identification of patients with recurrence, while identification of patients without recurrence are approximately 50%. Suggesting this descriptor can be used to identify patients with recurrence. Prediction using local variance achieve better than random-guessing using a linear normalization of images, but overall results are low. Prognostic value for progression using both descriptors are low, with no clear identification of patients with progression found.

Preface

This thesis was written at the Department of Electrical Engineering and Computer Science, University of Stavanger. I would like to thank my supervisor Prof. Kjersti Engan for her advice and valuable feedback. I would also like to thank Emiel Janssen and Vebjørn Kvikstad at Stavanger University Hospital for the data material and their medical insight. Finally, I would like to thank Anne-Marthe Hovden for her advice and support throughout the semester.

Contents

1	Introduction	1
1.1	Thesis outline	2
2	Background	3
2.1	Urothelial carcinoma	3
2.1.1	Epidemiology	3
2.1.2	Stage	3
2.1.3	Current grading systems	4
2.2	Material	6
2.3	Histogram equalizing	7
2.4	Image thresholding	8
2.5	Local Binary Pattern	9
2.5.1	Principle	9
2.5.2	Rotation invariance	10
2.5.3	Uniform	11
2.5.4	Local variance	12
2.5.5	Multiresolution analysis	13
2.6	Classification	14
2.6.1	Class imbalance	14
2.6.2	A hybrid approach	15
2.7	Similarity measures	16
2.7.1	Sum of squared error	16
2.7.2	Chi-squared	17
2.8	Performance Measures	17
3	Implementation	19
3.1	Preprocessing	19
3.1.1	Grayscale expansion	20
3.1.2	Histogram equalizing	21
3.2	Masking	21
3.2.1	Identification of regions with tissue	21
3.2.2	Identification of regions with connective tissue	22
3.3	Image descriptors	24
3.3.1	Local binary pattern	24
3.3.2	Local variance	24
3.4	Feature extraction	24
3.4.1	Model histograms	24
3.4.2	Feature vectors	25
3.5	Classification	27
3.6	Proposed system	28

4	Experiments and results	31
4.1	Masking	31
4.1.1	Identification of regions with tissue	31
4.1.2	Identification of regions with connective tissue	32
4.2	Experiments	33
4.2.1	Pre-experiment: Choosing system parameters	34
4.2.2	Experiment 1: Recurrence	37
4.2.3	Experiment 2: Progression	38
4.2.4	Experiment 3: Cost functions	40
4.2.5	Experiment 4: Masking - removing suburothelial connective tissue	42
4.2.6	Experiment 5: Preprocessing	44
5	Discussion	47
5.1	Data set	47
5.2	System parameters	47
5.3	Feature extraction	48
5.4	Preprocessing	48
5.5	Masking	49
5.6	Other texture descriptors	50
6	Conclusion and future work	51
6.1	Future work	51
A	Matlab code	57
B	Data set	59
C	Patient follow-up	71

List of Figures

1.1	Urothelial carcinoma, example images	1
1.2	Urothelial carcinoma, example images	2
2.1	Bladder cancer incidence	4
2.2	Extent of primary bladder cancer	5
2.3	Comparison of WHO73 and WHO04 grading system	5
2.4	Histogram equalizing, example image	7
2.5	Histogram equalizing, transformed image	8
2.6	Example, Otsu's method	9
2.7	LBP: principle	9
2.8	LBP: weights	10
2.9	LBP example: texture, source: MATLAB test images	10
2.10	LBP example: histograms	11
2.11	LBP: Rotation invariance principle	11
2.12	LBP: Uniform rotation invariant descriptor bins	12
2.13	Exponential multi-resolution LBP	13
3.1	System overview	19
3.2	Variance in image colours	20
3.3	Graphical scale	20
3.4	[0, 1] normalized grayscale images	20
3.5	Histogram equalized grayscale images	21
3.6	Patient ID 147, cell and cell mask	22
3.7	Variations in connective tissue	23
3.8	LBP model histograms	25
3.9	Example: New unknown texture with its corresponding LBP histogram	26
3.10	Example: Chi-squared for the texture	26
3.11	Proposed system	29
4.1	Identification of tissue: Patient ID 28	31
4.2	Identification of tissue: Patient ID 98	32
4.3	Identification of connective tissue: Patient ID 28	32
4.4	Identification of connective tissue: Patient ID 98	33
4.5	Experiment 1: Recurrence, multiple descriptors	37
4.6	Experiment 1: Recurrence, LBP	38
4.7	Experiment 1: Recurrence, local variance	38
4.8	Experiment 2: Any progression, multiple descriptors	39
4.9	Experiment 2: Muscle invasive progression, multiple descriptors	40
4.10	Experiment 3: Recurrence, multiple descriptors using cost function	41
4.11	Experiment 3: Recurrence, single descriptors using cost function	41
4.12	Experiment 3: Recurrence, multiple descriptors using transposed cost function	42

4.13	Experiment 4: Removing connective tissue (case ID: 28)	42
4.14	Experiment 4: Recurrence, multiple descriptors using alternative masking	43
4.15	Experiment 4: Recurrence, single descriptors using alternative masking	43
4.16	Experiment 4: Progression, multiple descriptors using alternative masking	44
4.17	Experiment 5: Preprocessing alternatives (case ID: 28)	44
4.18	Experiment 5: Recurrence, multiple descriptors using alternative preprocessing	45
4.19	Experiment 5: Recurrence, single descriptors using alternative preprocessing	45
4.20	Experiment 5: Progression, multiple descriptors using alternative normalization	46
5.1	Block wise LBP, comparison	49

List of Tables

2.1	Prognostic follow-up information	6
2.2	Follow-up regime	6
2.3	Confusion matrix for a 2-class problem	17
4.1	Identification of any tissue	32
4.2	Identification of connective tissue	33
4.3	Pre-experiment: System performance using a single scale	34
4.4	Pre-experiment: System performance using two scales	35
4.5	Pre-experiment: System performance using multiple scales	35
4.6	Classifier parameters	36

List of abbreviations

ASR	Age-standardized rate
HES	Haematoxylin-Erythrosine-Saffron
LBP	Local Binary Pattern
MSE	Mean Squared Error
SSE	Sum of Squared Errors
WHO04	2004 World Health Organization classification of papillary urothelial neoplasms
WHO73	1973 World Health Organization classification of papillary urothelial neoplasms

1. Introduction

Bladder cancer is the 6th most common cancer in the world, with 429.000 new incidents reported in 2012 [1]. While several types of bladder cancer exists, urothelial carcinoma is the most common [2]. An extensive follow-up is given to patients diagnosed with urothelial carcinoma, as recurrences are found in 50 – 70% of patients, and progression to a higher stage in 10 – 30% of patients [2]. Grading of the tumours are currently done by analysing multiple features from extracted cell-tissue by a trained pathologist, where each feature are graded using the WHO73¹ or WHO04² guidelines. As these gradings are done manually, interobserver variability has been a concern and has been reported as a challenge for both grading systems [2]. Prognostic values of the grading system are also low, with no significant differences found between grades using WHO73 for recurrence- and progression free survival. Using WHO04, a significant difference was only found for progression free survival between the lowest and highest grade [3].

By introducing image processing, interobserver variability can be improved as distinct cut-off areas between grades are defined. The use of of image processing also introduces the opportunity of implementing new features, which can be used to describe texture and patterns not possible to detect by visual observation. Analysis based on local textures are used in multiple disciplines ranging from identification of liver disease [4] to classification of breast cancer in mammographies [5] with promising results.

A local texture analysis will be done using a data set of scanned tissue from 42 patients originally diagnosed with non-invasive urothelial carcinoma. The analysis will be done through the use of Local Binary Pattern (LBP) and local variance texture descriptors, with the goal of identifying prognostic information from tissue samples. Apparent intersample variations of the extracted tissue using a 40x magnification are observed, illustrated in figure 1.1 and 1.2. Of the samples shown, case ID 18 and 31 had progression, while ID 28 had recurrence but no progression and ID 182 did not have recurrence.

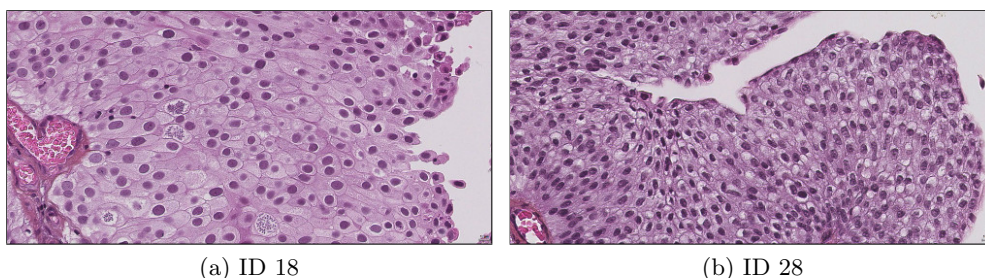


Figure 1.1: Urothelial caricnoma, example images

¹1973 World Health Organization classification of papillary urothelial neoplasms

²2004 World Health Organization classification of papillary urothelial neoplasms

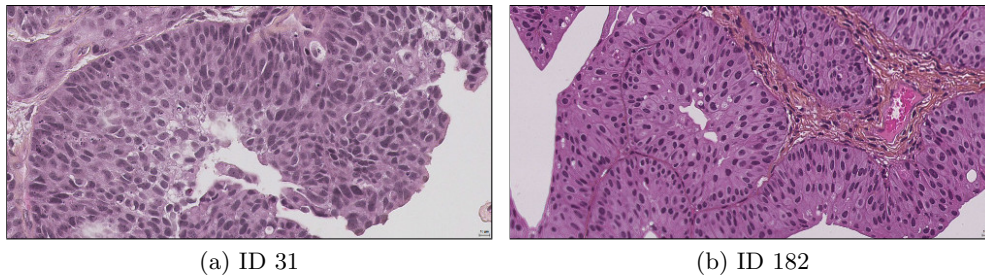


Figure 1.2: Urothelial carcinoma, example images

The objective of this thesis is to determine if a local texture analysis on microscopic tissue sample images can be used as an aid for predicting recurrence and progression of patients with urothelial carcinoma.

1.1 Thesis outline

Chapter 2 - Background:

This chapter contains a brief presentation of urothelial carcinoma, material used and concepts the designed system is based upon.

Chapter 3 - Implementation:

The proposed system is presented, where each module in addition are presented in closer detail.

Chapter 4 - Experiments and results:

Outline of preliminary experiments and the system parameters chosen are presented, experiments for prediction of recurrence and progression and the corresponding results are then presented.

Chapter 5 - Discussion:

The chapter contains a discussion of material, concepts and the results obtained throughout this thesis.

Chapter 6 - Conclusion and future work:

The chapter contains a conclusion of the prognostic value of the analysed methods, possible improvements of the devised system and directions for future research.

Appendix A - Matlab code:

A list of devised functions and their behaviour are presented. All code described are found in the embedded file *matlab.7z*.

Appendix B - Data set:

Used images in reduced resolution.

Appendix C - Patient follow-up:

Prognostic follow-up information and stage on first presentation of the carcinoma.

2. Background

This chapter contains the background information this thesis is based upon. Urothelial carcinoma and the data set are first presented. Image normalization and thresholding are then presented, followed by local binary pattern. Finally, the similarity and performance measures used throughout this thesis are described.

2.1 Urothelial carcinoma

Urothelial carcinoma, previously known as transitional cell carcinoma, arises in the urinary bladder walls. The bladder wall consist of several layers, urothelium, suburothelial connective tissue and muscles, where the carcinoma arises in the outer urothelial layer. Tumours can be found anywhere inside the bladder, but are most commonly found on the lateral walls [6]. The tumours can form papillary propulsions ¹, solid nodules or grow into the bladder wall [2].

2.1.1 Epidemiology

Bladder cancer is the 6th most common cancer in the world with 429 000 incidents and a mortality rate of 165 000 reported in 2012 [1]. Large differences based on both regions and sexes exist, with incident rates for men over three times higher than for females. In more developed regions men have an age-standardized rate (ASR) of 16.9 incidents per 100 000 inhabitants, while an ASR of 3.7 per 100 000 inhabitants are found for females. In less developed regions men have an ASR of 5.3 and females 1.5. These regional differences are illustrated in figure 2.1 where Europe and North America have a high incident rate, while parts of Africa and Asia have a lower incident rate.

In Norway, 1378 incidents and a mortality rate of 334 were reported in 2012 [1]. Including 1021 men, resulting in an ASR of 21.6, which is well above the rate for the more developed regions. With the reported number of incidents, bladder cancer ranks 4th for the most common cancer for men and 9th most common cancer for females.

2.1.2 Stage

The Tumor Node Metastasis classification system (TNM) are used to describe a patients stage of cancer. Where the T-parameter describes the original tumor size, and possible invasion into nearby tissue. Papillary tumours confined to the urothelial layer are described as pTa, where the prefix *p* denotes stage is given by pathological examination. A papillary tumor with invasion

¹projection of tissue into the bladder

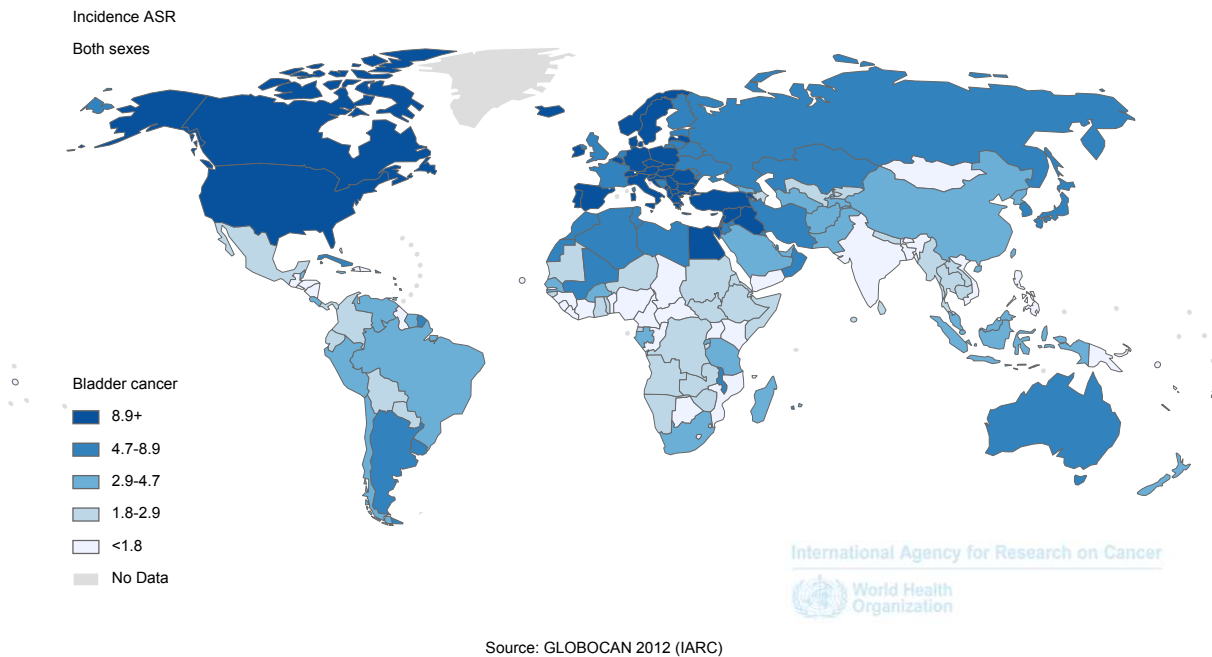


Figure 2.1: Bladder cancer incidence ASR, [1]

into the suburothelial connective tissue is described as pT1. Both of these classifications are often grouped together and described as non-muscle-invasive tumors. Stage pTa and pT1 are diagnosed in approximately 66 – 75% of new patients on first presentation of the carcinoma [7]. A third class of non-muscle-invasive tumor is carcinoma in situ, known as Tis. Tis is however confined to the urothelial layer with no papillary expansions into the urinary bladder.

Deeper stages are grouped together and labelled muscle-invasive, pT2 with invasion to the muscle, pT3 with invasion into fat and connective tissue surrounding the urinary bladder and pT4 with invasion to surrounding organs. The different stages can also be divided into substages based on invasion. A visual representation of invasion in the different stages are illustrated in figure 2.2.

2.1.3 Current grading systems

Two grading systems defined by the World Health Organization are currently in use, WHO73 and WHO04 [2]. Both systems are based on a visual analysis of known features by a pathologist to identify the grade.

The WHO73 system grades tumours based on cellular differentiation into three distinct grades. Grade 1 is given to tumours with the least cellular differentiation, and grade 3 to tumours with severe degrees of cellular differentiation. While grade 2 is given to tumours in between [9]. The system does however not describe clear cut-off regions between grades. Due to this unclear cut-off, interobserver variability has been a concern. And variations of grade 2 incidents ranging from 13% to 69% has been reported [10].

The WHO04 system was originally presented at the International Society of Urological Pathology 1998 and included in the WHO Blue Book [11] in 2004. The system was developed to give a clearer criteria for each grade to improve reproducibility. Grading is done using the terms *papillary urothelial neoplasm of low malignant potential (PUNLMP)*, *low grade* and *high grade*.

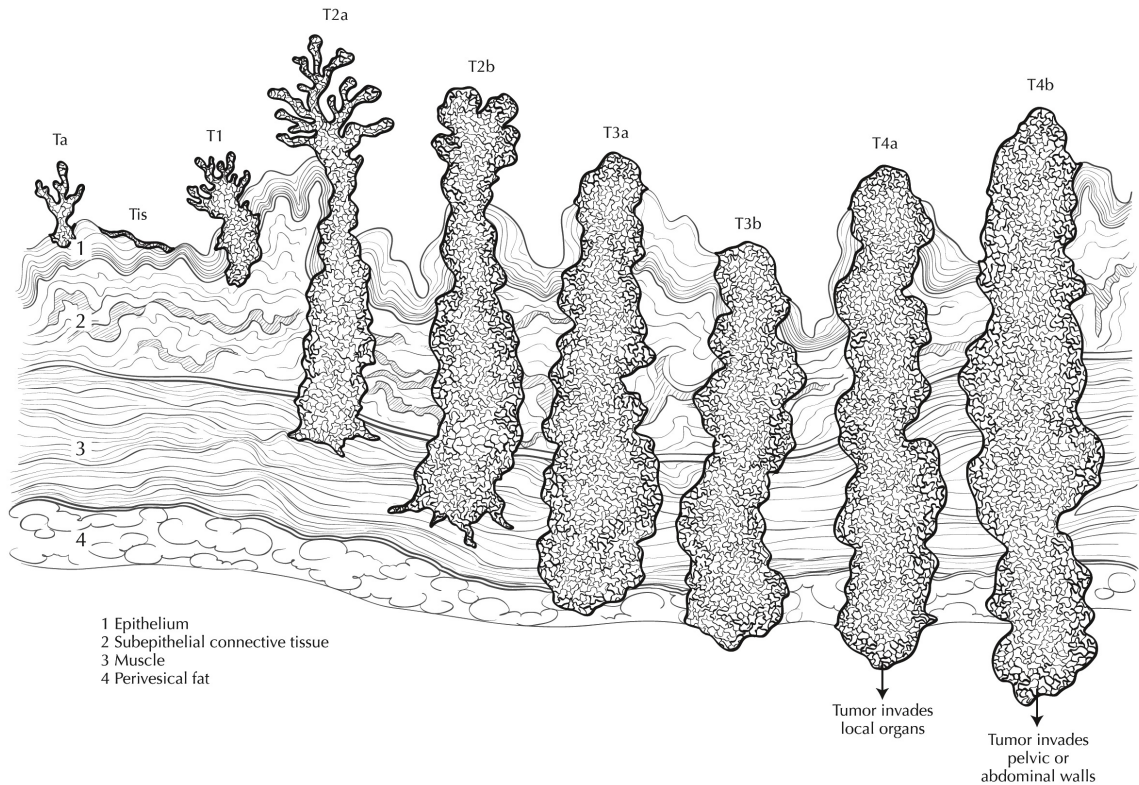


Figure 2.2: Extent of primary bladder cancer ¹ [8]

PUNLMP was introduced to describe the low probability of recurrence and progression of this abnormal growth of tissue [10].

Both classification systems use three distinct grades, they are however not interchangeable. Some WHO73 grade 1 will correspond to cases in both PUNLMP and *low grade*, WHO73 grade 2 corresponds to cases in both *low* and *high grade*, figure 2.3. TCC denotes *Transitional cell carcinoma*.

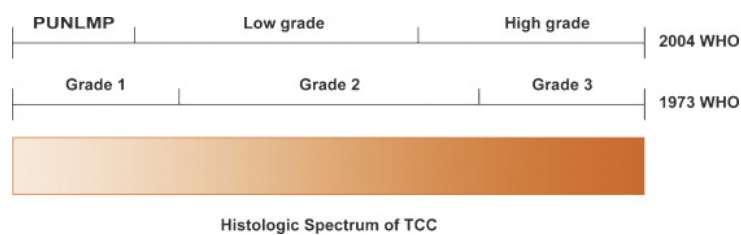


Figure 2.3: Comparison of WHO73 and WHO04 grading system [12]

Improvement of interobserver variability was one of the goals of WHO04, it was however not increased[13]. A study on interobserver variability using the WHO73 and WHO04 guidelines with three pathologists, found only a perfect agreement between two of the pathologists on a single feature out of 22 studied. While a good agreement was found on three features, and only between two of the pathologists [14].

¹Used with permission of the American Joint Committee on Cancer (AJCC), Chicago, Illinois. The original and primary source for this information is the AJCC Cancer Staging Manual, Seventh Edition (2010) published by Springer Science+Business Media.

2.2 Material

The material used in this study is a subset of a larger data set first used by Mangrud et al. [15]. The original study was approved by the Norwegian Regional Ethics Committee (#106/09). It contains 249 consecutive cases of primary non-muscle invasive urothelial carcinoma diagnosed at the Department of Urology and Pathology at Stavanger University Hospital during the period January 1. 2002 - December 31. 2006. 56 cases were excluded by Mangrud et al., primarily due to inadequate sample quality [15].

Tumor tissue was fixed in 4% buffered formaldehyde, dehydrated and embedded in paraffin, and four μm thick sections stained with Haematoxylin-Erythrosine-Saffron (HES) was used [15]. Digitalization of these tissue samples were done at the Department of Urology and Pathology at Stavanger University Hospital.

A subset of 42 cases were chosen from the original dataset by Mangrud [14]. Where the subset includes all cases with progression (14 cases) and 28 cases randomly selected from the remaining 179 cases without progression. No significant differences were found between the randomly selected cases and the remaining 151 cases [14]. Recurrence was defined as reappearance of urothelial carcinoma in the bladder. Progression was defined as an advancement in stage, histologically proven metastasis or death of disease within a five year period after first diagnosis. This study will however consider all available follow-up information. The subset are then changed, as progression occurred for one patient after 5 years and 1 month. Resulting in 15 cases with progression and 27 cases without progression, table 2.1.

	No recurrence	Recurrence	Progression
# of patients	13	14	15

Table 2.1: Prognostic follow-up information

Follow-up information was acquired from medical records at the Department of Pathology at Stavanger University Hospital. Provided that the original diagnosis was pTa grade 1 and cystoscopies acquired during follow-up were negative, follow-up cystoscopies were done using the *low follow-up regime* shown in table 2.2. Follow-up for all other cases were done using the *high follow-up regime*.

	low follow-up regime	high follow-up regime
Year 1	After 3 months	every 3 months
	After 9 months	
Year 2	Annually for 5-10 years	every 4 months
Year 3		every 6 months
Year 4		Annually
Year 5		Annually
Year 6		thereafter

Table 2.2: Follow-up regime

Original images in the hospital database are stored as single-file pyramidal tiled BigTiff with non-standard metadata through the use of the file format *scn*. With file sizes of approximately 1GB per image, image processing using complete images were not considered feasible due to the high computational requirement. As a consequence of this, smaller sections of the images were chosen in collaboration with a pathologist with the goal of illustrating cell maturation from

the connective tissue to the outer cell layer. Regions with artifacts due to heating damage or other external causes were avoided. Folded tissue was included in one image (case ID116), as the chosen section was otherwise a good representation of the tumour. Exporting images from the hospital database were done using a *export current view* function, saving current view of the issue using the monitor resolution. As a result of this, image resolution are 1918×921 .

2.3 Histogram equalizing

Histogram equalizing is a well-known method used to enhance contrast in images. The method is used in a wide range of applications ranging from medical image processing to radar image processing [16].

Given an image X of L levels, where the individual luminance levels are described by (X_0, \dots, X_{L-1}) and $X\{i, j\}$ are the grayscale value in location $\{i, j\}$. A probability density function, $p_x(k)$, for the image is defined as

$$p_x(k) = \frac{n_k}{n}, \quad 0 \leq k \leq L \quad (2.1)$$

Where n are the total number of pixels in the original image X , n_k the number of pixels of value k . The number of levels, L , are normally 256 (8 bit). A cumulative distribution function, c_x is then defined from p_x

$$c_x(x) = \sum_{j=0}^k p(j) \quad (2.2)$$

The grayscale levels in the input image are then mapped to cover the entire dynamic range, $[x_0, x_{L-1}]$, using the cumulative distribution function as a transform function.

$$f(x) = X_0 + (X_{L-1} - X_0)c_x(x) \quad (2.3)$$

The equalized output image, Y , can then be expressed using

$$Y = f(X) \quad (2.4)$$

$$= f(X(i, j)) | \forall X(i, j) \in X \quad (2.5)$$

One example of histogram equalization is illustrated in figure 2.4 and 2.5. Figure 2.4 show a small section of tissue taken from case ID 98, with its corresponding histogram. Where blue illustrates p_x and the black line illustrates c_x .

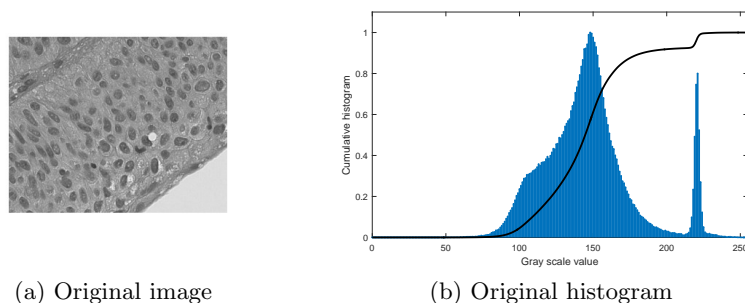


Figure 2.4: Histogram equalizing, example image

The resulting image and corresponding histogram achieved using histogram equalizing is shown in figure 2.5. The cumulative distribution function follows a linear trend, illustrating a similar amounts of pixels at all possible values. An increase in contrast are also seen in the resulting image.

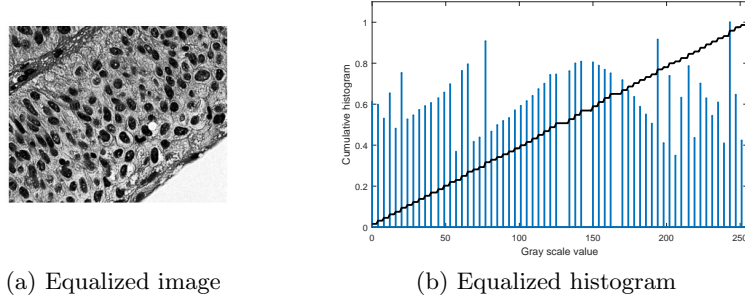


Figure 2.5: Histogram equalizing, transformed image

2.4 Image thresholding

Binary thresholding is a simple approach of image segmentations, where an image is divided into two classes to describe the distinct regions. A well-known method for automatically selecting the threshold value for a grayscale image is Otsu's method [17], which determines the optimal threshold value. The method can also be used to determine multiple thresholding values, assuming the image histogram has a multimodal distribution.

Given a normalized histogram p . Otsu's method define two classes to describe the binary image, where the probability of each class is given by

$$w_0 = \sum_{i=1}^t p_i, \quad w_1 = \sum_{i=t+1}^L p_i \quad (2.6)$$

where t describes the chosen threshold value, p_i bin i of the image histogram and L the number of levels in the image. Mean of the two classes are given by

$$\mu_0 = \sum_{i=1}^t \frac{ip_i}{w_0}, \quad \mu_1 = \sum_{i=t+1}^L \frac{ip_i}{w_1} \quad (2.7)$$

Variance in each class is then given by

$$\sigma_0^2 = \sum_{i=1}^t (i - \mu_0)^2 \frac{p_i}{w_0}, \quad \sigma_1^2 = \sum_{i=t+1}^L (i - \mu_1)^2 \frac{p_i}{w_1} \quad (2.8)$$

And the within-class variance given by

$$\sigma_w^2 = w_0\sigma_0^2 + w_1\sigma_1^2 \quad (2.9)$$

Otsu's method uses exhaustive search to determine the optimal threshold value t by find the threshold maximizing the within-class variance. An example of the method is illustrated using a tissue sample from the data set and its corresponding thresholded image, shown in figure 2.6.

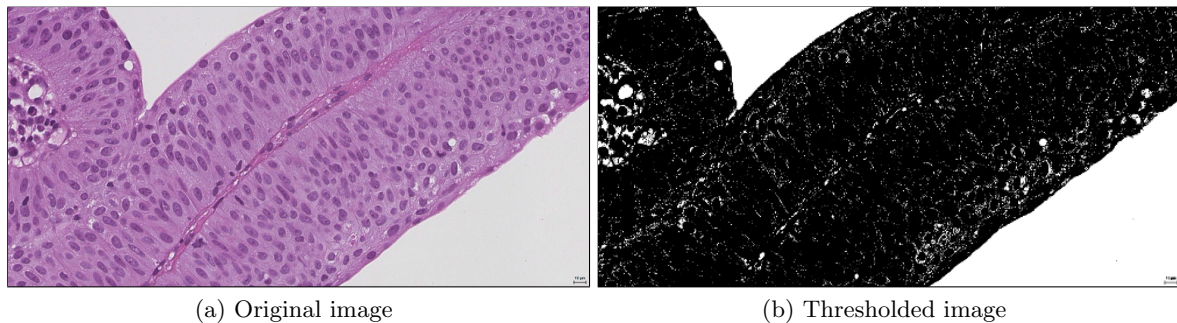


Figure 2.6: Example, Otsu's method

2.5 Local Binary Pattern

The original local binary pattern (LBP) operator, introduced by Ojala et al., provides a robust way of describing local texture in a 3×3 neighbourhood [18]. For each pixel in the image, a binary label can be obtained by comparing the center pixel with each of its neighbours. The binary label can then be represented by a numeral label using predefined weights for each of the neighbouring samples. Using a 3×3 neighbourhood, a maximum of 256 (2^8) textures can be described. The occurrence of texture labels over a predefined region, are used to create a histogram describing the local texture. LBP have previously been applied with promising results in multiple areas, ranging from face detection [19] to identification of liver disease [4].

A limitation of the original operator has been its small spatial support area, due to its 3×3 neighbourhood. Features found in this neighbourhood cannot capture large structures in the texture, which could be a dominant. As a solution, extensions of the operator was introduced by Ojala et al. to facilitate for uniform patterns and a rotation invariant analysis of image textures at multiple scales [20].

2.5.1 Principle

Given an arbitrary image pixel, its circular neighbourhood can be described using a radius r and a fixed number of samples along the circle. By a comparison of each sample in the circular neighbourhood with the center sample, a binary label is set to illustrate if samples are above or below the threshold defined by the center sample, as shown in figure 2.7.

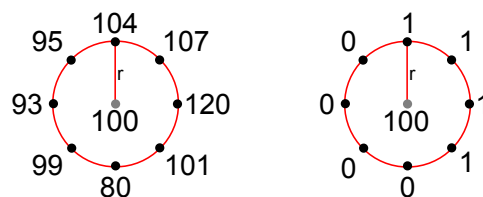


Figure 2.7: LBP: principle

A numeral label is found using individual weights to each sample in the neighbourhood, corresponding to figure 2.8.

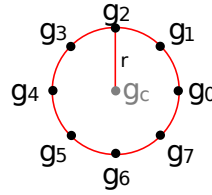


Figure 2.8: LBP: weights

Where each weight is defined by

$$g_p = 2^p \quad (2.10)$$

The LBP operator is denoted $LBP_{P,R}$, where P is the number of neighbours and R is the operator radius. The operator can then be defined as:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \quad (2.11)$$

where $s(x)$ is the logic function defined by

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (2.12)$$

In the case where neighbouring samples does not fall in the center of an image pixel, bilinear interpolation is used. A histogram of the texture descriptors in a defined region is used to identify regions or known textures in images. Figure 2.9 shows the MATLAB test image *bag.png*, where two textures are visually distinguishable. By doing LBP on both textures separately, using the parameters $P = 8, R = 3$, the histograms shown in figure 2.10 are obtained. The distribution has clearly different characteristics, and can used to identify the two regions.

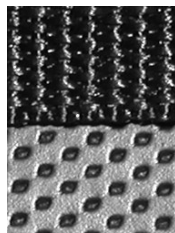


Figure 2.9: LBP example: texture, source: MATLAB test images

2.5.2 Rotation invariance

An extension of the original operator is the introduction of rotation invariance [20]. As texture orientation often can be arbitrary, a rotation invariant way of describing the texture is desired. Using figure 2.7 as an example, LBP gives the binary pattern 10000111. With various orientation

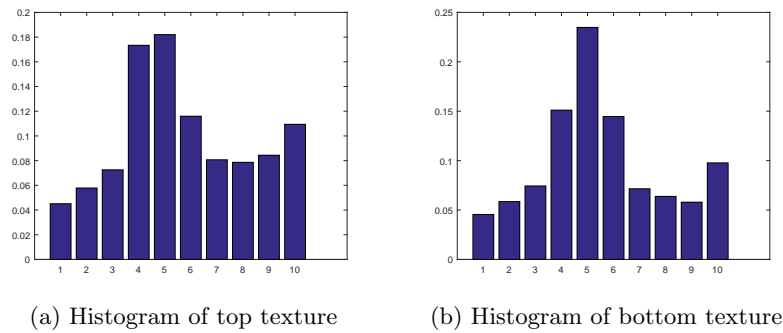


Figure 2.10: LBP example: histograms

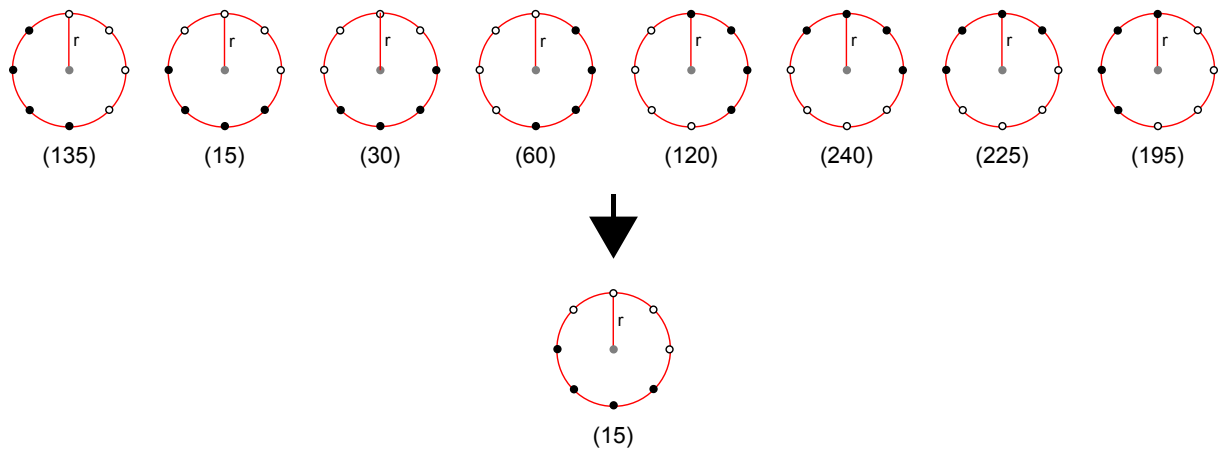


Figure 2.11: LBP: Rotation invariance principle

of the texture, the center pixel can be described by a total of eight different neighbourhoods, illustrated in figure 2.11.

The grey circle is defined as the center pixel, black circles as zeros and white circles as ones. By applying weights, the eight numeral values are obtained. The rotation invariant descriptor using the minimum of the possible descriptors are then be described using:

$$LBP_{P,R}^{riu} = \min \{ROR(LBP_{P,R}, i) | i = 0, \dots, P - 1\} \quad (2.13)$$

Where ROR is a rotate operation, used to find the P possible rotations of the LBP label. The minimum value found is 15, corresponding to the binary pattern 00001111 illustrated in figure 2.11.

2.5.3 Uniform

A second extension of the original operator is uniform patterns [20]. Ojala et al. observed in their experiments that the nine uniform patterns in $LBP_{8,1}^{riu2}$, where $riu2$ denotes rotation invariant uniform patterns with a maximum of 2 binary transition, contributed on average to 87.2 and 89.7 percent of all patterns. In the case of $LBP_{16,2}^{riu2}$, the 17 uniform patterns contributed on average to 66.9 and 70.7 percent of all patterns in the image.

Pattern uniformity, $U(LBP_{P,R})$, is given by equation 2.14, and describes the number of transitions between zero and one in the local binary pattern. For example, the pattern 00011000 and 00001111 contains two transitions while 01101100 contains four transitions.

$$U(LBP_{P,R}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)| \quad (2.14)$$

The local binary pattern is denoted $u2$ if the pattern uniformity is equal to two or less. With P neighbours, $P + 2$ possible bins for rotation invariant uniform patterns are defined. Where the number of uniform patterns are $P + 1 = 9$ and the last bin describes all other patterns.

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c) & \text{if } U(LBP_{P,R}) \leq 2 \\ P + 1 & \text{otherwise} \end{cases} \quad (2.15)$$

Figure 2.12 illustrates the proposed bins and their corresponding uniform patterns, using the LBP operator with eight neighbours. Black circles illustrates zeros and white circles illustrates ones.

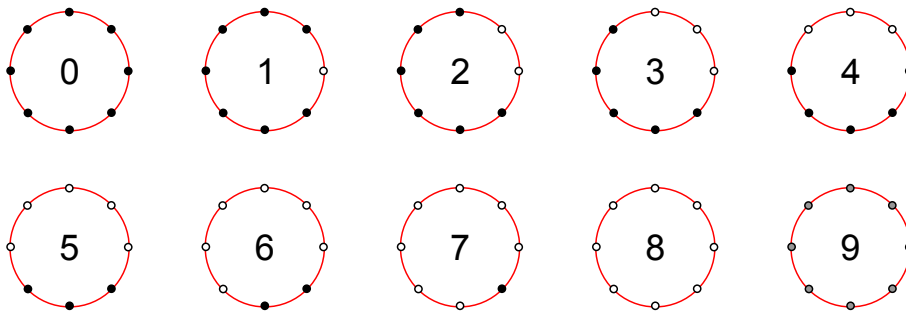


Figure 2.12: LBP: Uniform rotation invariant descriptor bins

2.5.4 Local variance

The LBP operator, defined in section 2.5.1, is a binary label of neighbouring samples in regard to the center sample itself, and is therefore not affected by monotonic transformations of the grayscale. A second rotation invariant measure of the local variance, using the same neighbourhood defined by LBP, was proposed by Ojala et al. [20].

The local variance of a neighbourhood, denoted $VAR_{P,R}$, is given by

$$VAR_{P,R} = \frac{1}{P} \sum_{p=0}^{P-1} (g_p - \mu)^2 \quad (2.16)$$

where

$$\mu = \frac{1}{P} \sum_{p=0}^{P-1} g_p \quad (2.17)$$

Since $LBP_{P,R}^{riu2}$ and $VAR_{P,R}$ are complementary, the joint distribution $LBP_{P,R}^{riu2}/VAR_{P,R}$ is expected to be a powerful rotation invariant measure [20].

2.5.5 Multiresolution analysis

Texture identification using a single predefined neighbourhood and radius can be challenging as features on multiple scales describe different properties of the texture. A multiresolution analysis can therefore be accomplished by combining information from multiple operators with an arbitrary radius and number of pixels in the neighbourhood [20]. There is however a problem in this direct approach. Keeping the number of neighbouring pixels fixed while increasing the radius, an increasingly sparse representation of the image is obtained. Thus introducing aliasing unless a low-pass filter is applied to the image. To solve this problem, Mäenpää et al. [21] proposed a multi-resolution LBP combined with Gaussian filtering. Using a low-pass Gaussian filter, the pixel intensity information for a sample is collected from a larger area. In addition to removing aliasing, this will also reduce the effect of noise.

The proposed method by Mäenpää uses an exponentially growing radius and non-overlapping effective areas. Figure 2.13 illustrates the effective areas for each sample using various scales. Operator radii are illustrated as the dotted circles, while the effective areas around each sample are illustrated as solid circles.

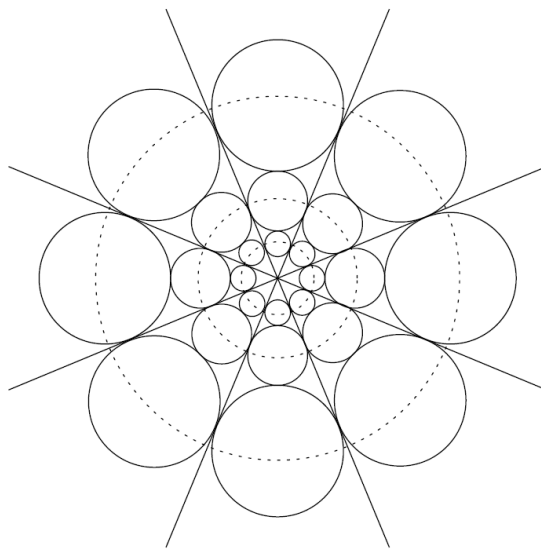


Figure 2.13: Exponential multi-resolution LBP[21]

By defining the effective area as touching and non-overlapping, the operator radius, R_n , at scale n ($n \geq 2$) are defined as:

$$R_n = \frac{r_n + r_{n-1}}{2} \quad (2.18)$$

Where r_n is given as the outer radius of the effective area on scale n

$$r_n = r_{n-1} \left(\frac{2}{1 - \sin(\pi/P_n)} - 1 \right), n = 2, \dots, N \quad (2.19)$$

Where N is the number of scales and P_n is the number of samples at scale n . Since low-pass filtering is only useful when using a radius above 1, r_1 is defined as 1.5 which is equal to the shortest distance from the center to the 3×3 neighbourhood border. The effective areas are then realized using Gaussian low-pass filters with 95% of it's mass inside the circle [21].

The operators obtained using varying scales (P, R) can be used individually, or be combined to capture larger texture structures [22]. By combining multiple operators, similarity between a pair of multiscale LBPs can be computed as the sum of similarity of each scale.

$$f = \sum_{r=1}^R \chi_r^2 \quad (2.20)$$

Where R is the number of operators of varying P, R , and χ^2 is the similarity measure between an observed histogram and a model. The combination of multiple operators in equation 2.20 assigns a uniform weight to each scale. A more general selection scheme is the use of individual normalized weights $w \in [0, 1]$ for each scale [22].

$$f = \sum_{r=1}^R w_r \chi_r^2 = \mathbf{w}^T \boldsymbol{\chi}_r^2 \quad (2.21)$$

Where \mathbf{w}^T is a $1 \times R$ weight vector and $\boldsymbol{\chi}_r^2$ is a column vector of χ_r^2 .

2.6 Classification

Training a classifier using data sets with imbalanced classes can be challenging. When a single class greatly outnumber other classes, traditional methods tend to create models where the largest class is favored [23]. Multiple algorithms have been proposed to alleviate this challenge, including data sampling and boosting [24].

The section first presents the principles of two algorithms used for alleviating class imbalance, before a hybrid approach is presented.

2.6.1 Class imbalance

Boosting

Boosting is the process of using weak learners to create a collection of models for a more complex classifier. With a reduced classification error, bias and variance in the complex classifier [25]. The weak learner can use weights or a subset of the training set to focus on the misclassified examples [26].

The well-known boosting algorithm AdaBoost [27], uses an iterative approach for building an ensemble of models. Weights are updated during each iteration for all examples in the training set. Weights are initially set equal, and in each iteration weights associated with misclassified examples are increased, while they are decreased for correctly classified examples. This forces the learner to focus more on the misclassified examples in the next iteration. Classification of new examples are then done using a weighted combination of all learners in the model.

Data sampling

Data sampling is used to alleviate class imbalance. This can be done using two different approaches, undersampling and oversampling.

Undersampling is based on removing enough examples from the larger classes to obtain a balanced data set with n examples in each class, where n is the number of examples in the smallest class. A data set with 90-10 distribution between two classes, will be reduced to 20% of its original size. As a consequence of this, the computational requirement to train the classifier will be reduced. The big drawback of using this approach is the possibility of losing information from examples which could be vital to identification of the relevant class [28].

Oversampling on the other hand, is based on increasing the number of examples, n , in the smaller classes. This can be done by duplicating existing examples or using interpolation to obtain new examples. Oversampling by duplicating examples can lead to overfitting [29]. A well-known oversampling approach using interpolation is SMOTE (Synthetic Minority Oversampling Technique (SMOTE)), which creates new interpolated samples in-between existing samples [30]. Using the same data set with a 90-10 distribution, the resulting size after oversampling would be 180% of the original data set. Thus increasing the computational requirement to train the classifier. Another challenge also occurs when oversampling extremely small classes, where the interpolated examples might not follow the actual distribution of the class.

2.6.2 A hybrid approach

Chawla et al. [31] proposed the novel approach SMOTEBoost, using data sampling in combination with AdaBoost. The technique create new training samples by taking the difference between a sample and one of its nearest neighbours of the same class, multiplying the difference by a random number between 0 and 1 before adding it to the original sample value. This causes a new synthetic training sample along the line between the original sample and its neighbours [30]. Improved classification results are shown by Chawla et al. when using SMOTE on a data set compared to AdaCost, an AdaBoost deviation with incorporated cost of misclassification examples. Improved results using SMOTEBoost were also shown compared to using SMOTE alone [31].

The drawback of SMOTEBoost is however the increased computational requirement due to its oversampling technique. As a solution to this, Seiffert et al. proposed RUSBoost [23]. A hybrid approach using random undersampling in combination with AdaBoost to alleviate class imbalance. Unlike AdaBoost, RUSBoost utilize boosting by resampling the training data according to the assigned weights. The primary drawback of using undersampling techniques is however the loss of information. The combination of undersampling and boosting overcome this drawback as the removed examples are likely to be used in other iterations of the boosting technique. Both SMOTEBoost and RUSBoost outperform other classifiers, and while RUSBoost is a simpler and faster technique it performs comparably to SMOTEBoost [23].

An overview of RUSBoost is shown in algorithm 1. Given an example (x_i, y_i) , where x_i is a point in feature space X and y_i is a class label in a set of labels Y . The principle of the algorithm is training a weak hypothesis using a selection of examples in the data set for each iteration t . Weak learners are created using the classifier algorithm *Weaklearn*, until the final ensemble consists of T weak learners. A strong learner is then created using a weighted combination of the weak learners. In the algorithm, D_t describes the normalized weight distribution for examples, and $h_t(x_i)$ the weak hypothesis for x_i . In step 1 in the algorithm, weights for all examples are initialized as $1/m$, where m is the number of examples in the training set. In step 2 (shown as

2a-2g), T weak learners are trained. Random undersampling is done in step 2a to reduce the size of the largest class, until $N\%$ of the new temporary data set S'_t belongs to the smallest class. The temporary data set will have a new weight distribution D'_t . The temporary data set and weight distribution are sent to the base learner *WeakLearn* in step 2b, and a weak learner is returned from the base learner in step 2c. A pseudoloss ϵ_t is then computed using the original training set S and its corresponding weight distribution in step 2d. A weight update parameter α is computed as $\epsilon_t/(1 - \epsilon_t)$ in step 2e. A new weight distribution for the next iteration, D_{t+1} , is computed in step 2f, and then normalized in step 2g. After T iterations of step 2, the final hypothesis $H(x)$ is returned as a weighted combination of all T weak hypotheses in step 3 [23].

Algorithm 1: RUSBoost[23]

Given:Set S of examples $(x_1, y_1), \dots, (x_m, y_m)$ with minority class $y^r \in Y$, $|Y| = 2$ Weak learner, *WeakLearn*Number of iterations, T Desired percentage of total instances to be represented by the minority class, N

1. Initialize $D_1(i) = \frac{1}{m}$ for all i .
 2. Do for $t = 1, 2, \dots, T$
 - (a) Create temporary training dataset S'_t with distribution D'_t using random undersampling
 - (b) Call *WeakLearn*, providing it with examples S'_t and their weights D'_t
 - (c) Get back a hypothesis $h_t: X \times Y \rightarrow [0, 1]$.
 - (d) Calculate the pseudo-loss for S and D_t :

$$\epsilon_t = \sum_{(i,y):y_i \neq y} D_t(i)(1 - h_t(x_i, y_i) + h_t(x_i, y))$$
 - (e) Calculate the weight update parameter:

$$\alpha_t = \frac{\epsilon_t}{1 - \epsilon_t}$$
 - (f) Update D_t

$$D_{t+1}(i) = D_t(i)\alpha_t^{1+h_t(x_i, y_i) - h_t(x_i, y \neq y_i)}$$
 - (g) Normalize D_{t+1} : Let $Z_t = \sum_i D_{t+1}(i)$

$$D_{t+1}(i) = \frac{D_{t+1}(i)}{Z_t}$$
 3. Output the final hypothesis:

$$H(x) = \underset{y \in Y}{\operatorname{argmax}} \sum_{t=1}^T h_t(x, y) \log \frac{1}{\alpha_t}$$
-

2.7 Similarity measures

This section presents two measures for describing similarities between observations and a defined model.

2.7.1 Sum of squared error

Assuming a model with mean μ , similarity between an observation and the model can be described using its squared error.

$$SE = (X - \mu)^2 \quad (2.22)$$

Observations using the methods presented in section 2.5 produces a histogram of N bins, where each bin is considered an individual observation. The sum of squared errors for the histogram can then be used as a fitness measure for the observed histogram.

$$SSE = \sum_{i=1}^N (X_i - \mu)^2 \quad (2.23)$$

Where N is the number of bins in the histogram.

2.7.2 Chi-squared

Chi-squared, introduced by Pearson [32], can be used as a goodness of fit measure. Assume a model with normal distribution, $N(\mu, \sigma^2)$. Similarity of an observation X and the model can be described using

$$\frac{X - \mu}{\sigma} \quad (2.24)$$

where it follows a standard normal distribution. Observation using the methods described in chapter 2.5 gives a histogram of N bins. Where each bin in the histogram can be described as an individual observation, with its mean and variance. The sum of chi-squares for the histogram can then be used as a similarity of the histogram and a model.

$$\chi^2 = \sum_{i=1}^N \frac{(X_i - \mu_i)^2}{\sigma_i^2} \quad (2.25)$$

Where N is the number of bin in the observed histogram.

2.8 Performance Measures

Evaluation of machine learning performance is typically done using a confusion matrix [30]. Figure 2.3 illustrates a 2-class confusion matrix with true class on rows and predicted class on columns.

	Predicted Not A	Predicted A
True not A	tn	fp
True A	fn	tp

Table 2.3: Confusion matrix for a 2-class problem

Where tp is the number of correctly identified examples of class A (true positives), tn the number of correctly identified examples of class $notA$ (true negative), fn the number of incorrectly

identified examples from class A (false negative) and fp the number of incorrectly identified examples of class $notA$ (false positive).

Several useful measures can be defined using the confusion matrix. Overall accuracy of the used algorithm is defined as the proportion of correctly identified examples in all classes.

$$ACC = \frac{tp + tn}{tp + tn + fp + fn} \quad (2.26)$$

In the event of misclassifying one class is considered more costly, some additional measurements can be useful for performance evaluation. True positive rate (also known as sensitivity) is the proportion of correctly identified examples in class A .

$$TPR = \frac{tp}{tp + fn} \quad (2.27)$$

Corresponding, the true negative rate (also known as specificity) is the proportion of the correctly identified examples in class $notA$

$$SPC = \frac{tn}{tn + fp} \quad (2.28)$$

3. Implementation

In this chapter, each module in the designed system are presented in closer detail. First in the chapter, preprocessing of the input images will be described. Following this, a mask algorithm for automatic identification of relevant regions in the images will be presented. Extraction of features, classification and validation are then described in closer detail. Finally, a proposed system for classifying new images will be presented. An overview of the devised system are shown in figure 3.1.

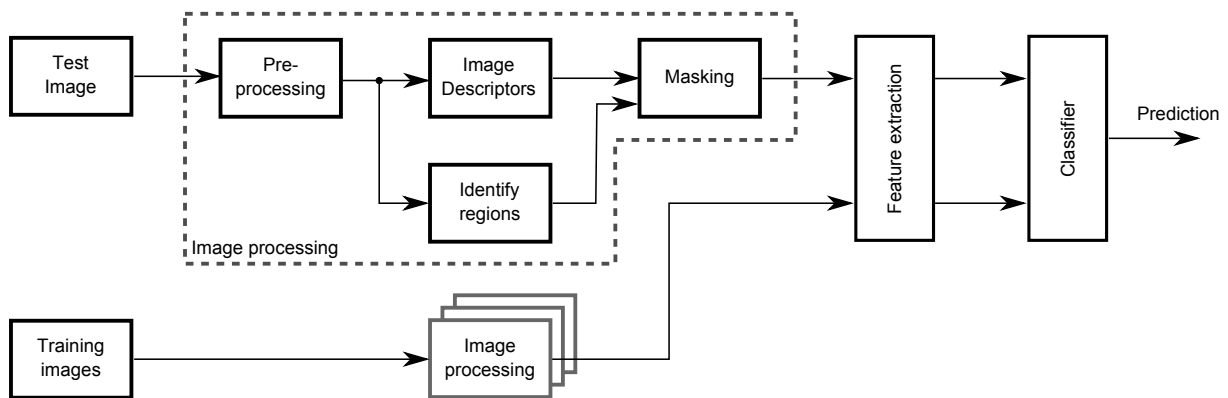


Figure 3.1: System overview

3.1 Preprocessing

This section describes the various steps used in the *preprocessing* block in figure 3.1.

Images used in this study are scans of four μm thick cell sections. Due to variations in actual thickness of the cell sections and amount of HES used, both shade and brightness of the tissue vary between samples, illustrated in figure 3.2. An overall darker colour is observed in the left image, this could be caused from a combination of the two subproblems. In a direct comparison between the two images, the cytoplasm in the left image appear to be darker than the cell nucleons in the right image. As these variations are caused by cutting and preparing the tissue samples, their effect on images are minimized using two methods. As shade depend on the amount of HES used, all images are converted to grayscale. In addition two approaches for image normalization of the resulting grayscale images are considered.

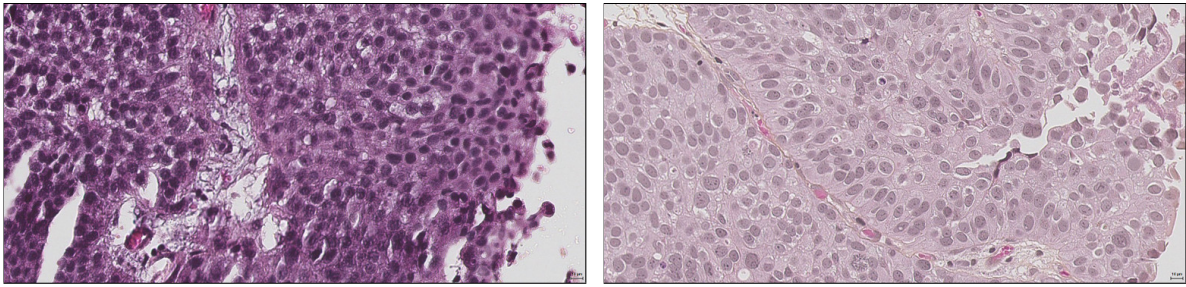


Figure 3.2: Variance in image colours

3.1.1 Grayscale expansion

Image normalization on grayscale images can be done by expanding the used grayscale spectre to cover the possible range of grayscale values, this are implemented using equation 3.1.

$$Img_{01norm} = \frac{Img - Img_{min}}{Img_{max} - Img_{min}} \quad (3.1)$$

Using this method directly on our dataset will however be challenging. A lack of low level values are observed in the right image in figure 3.2. There is however a black graphical scale embedded in the bottom right corner for all exported images, searching for the minimal grayscale value will then return the dark pixels in this scale. A close-up of the graphical scale are shown in figure 3.3.

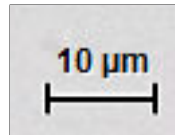


Figure 3.3: Graphical scale

A proposed solution to this, is to exclude the bottom right corner when searching for the minimum grayscale value in the image. Figure 3.4 illustrates the resulting images from figure 3.2 using this normalization.

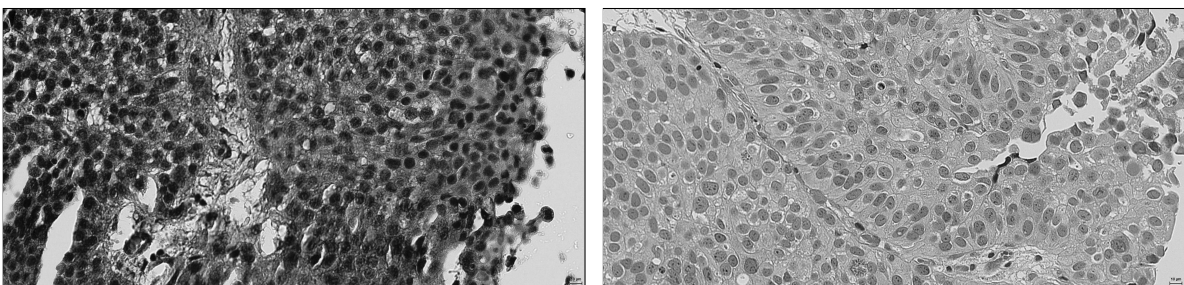


Figure 3.4: $[0, 1]$ normalized grayscale images

3.1.2 Histogram equalizing

An alternative approach of normalizing images is the use of histogram equalization. The method increase global contrast by changing the distribution of grayscale intensities in the image. Implementation of histogram equalizing, equation 2.1- 2.5 are done through the use of the included MATLAB function for histogram equalizing. Figure 3.5 illustrates the resulting images from figure 3.2 using this normalization. Both nucleons and cytoplasms are now visually similar in both images.

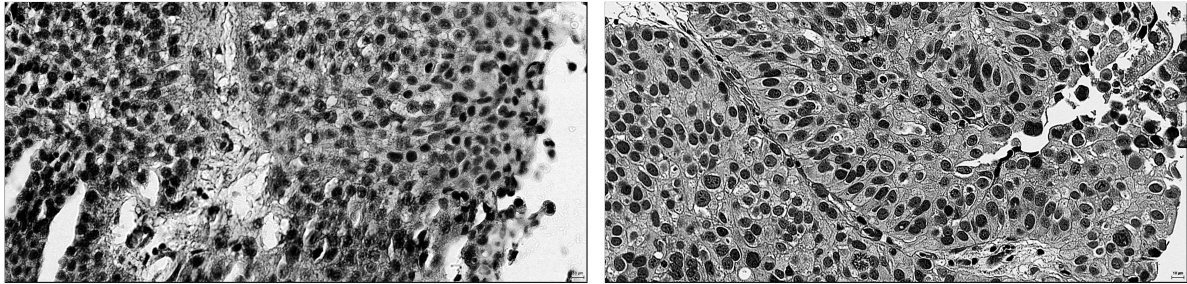


Figure 3.5: Histogram equalized grayscale images

3.2 Masking

This section describes the various steps used in the *Identify regions* and *Masking* blocks in figure 3.1. The proposed masking algorithms are devised using a microscopic images with a high magnification.

3.2.1 Identification of regions with tissue

As a consequence of how images are chosen, most images will include areas not covered by any cells. Textures found in these areas will not be representative for the patient, and should therefore not be taken into account when computing features. To easily exclude these areas from further analysis, an algorithm is devised to automatically create a logic mask to identify regions consisting of tissue.

Figure 3.6 illustrates an arbitrary image, and its corresponding mask. The generated masks are the same size as the original image, and use zeros to describe regions without tissue. It can therefore be applied by doing a pixel-wise multiplication of the image and its corresponding mask to remove regions without tissue.

A binary threshold based on the Otsu's method, described in section 2.4, is chosen to define major trends in the image. The resulting mask will have numerous holes in areas covered by tissue. Some additional regions will also be identified as tissue when true nature of the regions are not tissue. To remove these holes and additional regions, the mask are first dilated using a 9 pixel wide circular disk. Holes below a fixed size are then removed, before regions below a fixed size areas are removed. The mask is finally eroded with the same 9 pixel disk that was used for dilation.

A graphical scale, described in section 3.1.1, is embedded onto images when exporting from the medical image database. As a result of this, its size and position will always be known and the

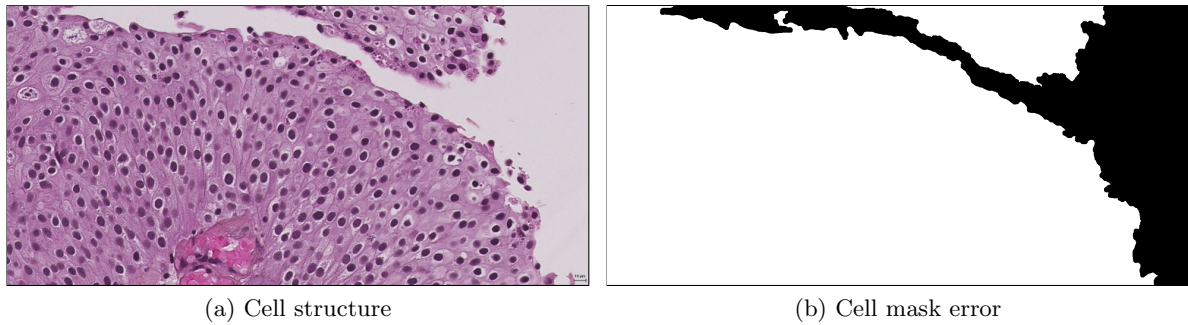


Figure 3.6: Patient ID 147, cell and cell mask

area it covers are therefore considered to not be relevant for a texture analysis. To avoid any effects from the graphical scale in the analysis, a mask is manually defined to remove the regions closer than 10 pixels of the graphical scale itself.

An overview of the devised masking function is shown in algorithm 2.

Algorithm 2: Generate cell masks

```

for all images do
    1. Threshold image using Otsu's method
    2. Dilate mask
    3. Fill all holes below fixed size
    4. Remove all regions below a fixed size
    5. Erode mask using same kernel as for dilation
    6. Remove graphical scale (in bottom right corner)
end

```

3.2.2 Identification of regions with connective tissue

Urothelial carcinoma arise in the outer cell layer, and connective tissue underneath can therefore be considered not relevant for the analysis. An automatic identification of connective tissue are therefore desired. Two approaches of identifying these regions are proposed. Variations in texture of the connective tissue are also observed throughout the data set, illustrated in figure 3.7 for two of the images in the data set.

LBP

An approach of identifying connective tissue is with the use of LBP. Using histograms in a small window, local texture can be described and identified. The devised method divides each image into blocks of size $N \times N$. If the mean squared error (MSE) between the window histogram and a model histogram are above a set threshold, pixels in the block are marked as not relevant. Due to the small sections of connective tissue in each image, the model histogram is defined using labels of all tissue in the image. An overview of the method are shown in algorithm 3.

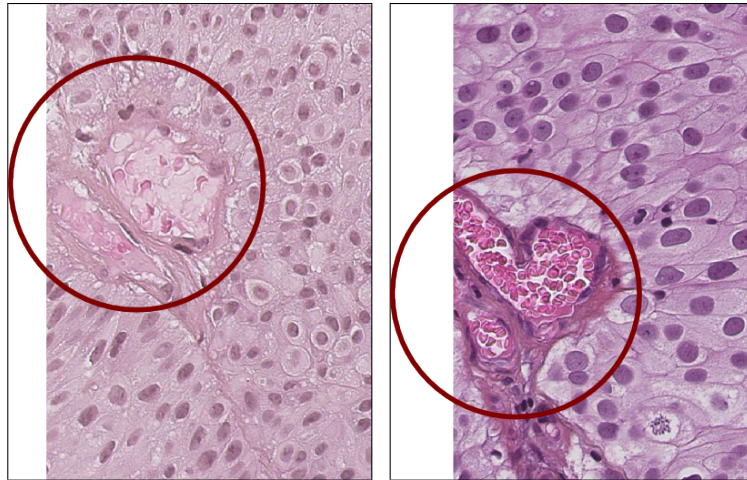


Figure 3.7: Variations in connective tissue

Algorithm 3: Identify connective tissue using LBP

```

Compute LBP for all image pixels
model = normalized LBP histogram of tissue
for blocks of size  $N \times N$  do
  create normalized histogram of block
  if  $MSE(blockHistogram, model) > threshold$  then
    | Mark block as not relevant
  end
end

```

Multi-level thresholding

Identification of regions with no tissue in the image is done using Otsu's thresholding, presented in section 2.4. An extension of this algorithm is the use of multi-level Otsu's thresholding to also identify connective tissue. A quantized version of the image of N levels are computed using the thresholds found using Otsu's method, a mask is then created using a range of the quantized levels. The resulting mask will have numerous holes, these are eliminated by removing all holes below a set size, s , to avoid removing holes created by connective tissue. An overview of the function is shown in algorithm 4. Implementation of Otsu's multi-level thresholding is done using the available MATLAB implementation.

Algorithm 4: Identify connective tissue using Otsu's multilevel threshold

1. $thresh =$ Compute N thresholds using Otsu's method
 2. $img_Q =$ Quantize image using $thresh$
 3. $Mask = a \leq img_Q \leq b$
 4. Remove holes in mask smaller than s
-

3.3 Image descriptors

This section presents how the various image descriptors are defined in the block *Image descriptor* in figure 3.1.

For analysis using radii higher than 1, images are preprocessed using a gaussian lowpass filter. As the devised system does not implement the exponentially increasing operator radii described by Mäenpää in section 2.5.5, radius of the gaussian filter is found using

$$R = \frac{\pi r_n}{P_n} \quad (3.2)$$

Where P_n are the number of neighbours, and r_n the radius of scale n .

3.3.1 Local binary pattern

A rotation invariant uniform local binary pattern is computed for each pixel in an image, the image is then masked using the automatic cell masks presented in section 3.2. Remaining LBP values are grouped into a normalized $P + 2$ bin histogram, where P is the number of neighbours used in the computation of LBP labels. For further information about the LBP operator, see section 2.5. LBP is computed using the implementation available from the University of Oulu (lbp.m, version 0.3.3) [33].

3.3.2 Local variance

A rotation invariant variance is computed for each pixel in an image using the same neighbourhood defined by the LBP operator, the image is then masked using the automatic cell masks presented in section 3.2. Remaining variance values are then grouped into a normalized B bin histogram. As variance have a continuous distribution, a large B will keep the distribution form intact. For further information about the pixel variance, see section 2.5.4. Local variance is computed using the implementation available from the University of Oulu (cont.m, version 0.1.0) [33].

3.4 Feature extraction

This section presents the block *feature extraction* in figure 3.1.

3.4.1 Model histograms

Model histograms are introduced based on known prognosis of patients in the validation set. Where models can be defined as 1. *not recurring*, 2. *recurring, not progressive* and 3. *recurring and progressive* or 1. *not recurring* and 2. *recurring*. A model histogram for each prognosis, ωx , is computed using

$$\mu_{\omega x} = \frac{1}{n} \sum_{i=1}^n h_{\omega x, i} \quad (3.3)$$

where $h_{\omega x, i}$ describes the i th histograms of class ωx , and n the number of patients in the training set in class ωx . Variance for each bin in the histogram are found using

$$\sigma_{\omega x}^2 = \frac{1}{n} \sum_{i=1}^n (h_{\omega x, i} - \mu_{\omega x})^2 \quad (3.4)$$

where $\omega x, i$ describes the i th histogram of class ωx , n the number of patients in class ωx and $\mu_{\omega x}$ is the mean histogram of patients belonging to class ωx .

A chi-squared measure, equation 2.25, is then computed to find the similarity between the image histogram and models, ωx , for each descriptor using

$$\begin{aligned} f_{i, \omega x} &= \chi_{i, \omega x}^2 \\ &= \sum_{n=1}^N \frac{(h_{i, n} - \mu_{n, \omega x})^2}{\sigma_{n, \omega x}^2} \end{aligned} \quad (3.5)$$

where $h_{i, n}$ describes bin n of the histogram from image i , $\mu_{n, \omega x}$ mean of bin i , $\sigma_{n, \omega x}^2$ variance of bin i , and N the number of bins in h .

Using the MATLAB test image *bag.png* as an example, shown in figure 2.9. The two textures are denoted *texture 1* for the top texture, and *texture 2* for the bottom texture. A model histograms for each texture can be found using equation 3.3 and its corresponding variance using equation 3.4. A model computed from five regions of each texture, using LBP with parameters $P = 8$ and $R = 1$, are shown in figure 3.8.

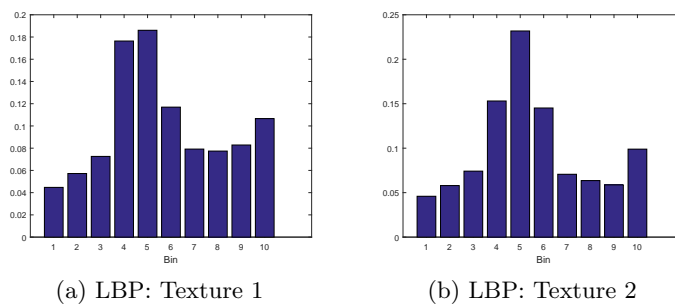


Figure 3.8: LBP model histograms

Given a new image of unknown texture, figure 3.9a, LBP labels are computed for all pixels. The resulting histogram of uniform rotation invariant LBP labels are shown in figure 3.9b.

A chi-squared measure is then computed for the new histogram and the two known texture model histograms using equation 3.5, shown in figure 3.10. The sum of chi-squared are then used to describe similarity between the new texture and the known models. A higher summed chi-squared is clearly observed for the bottom texture, shown as red in figure 3.10. With a better similarity to the top texture, the new texture are classified as this.

3.4.2 Feature vectors

The number of features used in experiments varies, both between experiments and within each experiments. Computation of features are however predefined. Each feature describes similarity

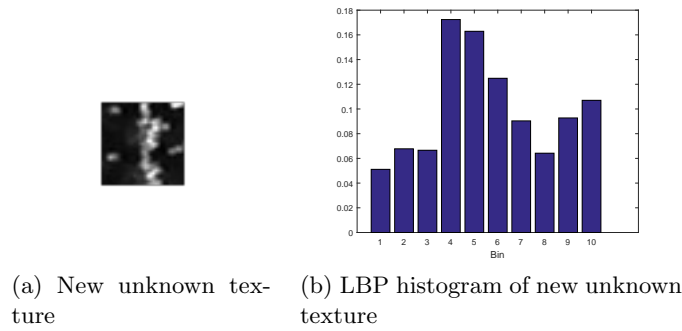


Figure 3.9: Example: New unknown texture with its corresponding LBP histogram

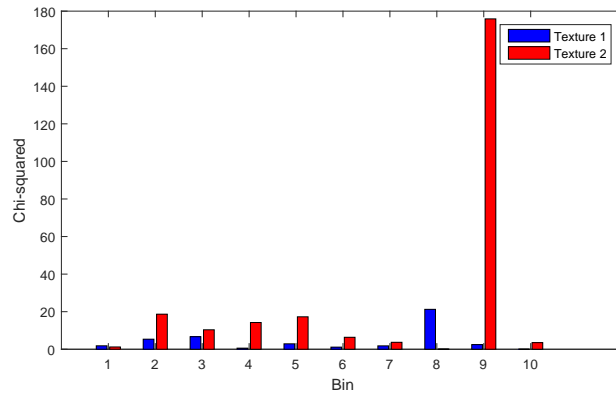


Figure 3.10: Example: Chi-squared for the texture

between a new image histogram, and computed models from the training set using chi-squared, equation 3.5. To capture larger texture structures, the combinations of multiple scales P, R are done using a weighted sum of chi-squares, equation 2.20. Models are defined for known prognosis of patients in the training set (no-recurrence, recurrence and progression). Each model is then created as the mean histogram of all patients with a predefined prognosis, equation 3.3. Variance for each bar in the histogram is computed using equation 3.4.

Several feature vectors are proposed for use in experiments. A combined vector, using both LBP and local variance in combination with all known prognosis, results in a six feature vector, equation 3.6.

$$f_i = \begin{bmatrix} \mathbf{w}^T \chi_{LBP, nR}^2 \\ \mathbf{w}^T \chi_{LBP, RnP}^2 \\ \mathbf{w}^T \chi_{LBP, P}^2 \\ \mathbf{w}^T \chi_{Var, nR}^2 \\ \mathbf{w}^T \chi_{Var, RnP}^2 \\ \mathbf{w}^T \chi_{Var, P}^2 \end{bmatrix} \quad (3.6)$$

Where \mathbf{w}^T are a weight vector, χ^2 a vector of similarities between a new image and a known model using multiple scales. Prognosis models are denoted nR for *no-recurrence*, RnP for *recurrence no-progression* and P for *progression*. LBP and variance can also be used individually in combination with all known prognosis, equation 3.7.

$$f_i = \begin{bmatrix} \mathbf{w}^T \chi_{nR}^2 \\ \mathbf{w}^T \chi_{RnP}^2 \\ \mathbf{w}^T \chi_P^2 \end{bmatrix} \quad (3.7)$$

By estimating no-recurrence/recurrence in patients, the feature vector can be reduced from the original six feature vector to a four feature vector, equation 3.8. This are done by grouping all patients with recurrence and progression into one group.

$$f_i = \begin{bmatrix} \mathbf{w}^T \chi_{LBP,nR}^2 \\ \mathbf{w}^T \chi_{LBP,R}^2 \\ \mathbf{w}^T \chi_{Var,nR}^2 \\ \mathbf{w}^T \chi_{Var,R}^2 \end{bmatrix} \quad (3.8)$$

The feature vector in equation 3.8 can be further reduced to two features by using LBP or variance individually.

$$f_i = \begin{bmatrix} \mathbf{w}^T \chi_{nR}^2 \\ \mathbf{w}^T \chi_R^2 \end{bmatrix} \quad (3.9)$$

Identification of patients with progression is highly desired, as a high follow-up regime will increase the possibility of identifying recurring tumours at an early stage. To identify this prognosis, both feature vectors consisting of all three models, equation 3.6 and 3.7 can be used. In addition, a feature vector to distinguish between patients without recurrence and patients with progression are proposed.

$$f_i = \begin{bmatrix} \mathbf{w}^T \chi_{LBP,nR}^2 \\ \mathbf{w}^T \chi_{LBP,P}^2 \\ \mathbf{w}^T \chi_{Var,nR}^2 \\ \mathbf{w}^T \chi_{Var,P}^2 \end{bmatrix} \quad (3.10)$$

3.5 Classification

This section presents the steps used in block *Classifier* in figure 3.1.

Implementation of the *RUSBoost* classifier is done using the available MATLAB implementation through the use of the *ensemble function*. Due to the limited data available, evaluation of system performance using a dedicated training set for the classifier is not considered feasible. A leave-one-out cross validation is chosen over a k-fold approach to maximize data available for training the classifier. A consequence of computing features as the similarity between a new image and histogram models, is that the feature vector will change for each iteration of the cross validation. As the image itself is excluded from computation of the models. To overcome this, a cross validation is implemented manually, with computation of new feature vectors added to each iteration. An overview of the classification is shown in algorithm 5.

For each classification in the leave-one-out cross validation, features are computed using pre-computed LBP and local variance descriptors. Features are calculated as the chi-squared of a descriptor histogram for an image and the internal models. Internal models are defined as

Algorithm 5: Leave-one-out crossvalidation

```

Given features  $X$ 
for each image  $i$  do
  for each image  $j$  do
    forall the features do
      Define models for each class using all images except  $i, j$ 
      Calculate chi-squared for image  $j$ 
    end
  end
  Train classifier using  $X$ 
  Predict class of image  $i$ 
end

```

the median histogram of patients with a defined prognosis, the image in consideration and the validation image are excluded when generating these models.

The use of pre-computed descriptor labels are chosen to reduce the computational requirement during classification.

3.6 Proposed system

The proposed system is designed to predict prognosis of new patients based on prior knowledge from a training set, an overview of the proposed system is illustrated in figure 3.11. System inputs consists of a new *test image* with unknown prognosis and a training set with corresponding prognostic follow-up information for the patients.

As images vary in both brightness and colour, image normalization is implemented to reduce the effect of these variations, section 3.1. Image descriptors are computed for all pixels using LBP and local variance, section 3.3. Masks are automatically created to identify regions consisting of tissue and the urothelial layer, 3.2. Features are then computed as the chi-squared of the masked descriptor histograms and models histograms, section 3.4. The model histograms for each feature are then created using the medical follow-up information and descriptor histograms, presented in section 3.4.1. The training images are however excluded from the model histogram generation when computing features for the image itself. Classification and prediction of the new image, is done using the RUSBoost classifier, described in section 3.5.

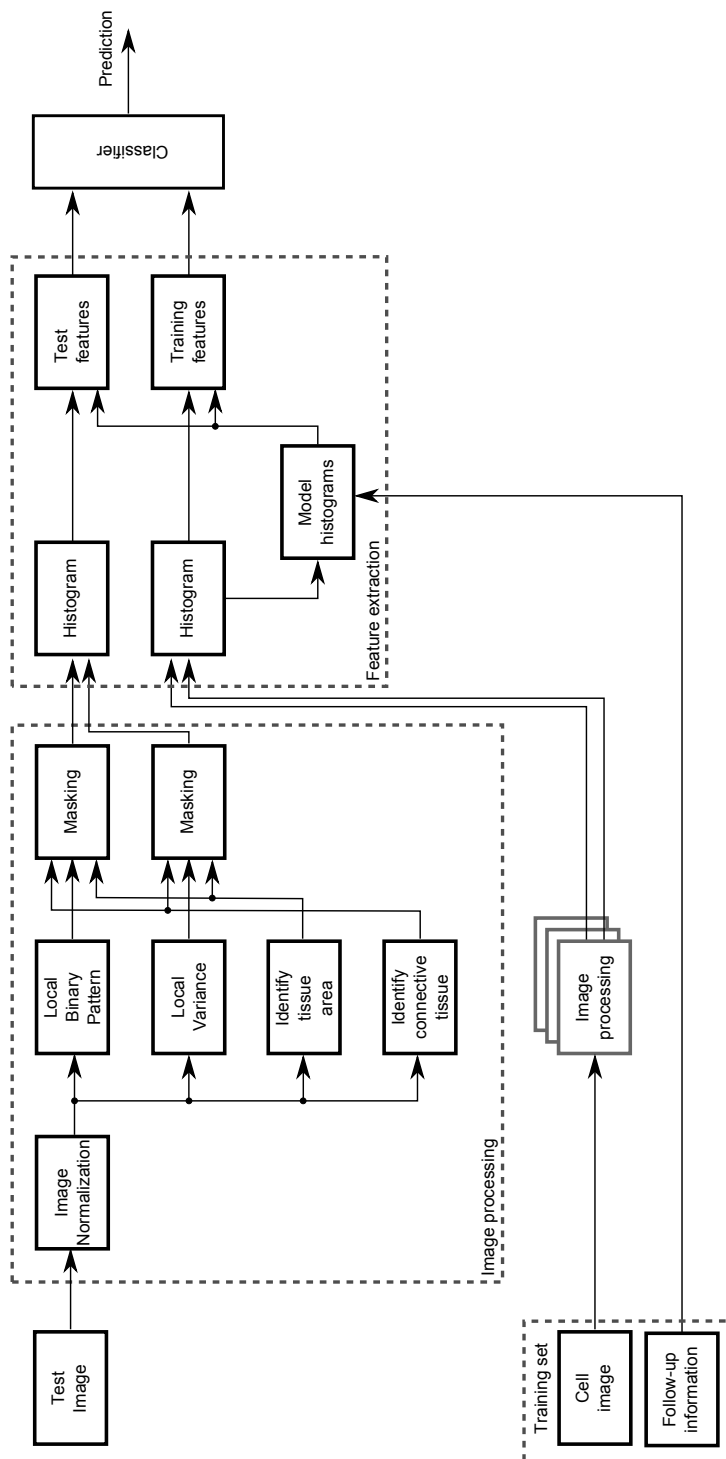


Figure 3.11: Proposed system

4. Experiments and results

This chapter presents results achieved for the proposed system. Validation of masks are first described, results achieved when predicting prognosis using varying masks and normalizations and then presented.

4.1 Masking

4.1.1 Identification of regions with tissue

For verification of the automatically generated masks presented in section 3.2, masks were compared with manually defined truth masks. The truth masks were created by manually defining points along the border with approximately 20-30 pixels between each point. A continuous curve was then creating using a piecewise polynomial between the points. As the border was defined by manually logging coordinates by the author of this paper, it should be taken as an approximation of the actual truth masks. As a consequence of this, only three images were used for validation. An overview of the algorithm used to generate the masks, can be found in section 3.2.

Figure 4.2 and 4.1 shows two scanned images with their corresponding cell masks error. Blue illustrates area identified as tissue by the truth mask but not by the generated mask (false negative), while red illustrates area identified as tissue by the generated mask but not by the truth mask (false positive).

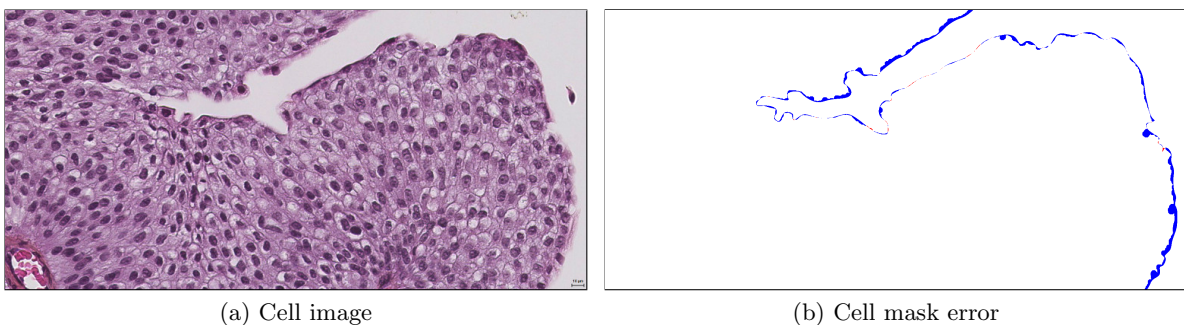


Figure 4.1: Identification of tissue: Patient ID 28

An accuracy of $\sim 99\%$ was found on all three masks tested, shown in table 4.1. Sensitivity was also high, with values of $> 99.9\%$. Small areas of false positives were observed, shown as red in figure 4.2 and 4.1. Specificity was high ($> 97.7\%$) for two of the images while it was lower for for third image (ID 28), shown as blue. The generated mask achieved a good fit in regions

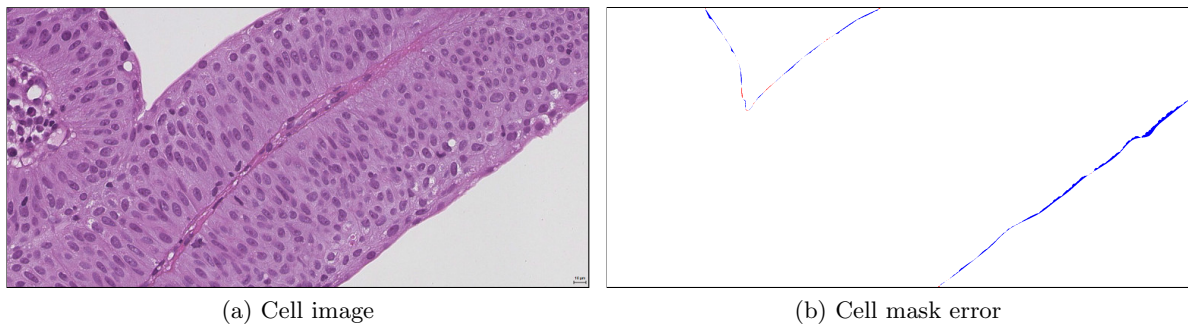


Figure 4.2: Identification of tissue: Patient ID 98

where a clear edge could be observed between tissue and areas without tissue, while it struggles in regions with a smooth transition between the two regions.

	ID28	ID 98	ID181
accuracy	0.9898	0.9965	0.9970
sensitivity	0.9998	0.9998	0.9999
specificity	0.9237	0.9826	0.9773

Table 4.1: Identification of any tissue

4.1.2 Identification of regions with connective tissue

Verification of the proposed methods used to identify connective tissue, presented in section 3.2.2, were done using the same images used for validation of tissue in section 4.1.1. Truth masks describing the connective tissue were manually defined by the author using GIMP, and were therefore considered an approximation of the actual truth masks. Evaluation of masks were done using the performance measures described in section 2.8.

For the method using multi-level thresholding, 12 levels were defined, where level 4 to 8 were used to create the mask. Holes below 3000 pixels were also removed. When using local LBP histograms, block size was set to 50×50 and MSE threshold to 0.0035. Parameters were experimentally chosen to minimize false negatives while still identifying sections with connective tissue. Images used for these experiments were not included in verification of the methods.

Figure 4.3 and 4.4 illustrates the masking error for two images using both methods. Red illustrates areas identified as relevant tissue by the automatically generated mask, but not by the truth mask (false positive). While blue illustrates areas marked as relevant by the truth, but identified as not relevant by the generated mask (false negative).

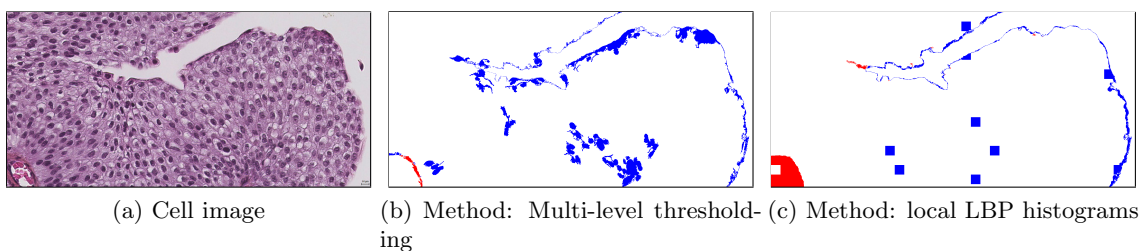


Figure 4.3: Identification of connective tissue: Patient ID 28

For patient ID 28, multi-level thresholding correctly identified most of the connective tissue, while local LBP histograms only identify parts of the connective tissue. Both methods struggled however with false negatives.

For patient ID 98, both methods correctly identified most of the large segment of connective tissue located on the left. Multi-level thresholding also correctly identified parts of the long thin section of connective tissue, while local LBP histogram struggled with identification of this region.

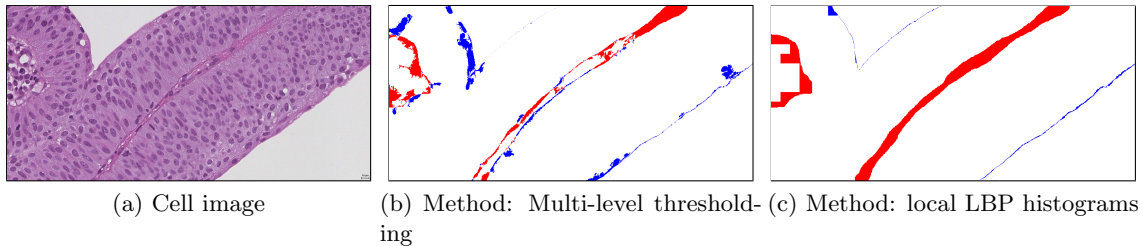


Figure 4.4: Identification of connective tissue: Patient ID 98

Performance of the three images tested are shown in table 4.2. Accuracies were in the region of 95% for patient ID 28 and 98 using both methods, while patient ID 181 scores lower with $\sim 91\%$ and $\sim 79\%$.

		ID 28	ID 98	ID 181
Multi-level thresholding	accuracy	0.9470	0.9507	0.9078
	sensitivity	0.9403	0.9698	0.9115
	specificity	0.9899	0.8965	0.8893
LBP histogram	accuracy	0.9645	0.9433	0.7844
	sensitivity	0.9740	0.9943	0.7923
	specificity	0.9027	0.7977	0.7446

Table 4.2: Identification of connective tissue

4.2 Experiments

Evaluation of system performance in the following experiments were done using classification errors of the various classes. During preliminary testing, three classes were used. Due to a high number of misclassifications of patients with recurrence and progression, illustrated in equation 4.1, a two class-problem to identify patients with recurrence from patients without recurrence was chosen.

$$C = \begin{bmatrix} 4 & 6 & 3 \\ 6 & 0 & 7 \\ 3 & 8 & 5 \end{bmatrix} \quad (4.1)$$

The use of random undersampling in the classifier, result in some variations for each computation done. Performance measures were therefore chosen illustrated as the mean of 5 computations, with the standard deviation illustrated using a paler colour around the computed mean. Due to the small size of the data set used, misclassification of one patient results in a $> 2\%$ change in

accuracy. Use of smoothing for both mean and standard deviation were therefore chosen to better illustrate the general trend in each region. Smoothing was done using a moving average filter. The filter size was experimentally chosen as 25 samples to reduce change between neighbouring number of weak learners while keeping the overall trend.

4.2.1 Pre-experiment: Choosing system parameters

To determine optimal system parameters for the LBP and local variance in combination with the optimal fitness measure, an initial pre-experiment was devised. Both single, and combinations of multiple scales using the same number of neighbours were considered. The devised system was configured using feature vector described by equation 3.6, prognosis predicted was no recurrence/recurrence.

In addition, the system was configured using *tree* and *learning rate 1* in the classifier. To avoid impact of random undersampling, the random generator was configured using the Mersenne Twister ¹ pseudorandom number generator with a defined seed. In addition, 50 weak learners were used throughout this pre-experiment.

Initial testing was done with and without prior probabilities defined in the classifier, an overall improvement of system performance was observed when using prior probabilities. As a result of this, only results with prior probabilities defined were evaluated.

Single neighbourhood

System performance using a single neighbourhood in combination with each distance measure, table 4.3, were overall low with some exceptions.

With the neighbourhood given using $P = 8$, system performance was low using the chi-squared measure. An increase in system performance was observed for radii three to five with the use of sum of squared errors, the increase was however small and only $R = 3$ and $R = 5$ achieves an accuracy above 60%. Increasing number of neighbours to $P = 16$, a higher accuracy was observed for $R = 5$ and $R = 9$ using chi-squared, and $R = 5$, $R = 9$ using SSE. Increasing the number of neighbours to $P = 24$, $R = 9$ achieves the highest accuracy. The optimal neighbourhood using chi-squared was $P = 8$ $R = 5$ with 29 correctly identified patients, and $P = 16$ $R = 7$ using SSE with 30 correctly identified patients.

Parameter	P=8	P=16	P=24	Parameter	P=8	P=16	P=24
R=3	$\begin{bmatrix} 4 & 14 \\ 9 & 15 \end{bmatrix}$	$\begin{bmatrix} 3 & 18 \\ 10 & 11 \end{bmatrix}$	$\begin{bmatrix} 2 & 18 \\ 11 & 11 \end{bmatrix}$	R=3	$\begin{bmatrix} 6 & 9 \\ 7 & 20 \end{bmatrix}$	$\begin{bmatrix} 5 & 13 \\ 8 & 16 \end{bmatrix}$	$\begin{bmatrix} 5 & 12 \\ 8 & 17 \end{bmatrix}$
R=5	$\begin{bmatrix} 5 & 16 \\ 8 & 13 \end{bmatrix}$	$\begin{bmatrix} 8 & 8 \\ 5 & 21 \end{bmatrix}$	$\begin{bmatrix} 7 & 9 \\ 6 & 20 \end{bmatrix}$	R=5	$\begin{bmatrix} 9 & 10 \\ 4 & 19 \end{bmatrix}$	$\begin{bmatrix} 6 & 10 \\ 7 & 19 \end{bmatrix}$	$\begin{bmatrix} 5 & 10 \\ 8 & 19 \end{bmatrix}$
R=7	$\begin{bmatrix} 5 & 13 \\ 8 & 16 \end{bmatrix}$	$\begin{bmatrix} 5 & 14 \\ 8 & 15 \end{bmatrix}$	$\begin{bmatrix} 7 & 14 \\ 6 & 15 \end{bmatrix}$	R=7	$\begin{bmatrix} 7 & 14 \\ 6 & 15 \end{bmatrix}$	$\begin{bmatrix} 9 & 8 \\ 4 & 21 \end{bmatrix}$	$\begin{bmatrix} 5 & 8 \\ 8 & 21 \end{bmatrix}$
R=9	$\begin{bmatrix} 5 & 16 \\ 8 & 13 \end{bmatrix}$	$\begin{bmatrix} 5 & 6 \\ 8 & 23 \end{bmatrix}$	$\begin{bmatrix} 5 & 13 \\ 8 & 16 \end{bmatrix}$	R=9	$\begin{bmatrix} 5 & 16 \\ 8 & 13 \end{bmatrix}$	$\begin{bmatrix} 7 & 10 \\ 6 & 19 \end{bmatrix}$	$\begin{bmatrix} 10 & 11 \\ 3 & 18 \end{bmatrix}$

(a) Distance measure: Chi-squared

(b) Distance measure: SSE

Table 4.3: Pre-experiment: System performance using a single scale

¹A widely used pseudorandom number generator [34]

Combining neighbourhoods

A multiresolution analysis was then considered, combining various neighbourhood and radii for an improved classification result with each distance measure. This experiment combines two, three and four radii using the same number of neighbours.

Using a combination of two scales, six possible combination exists for each neighbourhood, table table 4.4. The optimal combinations using chi-squared were $R = \{3, 5\}$ and $R = \{5, 9\}$ in combination with $P = 8$ with 28 correctly classified patients. A reduction of 1 correctly identified patients, compared to the 29 patients correctly identified using a single scale. The optimal combination using SSE achieves 30 correctly identified patients using the parameters $P = 8$ $R = \{3, 7\}$, corresponding to the accuracy found using only a single scale.

An overall increase in system performance were also observed when using sum of squared errors compared to chi-squared. A total of seven combinations achieves more than 26 correctly classified patients, compared to only two combinations achieving the same using chi-squared.

Parameter	P=8	P=16	P=24	Parameter	P=8	P=16	P=24
$R=\{3,5\}$	$\begin{bmatrix} 6 & 7 \\ 7 & 22 \end{bmatrix}$	$\begin{bmatrix} 4 & 14 \\ 9 & 15 \end{bmatrix}$	$\begin{bmatrix} 6 & 13 \\ 7 & 16 \end{bmatrix}$	$R=\{3,5\}$	$\begin{bmatrix} 6 & 8 \\ 7 & 21 \end{bmatrix}$	$\begin{bmatrix} 4 & 10 \\ 9 & 19 \end{bmatrix}$	$\begin{bmatrix} 5 & 16 \\ 8 & 13 \end{bmatrix}$
$R=\{3,7\}$	$\begin{bmatrix} 4 & 8 \\ 9 & 21 \end{bmatrix}$	$\begin{bmatrix} 6 & 10 \\ 7 & 19 \end{bmatrix}$	$\begin{bmatrix} 5 & 15 \\ 8 & 14 \end{bmatrix}$	$R=\{3,7\}$	$\begin{bmatrix} 4 & 3 \\ 9 & 26 \end{bmatrix}$	$\begin{bmatrix} 5 & 8 \\ 8 & 21 \end{bmatrix}$	$\begin{bmatrix} 5 & 13 \\ 8 & 16 \end{bmatrix}$
$R=\{3,9\}$	$\begin{bmatrix} 1 & 8 \\ 12 & 21 \end{bmatrix}$	$\begin{bmatrix} 4 & 8 \\ 9 & 21 \end{bmatrix}$	$\begin{bmatrix} 5 & 12 \\ 8 & 17 \end{bmatrix}$	$R=\{3,9\}$	$\begin{bmatrix} 4 & 11 \\ 9 & 18 \end{bmatrix}$	$\begin{bmatrix} 4 & 7 \\ 9 & 22 \end{bmatrix}$	$\begin{bmatrix} 6 & 14 \\ 7 & 15 \end{bmatrix}$
$R=\{5,7\}$	$\begin{bmatrix} 6 & 11 \\ 7 & 18 \end{bmatrix}$	$\begin{bmatrix} 4 & 14 \\ 9 & 15 \end{bmatrix}$	$\begin{bmatrix} 4 & 12 \\ 9 & 17 \end{bmatrix}$	$R=\{5,7\}$	$\begin{bmatrix} 7 & 12 \\ 6 & 17 \end{bmatrix}$	$\begin{bmatrix} 6 & 10 \\ 7 & 19 \end{bmatrix}$	$\begin{bmatrix} 5 & 9 \\ 8 & 20 \end{bmatrix}$
$R=\{5,9\}$	$\begin{bmatrix} 6 & 7 \\ 7 & 22 \end{bmatrix}$	$\begin{bmatrix} 5 & 13 \\ 8 & 16 \end{bmatrix}$	$\begin{bmatrix} 4 & 13 \\ 9 & 16 \end{bmatrix}$	$R=\{5,9\}$	$\begin{bmatrix} 8 & 13 \\ 5 & 16 \end{bmatrix}$	$\begin{bmatrix} 6 & 9 \\ 7 & 20 \end{bmatrix}$	$\begin{bmatrix} 5 & 6 \\ 8 & 23 \end{bmatrix}$
$R=\{7,9\}$	$\begin{bmatrix} 5 & 13 \\ 8 & 16 \end{bmatrix}$	$\begin{bmatrix} 4 & 9 \\ 9 & 20 \end{bmatrix}$	$\begin{bmatrix} 4 & 13 \\ 9 & 16 \end{bmatrix}$	$R=\{7,9\}$	$\begin{bmatrix} 7 & 12 \\ 6 & 17 \end{bmatrix}$	$\begin{bmatrix} 7 & 10 \\ 6 & 19 \end{bmatrix}$	$\begin{bmatrix} 5 & 12 \\ 8 & 17 \end{bmatrix}$

(a) Distance measure: Chi-squared

(b) Distance measure: SSE

Table 4.4: Pre-experiment: System performance using two scales

Combining additional scales, table 4.5, does not offer an increased accuracy compared to results obtained using a single or a combination of two scales. The neighbourhood $P = 16$ outperforms both $P = 8$ and $P = 24$ using combinations of multiple scales, with a maximum accuracy of $\sim 66\%$ found using chi-squared and $\sim 64\%$ found using SSE. The results were however reduced from the optimal results using a combinations of fewer scales.

Parameter	P8	P16	P24	Parameter	P8	P16	P24
$R=\{3,5,7\}$	$\begin{bmatrix} 3 & 13 \\ 10 & 16 \end{bmatrix}$	$\begin{bmatrix} 5 & 10 \\ 8 & 19 \end{bmatrix}$	$\begin{bmatrix} 7 & 15 \\ 6 & 14 \end{bmatrix}$	$R=\{3,5,7\}$	$\begin{bmatrix} 6 & 12 \\ 7 & 17 \end{bmatrix}$	$\begin{bmatrix} 6 & 8 \\ 7 & 21 \end{bmatrix}$	$\begin{bmatrix} 7 & 15 \\ 6 & 14 \end{bmatrix}$
$R=\{5,7,9\}$	$\begin{bmatrix} 4 & 10 \\ 9 & 19 \end{bmatrix}$	$\begin{bmatrix} 6 & 9 \\ 7 & 20 \end{bmatrix}$	$\begin{bmatrix} 4 & 13 \\ 9 & 16 \end{bmatrix}$	$R=\{5,7,9\}$	$\begin{bmatrix} 7 & 12 \\ 6 & 17 \end{bmatrix}$	$\begin{bmatrix} 7 & 10 \\ 6 & 19 \end{bmatrix}$	$\begin{bmatrix} 5 & 10 \\ 8 & 19 \end{bmatrix}$
$R=\{3,5,7,9\}$	$\begin{bmatrix} 4 & 10 \\ 9 & 19 \end{bmatrix}$	$\begin{bmatrix} 5 & 6 \\ 8 & 23 \end{bmatrix}$	$\begin{bmatrix} 4 & 12 \\ 9 & 17 \end{bmatrix}$	$R=\{3,5,7,9\}$	$\begin{bmatrix} 4 & 12 \\ 9 & 17 \end{bmatrix}$	$\begin{bmatrix} 6 & 9 \\ 7 & 20 \end{bmatrix}$	$\begin{bmatrix} 5 & 11 \\ 8 & 18 \end{bmatrix}$

(a) Distance measure: Chi-squared

(b) Distance measure: SSE

Table 4.5: Pre-experiment: System performance using multiple scales

Multi-scale weighting

Results in the previous section were obtained using a uniform weight for on each scale in the combination. The use of individual weights for each scale were considered interesting, as the optimal combination of each scale were unknown.

This experiment was devised to explore the possibility if individual weights could be used to increase system performance. A weight of $[0, 1]$ were added to each scale, resulting in a high number of possible weighting scenarios for combinations of n scales. Due to the computational requirement for this analysis, the study focused on combinations obtaining good results using uniform weighted combinations.

While some of the tested combinations did not increase system performance, i.e. $P = 8$, $R = \{5, 9\}$. Other combinations $P = 8$, $R = \{3, 7\}$, equation 4.2, offered a small increase in specificity without a change in sensitivity. Accuracy for the system was increased from $\sim 71\%$ to $\sim 74\%$

$$\begin{array}{ccc} \begin{bmatrix} 4 & 3 \\ 9 & 26 \end{bmatrix} & \xrightarrow{\text{weights}} & \begin{bmatrix} 5 & 3 \\ 8 & 26 \end{bmatrix} \\ & & P = 8, R = \{3, 7\} \end{array} \quad (4.2)$$

The highest accuracy using a combination of two scales were found using $P = 8$ $R = \{3, 5\}$ with $\sim 76\%$, equation 4.3. An increase of $\sim 10\%$ compared to no weighting.

$$\begin{array}{ccc} \begin{bmatrix} 6 & 7 \\ 7 & 22 \end{bmatrix} & \xrightarrow{\text{weights}} & \begin{bmatrix} 7 & 4 \\ 6 & 25 \end{bmatrix} \\ & & P = 8, R = \{3, 5\} \end{array} \quad (4.3)$$

Classifier parameters

Based on preliminary testing, parameters were chosen for the following experiments, table 4.6. Parameters the descriptors were chosen as $P8$ and $R = \{3, 5\}$ in combination with chi-squared, as this combination achieved the highest accuracy and sensitivity in the preliminary testing. The ideal number of weak learners were however unknown, an analysis using a wide range of learners were therefore chosen to visualize the ideal region in each experiment. Prior probabilities were set for each experiment using known information from the data set. Due to the computational requirement for each analysis using a wide range of weak learners, a learning rate of 1 was chosen.

	Parameter	Value
Classifier	Weak learner type	Tree
	Number of weak learners	Range: 1-200
	Learning rate	1
	Priors	Based on data set
	Cost	None
Features	Descriptor	LBP, Var
	Neighbours	8
	Radius	r1=3, r2=5
	Weights	w1=0.53, w2=0.47
	Distance measure	Chi-squared

Table 4.6: Classifier parameters

In addition, a full crossvalidation were chosen to be done for each number of weak learners used.

4.2.2 Experiment 1: Recurrence

This experiment was devised to identify if the presented descriptors could be used to predict recurrence in patients. The analysis was based on features computed using local binary pattern and local variance. Results using a combination of both descriptors will be presented first, before each descriptor is presented individually.

Classification error of patients without recurrence is shown in red, classification error of patients with recurrence in blue and overall classification error of the system in black. The lighter colours illustrate the standard deviation for its corresponding line.

Features: LBP and VAR

When using six features, defined in equation 3.6, the classification errors shown in figure 4.5a were obtained. Identification of patients with recurrence achieved a minimum mean classification error of approximately 27% when using approximately 40 weak learners. Misclassification of patients without recurrence as recurrence stays in the region of 55 – 60% independent of number of weak learners used. Using four features, equation 3.8, classification error are shown in figure 4.5b. A high prediction of recurrence was observed independent of true prognosis of the patients, resulting in a sensitivity of approximately 70% and a specificity of approximately 30%.

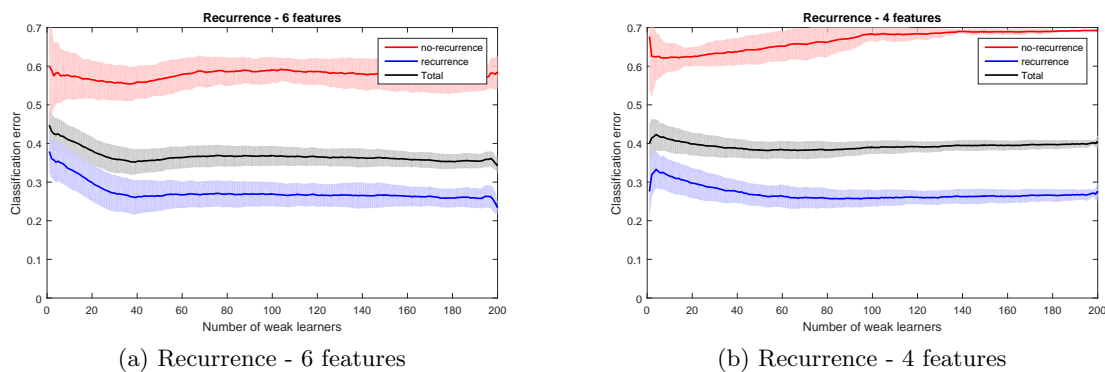


Figure 4.5: Experiment 1: Recurrence, multiple descriptors

Features: LBP

The use of a single descriptor was chosen to study performance of each descriptor individually. The proposed system remains unchanged from the previous analysis, with the only exception of now using a reduced feature vector. Classification errors using features computed from LBP are illustrated in figure 4.6.

Classification using three features, defined in equation 3.7, are shown in figure 4.6a. An accuracy of 65 – 70% was observed using 75 weak learners. Sensitivity for the system was 80% using 75 weak learners, corresponding to 4 – 5 patients with recurrence misclassified as patients without recurrence. Specificity on the other hand stays in the area of 50 – 55%, independent of the

number of weak learners used. Overall performance of the system was improved compared to using both feature descriptors, figure 4.5a.

Classification errors using two features, defined in equation 3.9, are shown in figure 4.6b. A high prediction of recurrence, independent of true prognosis of the patient were observed.

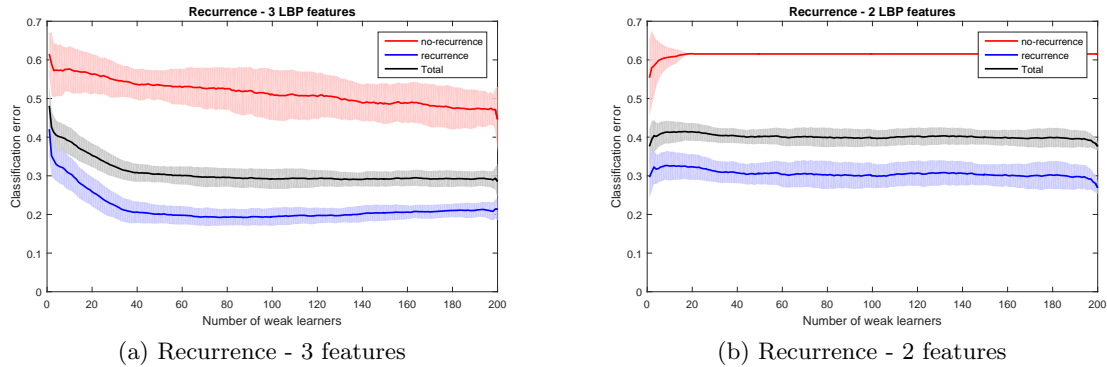


Figure 4.6: Experiment 1: Recurrence, LBP

Features: VAR

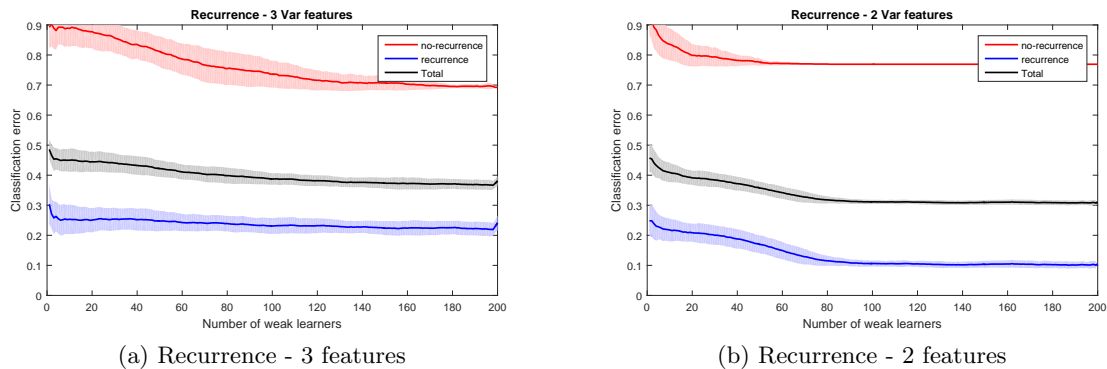


Figure 4.7: Experiment 1: Recurrence, local variance

With the use of three features computed from local variance, defined in equation 3.7, classification errors shown in figure 4.7a were obtained. And the use of two features computed from local variance, defined in equation 3.9, are shown in figure 4.7b. A high prediction rate of recurrence was observed with use of both feature vectors, with approximately 80% of all patients predicted as recurrence independent of true prognosis.

4.2.3 Experiment 2: Progression

Early identification of patients with progression are desired, as this information could lead to an increased follow-up regime to detection new tumours at an early stage. Two possible scenarios of progression were considered, recurrence with a stage higher than originally diagnosed and recurrence with a muscle invasive tumour. Where the former scenario include progression from Ta to T1 in addition to muscle-invasive progression. As patients with both recurrence and

progression require a higher follow-up regime than patients without recurrence, this experiment was devised to study if patients with progression could be distinguished from those without recurrence. As a result of this, patients with recurrence but not progression were excluded from this experiment.

Classification error of patients without recurrence is shown in red, classification error of patients with progression in blue and overall classification error of the system in black. The lighter colours illustrate the standard deviation for its corresponding line.

Any progression

The data set contains a total of 15 patients with progression, and 13 patients without recurrence. Where six patients with progression were originally diagnosed with Ta, while nine were originally diagnosed with T1. System performance using four features, defined in equation 3.10, are illustrated in figure 4.8.

A high prediction rate of patients as progression, independent of true prognosis of the patients were observed. With 85% of patients belonging to this class and 70% of patients without progression predicted as this. No correlation between original stage of the tumour and prediction were observed.

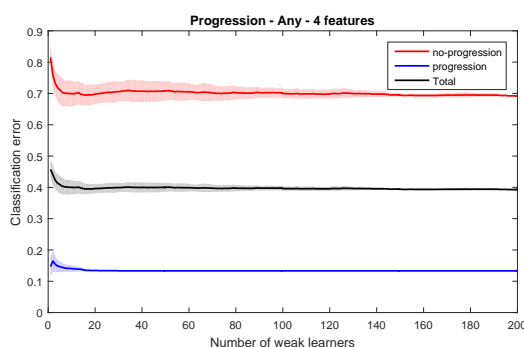


Figure 4.8: Experiment 2: Any progression, multiple descriptors

Muscle-invasive progression

By removing patients with non-invasive progression, the data set was reduced from 28 to 26 patients. 13 without recurrence and 13 with muscle-invasive progression. Of patients with progression four patients were originally diagnosed with Ta, while nine originally diagnosed with T1. System performance using four features, defined in equation 3.10, is illustrated in figure 4.9.

A high prediction rate of patients as progression, independent of true prognosis were observed. With 60% of patients belonging to this class, and 80% of patients without recurrence predicted as this. No correlation between original stage of the tumour and prediction was observed.

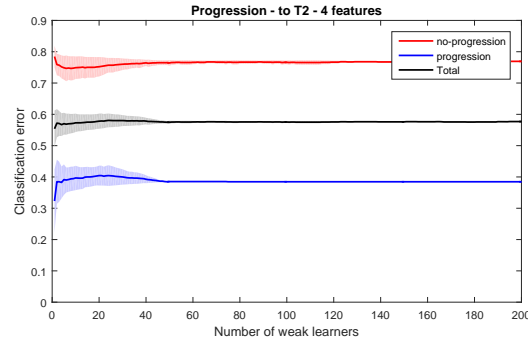


Figure 4.9: Experiment 2: Muscle invasive progression, multiple descriptors

4.2.4 Experiment 3: Cost functions

Misclassification of patients with recurrence or progression as no-recurrence is considered costly as this result in insufficient treatment of the patient. While the opposite misclassification result in unnecessary treatment of the patient, which is a costly procedure for the hospital but not a life-threatening situation for the patient. This experiment introduces a cost function to reduce the misclassification of patients with recurrence as no-recurrence.

The cost for misclassifying resulting in insufficient treatment were set five times the cost of misclassifying resulting in unnecessary treatment, equation 4.4. This ratio was chosen through preliminary experiments to obtain a sufficient reduction in the costly misclassification, without unnecessary decrease in specificity. The preliminary experiments were done with varying cost using 50 weak learners.

$$cost = \begin{bmatrix} 0 & 5 \\ 1 & 0 \end{bmatrix} \quad (4.4)$$

Classification error of patients without recurrence are shown in red, classification error of patients with recurrence in blue and overall classification error of the system in black. The lighter colours illustrate the standard deviation for its corresponding line. Solid lines represents results obtained using a cost function, while the dashed line represents results obtained without a cost function defined.

Classification error using six features, defined in equation 3.6, are shown in figure 4.10a. A 10% reduction in misclassification of patients with true recurrence were observed, while misclassification of patients without recurrence stays in the same region as found in experiment 1. Classification error using four features, defined in equation 3.8, are shown in figure 4.10b. An increase of approximately 5% in sensitivity was observed, while specificity stays within the standard deviation of the result found in experiment 1.

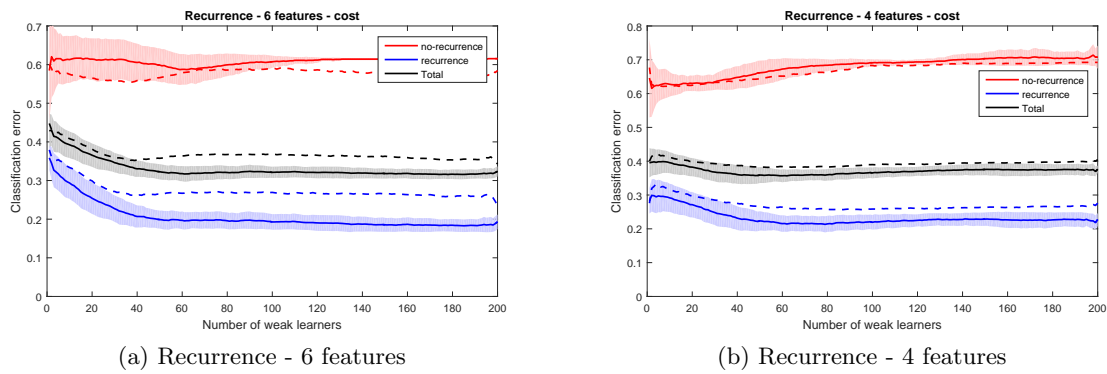


Figure 4.10: Experiment 3: Recurrence, multiple descriptors using cost function

Using each descriptor separately, a reduced feature vector was obtained, defined in equation 3.7. When using the LBP operator, figure 4.11a, accuracy stays within the standard deviation of the result found in experiment 1. A small improvement of 2–3% were observed for sensitivity, while specificity were reduced for all weak learners tested. For local variance, accuracy stays in the region of 40% independent of the number of weak learners used, figure 4.11b. A high prediction of recurrence was observed independent of true prognosis of patients.

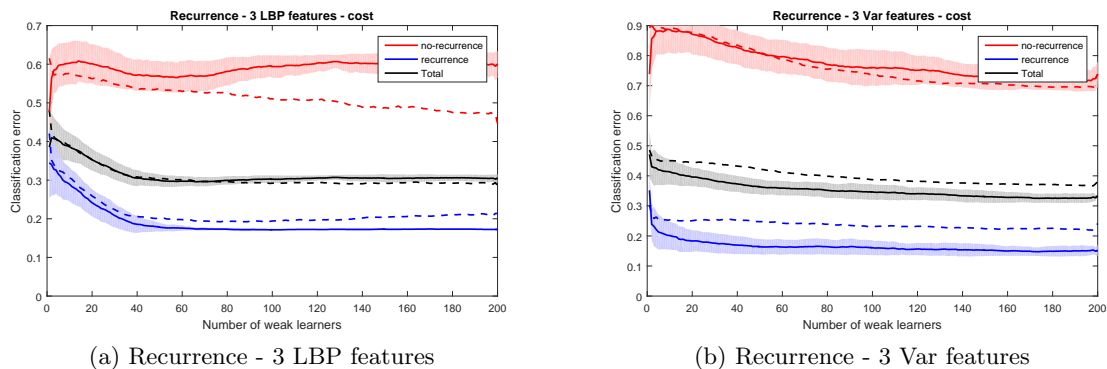


Figure 4.11: Experiment 3: Recurrence, single descriptors using cost function

Setting the cost of misclassifying patients with no-recurrence as recurrence five times higher than recurrence as no-recurrence, equation 4.5, and features defined in equation 3.6. The result shown in figure 4.12 was obtained.

$$cost = \begin{bmatrix} 0 & 5 \\ 1 & 0 \end{bmatrix} \quad (4.5)$$

Misclassification of patients without recurrence were reduced, while misclassification of patients with recurrence were increased. Due to the size of class 2 (29 patients), accuracy for the system was reduced to approximately 50%.

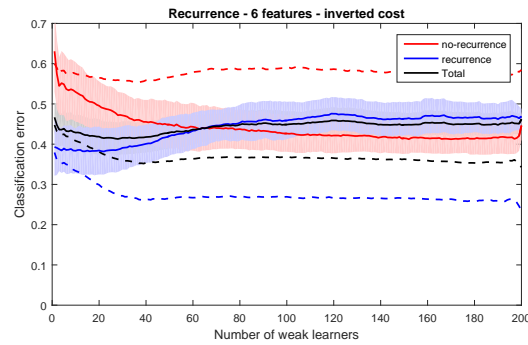


Figure 4.12: Experiment 3: Recurrence, multiple descriptors using transposed cost function

4.2.5 Experiment 4: Masking - removing suburothelial connective tissue

With the original masking algorithm used in the previous experiments, all tissue in the scanned images were used when computing features. The idea behind this experiment was to remove the suburothelial connective tissue from the analysis as the tumour itself originates in the urothelial layer. Suburothelial connective tissue could therefore be seen as not relevant regions for the analysis. Masks describing regions with connective tissue were manually defined by the author using *GIMP*¹. As a consequence of this, masks were considered an approximation of the truth masks describing connective tissue. Masks were then incorporated into the proposed system as an extension of the original masks.

Figure 4.13 illustrates case ID 98 with all tissue preserved as used in previous experiments, and only the urothelial layer preserved.

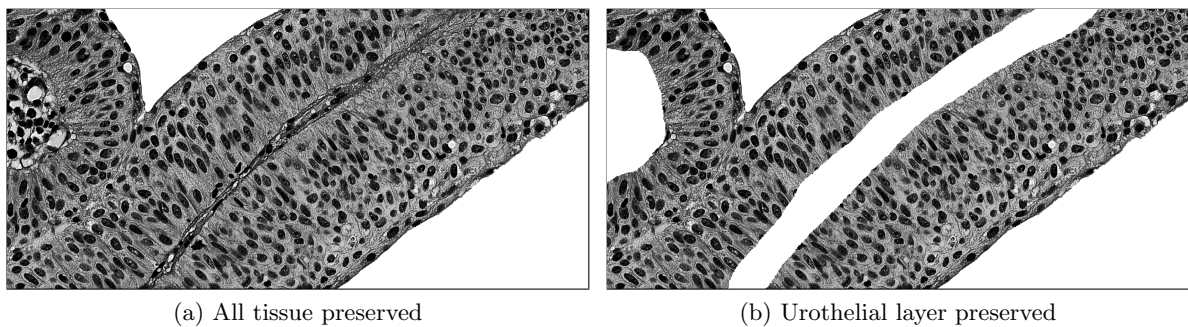


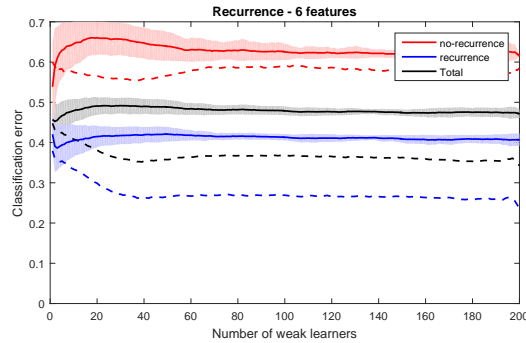
Figure 4.13: Experiment 4: Removing connective tissue (case ID: 28)

Classification errors of patients without recurrence are shown in red, classification error of patients with recurrence in blue and overall classification error of the system in black. The lighter colours illustrate the standard deviation for its corresponding line. Solid lines represent results obtained using only the urothelial layer, while the dashed line represents results obtained using all tissue.

¹GNU Image Manipulation Program

Recurrence

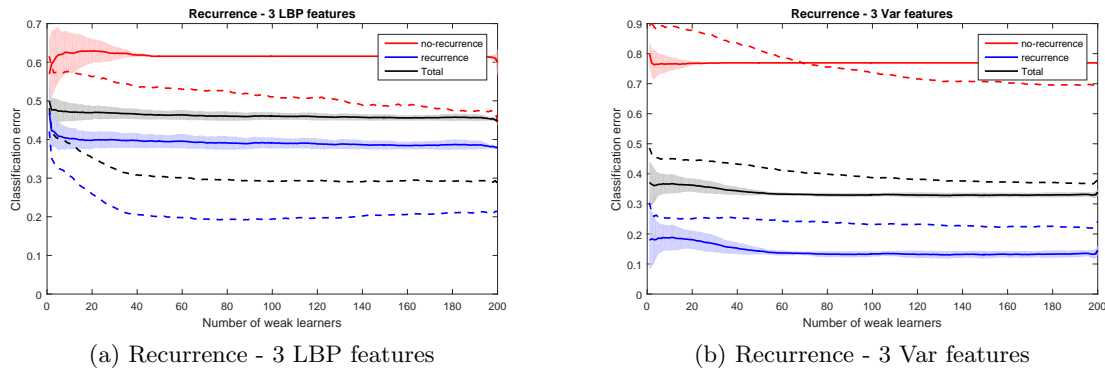
Identification of patients with recurrence using six features extracted from the urothelial layer, defined in equation 3.6, are shown in figure 4.14a. Results using both feature descriptors, defined in equation 3.6, are shown in figure 4.14. An increased misclassification rate of all patients were observed, resulting in a lower accuracy, independent of the number of weak learners used.



(a) Recurrence - 6 features

Figure 4.14: Experiment 4: Recurrence, multiple descriptors using alternative masking

Classification using features from a single descriptor, defined in equation 3.7, are shown in figure 4.15. Misclassification of both classes was increased when using only LBP, resulting in an accuracy of approximately 50% independent of the number of weak learners. For local variance, a high prediction rate of recurrence was observed, with approximately 80% of all patients predicted as this class independent of true prognosis of the patient.



(a) Recurrence - 3 LBP features

(b) Recurrence - 3 Var features

Figure 4.15: Experiment 4: Recurrence, single descriptors using alternative masking

Progression

Early identification of patients with progression was desired, as described in experiment 2. This experiment identifies patients with progression from no-recurrence using only the urothelial layer, and the feature vector were defined using equation 3.10. The data set was reduced corresponding to what was done in experiment 2, with a total of 28 patients included in the analysis for any progression and 26 patients included in the analysis for muscle-invasive progression. The classification errors achieved, are shown in figure 4.16.

A increase in misclassification of patients with progression were observed for both analysis, while misclassification of patients without recurrence stays in the same regions as in experiment 1. Resulting in an accuracy of 45% for any progression and 35% for muscle-invasive progression.

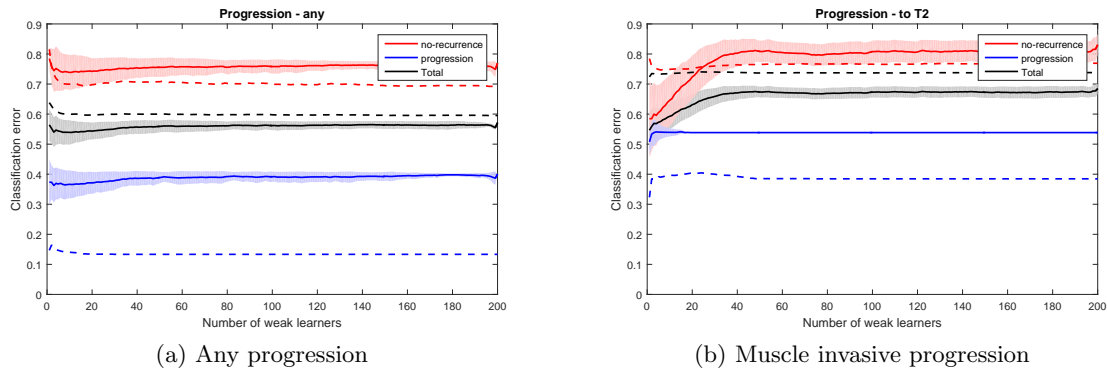


Figure 4.16: Experiment 4: Progression, multiple descriptors using alternative masking

4.2.6 Experiment 5: Preprocessing

Preprocessing of images in the previous experiments were done using histogram equalization. The method resulted in a visually good normalization across multiple images, with nucleons and cytoplasm being similar independent of original contrast in the image. While the LBP operator is not affected by monotonic transformations of the gray scale, and will thus be unaffected by this equalization. The local variance operator will be affected. This experiment was devised to test if an alternative preprocessing method could be used to improve identification of patients with recurrence and progression.

Preprocessing of images were in this experiment changed from the histogram normalization method used in the proposed system, illustrated in figure 4.17a, to a simple $[0, 1]$ normalization, illustrated in figure 4.17b. For more information of the two methods, see section 3.1.

Classification errors of patients without recurrence are shown in red, classification errors of patients with recurrence in blue and overall classification errors of the system in black. The lighter colours illustrate the standard deviation for its corresponding line. Solid lines represent results obtained using a $[0, 1]$ normalization, while the dashed line represents results obtained using histogram equalizing.

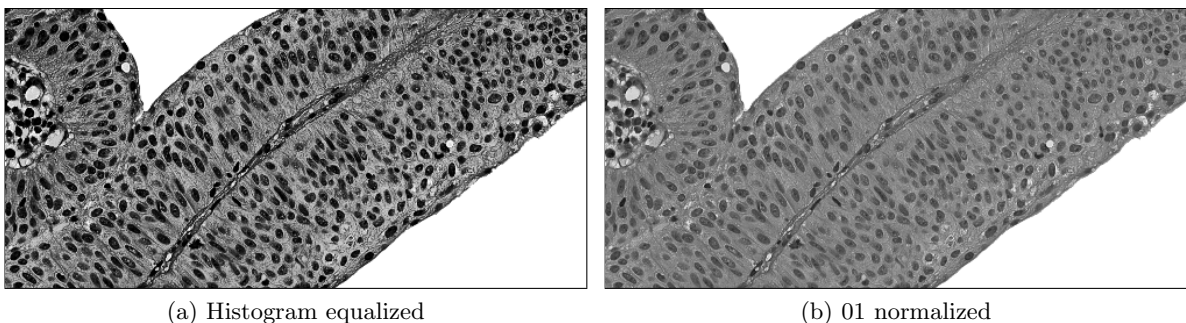


Figure 4.17: Experiment 5: Preprocessing alternatives (case ID: 28)

Recurrence

System performance using six features, defined in equation 3.6, is shown in figure 4.18. An improvement of accuracy was observed for all numbers of weak learners used, with a maximum value of approximately 70%. Misclassification of patients with recurrence were reduced from approximately 30% to below 20%. While misclassification of patients without recurrence were within the standard deviation, and can be considered a result of variations in the random undersampling.

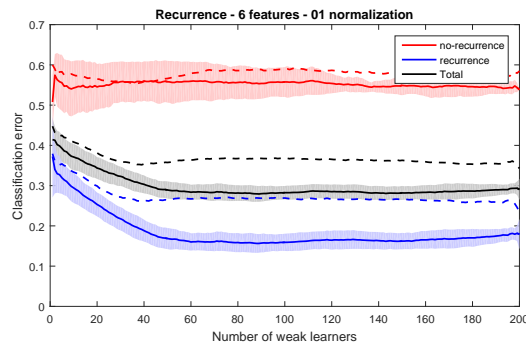


Figure 4.18: Experiment 5: Recurrence, multiple descriptors using alternative preprocessing

Results using three features from the LBP operator alone, defined in equation 3.7, are shown in figure 4.19a. A change in misclassifications of patients without recurrence were observed. The LBP operator is however invariant of monotonic transformation in the grayscale, and the change is therefore considered a result of variations in the random undersampling.

A larger difference was found using three features computed using local variance, defined in equation 3.7, shown in figure 4.19b. In experiment 1, a high prediction of recurrence independent of true prognosis was found. With the new preprocessing method, a misclassification error of approximately 40% were however found for both classes when using more than 40 weak learners.

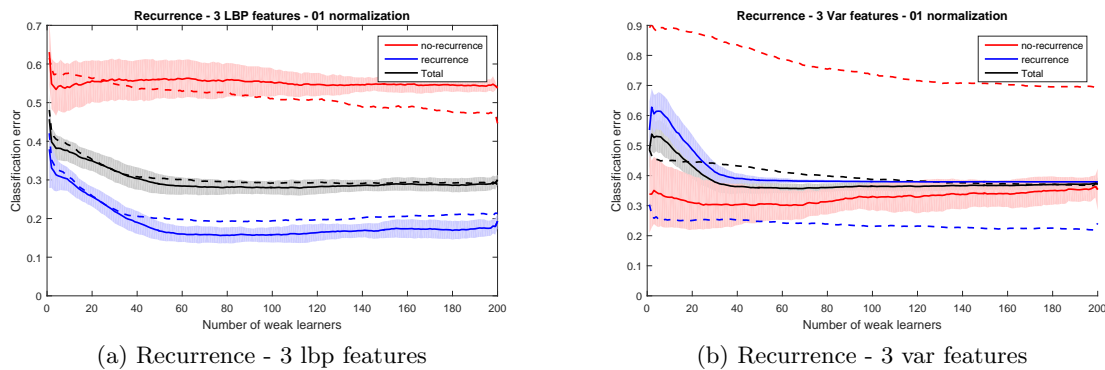


Figure 4.19: Experiment 5: Recurrence, single descriptors using alternative preprocessing

Progression

Early identification of patients with progression was desired, as described in experiment 2. This experiment was devised to identify patients with progression from patients without recurrence using a $[0, 1]$ grayscale normalization. Feature vectors were defined using equation 3.10. The

data set was reduced corresponding to what was done in experiment 2, with a total of 28 patients included in the analysis for any progression and 26 patients included in the analysis for muscle-invasive progression.

A high prediction of progression independent of true prognosis was observed for both analysis shown in figure 4.20. Where accuracy for any progression stays in the area of 60%, and accuracy for muscle-invasive progression in the area of 50%.

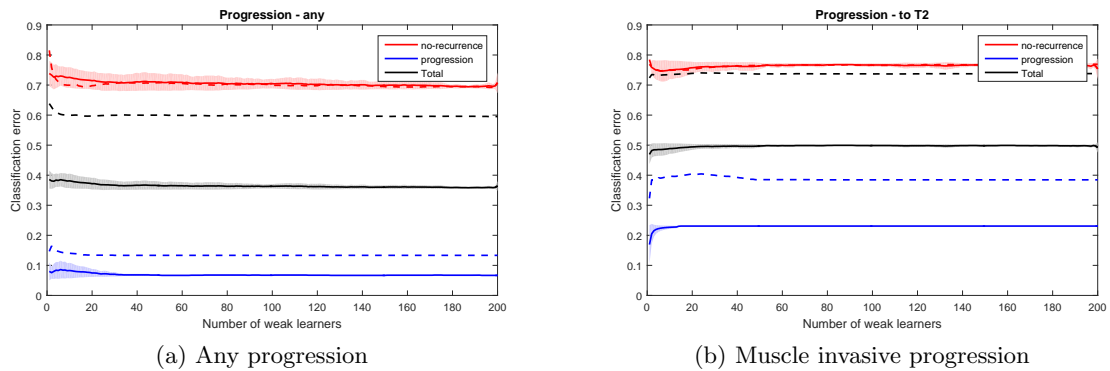


Figure 4.20: Experiment 5: Progression, multiple descriptors using alternative normalization

5. Discussion

The best results for predicting recurrence was found with the use of a $[0, 1]$ grayscale normalization in experiment 5. An identification of 83% was achieved for patients with recurrence and an identification of 45% for patients without recurrence with the use of both feature descriptors.

Prediction of progression was conducted in experiment 2, class separation using either of the texture descriptors were however low. Alternative weighting, normalization and masking were studied in experiment 3-5, but no improvement in performance was observed.

5.1 Data set

The data set was chosen in collaboration with Stavanger University Hospital as a result of its documentation and use in previous research. A larger data set was desired, but the 42 patients in this data set with equal distribution between classes were considered a good starting point for this analysis. Due to the large file sizes, smaller segments were chosen in collaboration with a pathologist to illustrate each patient.

The use of these segments could affect our texture analysis, as the chosen regions does not necessarily result in the best images for local texture analysis. Two possible approaches were considered interesting, exporting larger segments, or additional images using other regions of the scanned tissue. These additional images could aid in construction of the model histograms, and aid in the prediction as other regions could add vital information to the analysis.

5.2 System parameters

Preliminary experiments were run to determine system parameters as a base for further analyses. Initial testing was done to separate all three prognoses. Due to the high misclassification error between patients with recurrence and progression, a two-class problem was instead chosen to predict either recurrence or progression. To remove the impact of random undersampling, a fixed seed was defined in the pseudorandom number. An increase in accuracy was found with the use of higher number of weak learners. A test was done using 1, 10, 25, 50, 100 and 200 weak learners, but no improvement was found when using more than 50 weak learners. This number of weak learners was therefore chosen for the preliminary experiments.

Descriptor parameters with the highest accuracy and sensitivity were chosen for a further analysis, with the aim of testing if a weighted combination of parameters could further improve the classification. While improvements were found for some combinations, no changes were observed for others. As only combinations with a high accuracy using uniform weights were tested, it is

therefore uncertain if other combinations could further improve the results. In addition, combinations using multiple neighbourhoods were excluded from this experiment. System parameters for the following experiments were then chosen as the parameters achieving the highest accuracy during this testing.

It should be noted that multiple challenges using our approach exists. While the combination found using 50 weak learners achieved the highest accuracy, it is uncertain if this combinations will also perform best with other numbers of weak learners. Using additional weak learners during the preliminary experiments were considered, but discarded as the combination of weak learners resulting in a global optimal combination were unknown. A second challenge was the prognosis used to define classes in these experiments. As the best combination for predicting recurrence are not necessarily the best combination for predicting progression. Results achieved when identifying patients with progression from patients without recurrence were low, independent of masking and normalization used. The preliminary experiment should therefore been expanded to also identify useful combination for prediction of this prognosis.

5.3 Feature extraction

In the implemented system, each image were described by a histogram obtained using either LBP or local variance. Labels were computed for all pixels, where regions identified as relevant by generated masks were used to create the histograms. Features describing the image were then computed as the chi-squared of the image histogram, and predefined models. Defining optimal models for use in the analysis were however challenging. The use of images consisting of healthy cells were considered to allow classification based on differentiation from healthy cells. This direction was however discarded, as all images in the data set consists of tumours, and would therefore differ from the norm. Instead, models using the medical follow-up information available were chosen. Models were defined as the mean histogram of all patients with a known prognosis. Multiple features will thus be defined for each image, where they describe similarity of the image texture and each of the prognoses models. As histograms varied between patients with a defined prognosis, the standard deviation of each bar in the histogram were also used in the calculation of chi-squared.

This principle for defining features does however have a drawback, as the impact of texture variations within a single image will be minimized. As these variations might only contain a relative few number of pixels compared to the total image, information vital to the prediction could be lost. A possible solution to this was considered to be using a histogram of smaller blocks instead of the entire image. A small experiment was done using two blocks of 100×100 pixels with different textures from case ID 98, shown in figure 5.1a and 5.1b. Uniform rotation invariant LBP labels were computed using the parameters $P = 8$ and $R = 3$. Histograms for each block and of all tissue in image 98 are shown in figure 5.1c. A similarity was observed between block 1 and the total histogram, while block 2 has a clear difference on bin 5, 6 and 10. The experiment illustrates that smaller blocks could be useful to identify smaller regions with larger differentiations, and with that a potential increase in accuracy for the classification.

5.4 Preprocessing

Two preprocessing steps were implemented in the proposed system, a grayscale conversion and a normalization of the grayscale levels. As colour in images depend on the amount of HES

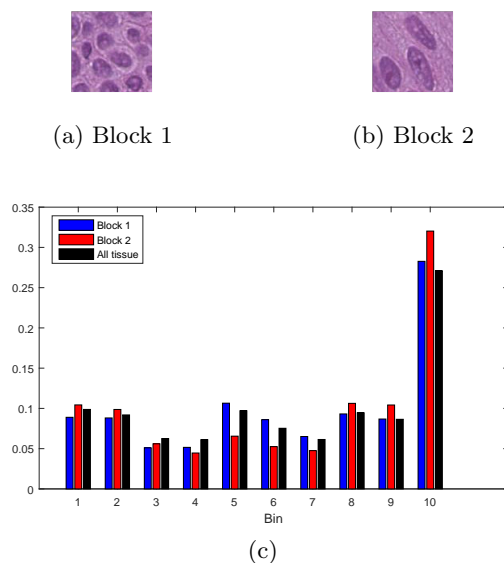


Figure 5.1: Block wise LBP, comparison

used, grayscale conversion was not considered problematic. Variation in brightness between images were also observed. This could be caused by variations in thickness of the samples, and the amount of HES used. A normalization between images were therefore desired. Two approaches of image normalization were considered and implemented, histogram equalizing and a $[0, 1]$ grayscale expansion. Histogram equalizing was chosen as this method resulted in visually similar images.

Neither of these normalizations affect LBP, as the descriptor is invariant of monotonic transformations in the grayscale levels. Differences were however observed for the local variance descriptor. Using histogram equalizing, a high prediction of recurrence independent of true prognosis of the patients were observed. While using a linear grayscale expansion, described in experiment 5, results were improved for both sensitivity and specificity. The improvements were considered to be a result of the linear transform of the grayscale.

5.5 Masking

Two masking scenarios were considered in this analysis, removing areas not covered by any tissue and removing connective tissue. The proposed algorithm for removing areas not covered by any tissue achieved good results, with an accuracy of $\sim 99\%$ for all three images used in the validation. The algorithm worked well when there was a clear edge between tissue and areas with no tissue, but struggles with smooth edges (gradients). In the latter, the algorithm could not identify all tissue as relevant. The exclusion of some regions were however considered preferable over including areas without tissue, as these regions would introduce noise in the analysis. The devised algorithm was optimized for use at $40\times$ magnification, and a change in parameters would be required when computing masks at other magnifications.

For identification of connective tissue, two algorithms were proposed, multi-level thresholding and LBP using smaller blocks. The idea behind multi-level thresholding was to identify the connective tissue based on the visual differences between connective tissue and urothelial layer observed throughout the data set. Through preliminary testing of the method, a 12 level thresh-

olding was chosen as this resulted in a reasonable identification of connective tissue on the test images. An accuracy of 90 – 95% was obtained for the tree validation images. Larger variations were however found for both sensitivity and specificity between the three images. The method correctly identifies sections of the connective tissue, but struggles with incorrectly identification of parts in the urothelial layer. The second proposed algorithm used LBP in smaller blocks, removing blocks where the histogram has a difference above a set threshold from desired tissue. The method achieves a high sensitivity ($> 97\%$) for two images, while the third achieves 79%. Large variations were also found for specificity, where one image achieved 90% while the two others achieved 75 – 80%. As neither of the designed algorithms achieved a good identification of areas with connective tissue, manually defined masks were chosen for experiment 4 where impact of connective tissue in the analysis was tested.

Classification using only the urothelial layer, described in experiment 4, resulted in a reduction in both sensitivity and specificity. The experiment was conducted as a consequence of pathologists considered connective tissue not relevant in the analysis. The results from this experiment raises a question of whether connective tissue contains information of prognostic value for local texture analysis.

5.6 Other texture descriptors

During early stages of research, multiple directions were considered. Where the most prominent direction was studying patterns in the cell structure, a feature currently used by pathologists. The ambition for this analysis was to describe these patterns and detect possible abnormalities. Initial tests were done using Riesz transform to identify the pattern [35], before using a directional filter to identify regions where multi directions in the pattern existed. The amount of variations in a region were then used as a feature in the system.

A second approach was the use of a Gabor filter bank [36], consisting of 8 oriented filters in the region $[0, \pi]rad$ to describe amount of edges in each directions. As the prominent direction in each image depend on orientation of the scanned sample, the result where made rotation invariant by defining the most prominent direction as $\pi/2$. Variance in the histogram was then used as a feature to describe the texture. Neither of these approaches gave promising results, and were therefore abandoned.

Defining other features based on visual observations of the data set were also explored. As a large variation in the amount of cells were observed between images, a feature was derived to illustrate the ratio of area by cells and area covered by cytoplasm. The derived feature achieved results similar to random guessing, and was therefore abandoned.

6. Conclusion and future work

This thesis elaborate upon a system to predict recurrence and progression of urothelial carcinoma with the use of a local texture analysis on microscopic images of extracted tissue. Images are described using a histogram of labels computed from local binary pattern and local variance. Features used to predict prognoses of patients are computed as the similarity between the image histograms, and model histograms created using a training set with known prognostic follow-up information.

Results in this thesis suggest that both local binary pattern, and to some extent local variance, can be used to identify patients with recurrence. Local binary pattern achieve 83% correct identification of patients with recurrence, while the identification rate of patients without recurrence are approximately 50%. While local variance achieve approximately 60% identification of each class. Prognostic value for progression using the chosen local texture descriptors are however low, with no clear identification of patients with progression found.

Evaluation of the various modules in the designed system point to some areas which can be studied further and improved upon. This includes normalization of images, where a simple $[0, 1]$ grayscale normalization achieves higher results than histogram equalizing. In addition, performance of the system are reduced when analysing only the urothelial layer. Raising a question of whether the suburothelial layer contains information of prognostic value.

6.1 Future work

Future work is primarily focused on validating the results achieved in this thesis. Reproducibility of class identification, and inclusion of connective tissue when predicting recurrence are interesting topics.

The implemented system for predicting prognosis of a patient based on a microscopic images of tissue samples also has room for improvements. The creation of models for each prognosis, and use of a single histogram to describe each patient being the largest concerns. Two possible directions are considered interesting for improving the prognoses models. Use of additional images from each patient to achieve a better identification of the true texture of each prognoses, as sections used in this analysis only covers a small portion of the available tissue. And creating model histograms based on regions with a distinct cell differentiation. Multiple models can then be used to identify if the image has regions similar to each texture. A consequence of using a single histogram to describe each patient is a low focus on smaller regions with a distinct different texture. By studying smaller segments of tissue individually, the system can achieve a higher focus on these regions

In addition to the features used, other texture properties could contain useful information when predicting prognosis based on microscopic tissue samples. Maturity of cells are currently used

as a feature by pathologists. Computed aided extraction of this information and the variation of maturity within a tissue sample are considered an interesting direction for future research.

Bibliography

- [1] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, and F. Bray, “Globocan 2012 v1.0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11 [internet].” Lyon, France: International Agency for Research on Cancer; 2013. Available from: <http://globocan.iarc.fr>, accessed on 28/04/2016.
- [2] O. Mangrud, *Identification of patients with high and low risk of progression of urothelial carcinoma of the urinary bladder stage Ta and T1*. PhD thesis, University of Bergen, 2014.
- [3] J. Oosterhuis, R. Schapers, M. Janssen-Heijnen, R. Pauwels, D. Newling, and K. F. ten, “Histological grading of papillary urothelial carcinoma of the bladder: prognostic value of the 1998 who/isup classification system and comparison with conventional grading systems,” *Journal of Clinical Pathology*, vol. 55(12), pp. 900–905, Dec 2002.
- [4] Y. m. Lei, X. m. Zhao, and W. d. Guo, “Cirrhosis recognition of liver ultrasound images based on svm and uniform lbp feature,” in *2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 382–387, Dec 2015.
- [5] E. T. Pereira, S. P. Eleutério, and J. M. Carvalho, “Local binary patterns applied to breast cancer classification in mammographies,” *Revista de Informática Teórica e Aplicada*, vol. 21, no. 2, pp. 32–46, 2014.
- [6] D. G. Bostwick and L. Cheng, *Urologic surgical pathology*. Elsevier Health Sciences, 2008.
- [7] L. S. J. Borden, P. E. Clark, and M. C. Hall, “Bladder cancer,” *Current Opinion in Oncology*, vol. May 2004 - Volume 16 - Issue 3, pp. 257–262, 2004.
- [8] S. B. Edge, D. R. Byrd, C. C. Compton, A. G. Fritz, F. L. Greene, and A. T. III, *AJCC cancer staging manual*. New York: Springer, seventh edition ed., 2010.
- [9] F. K. Mostofi, C. J. Davis, and I. A. Sesterhenn, *Histological typing of urinary bladder tumours*. Springer Science & Business Media, 2012.
- [10] D. G. Bostwick and G. Mikuz, “Urothelial papillary (exophytic) neoplasms,” *Virchows Archiv*, vol. 441, no. 2, pp. 109–116, 2002.
- [11] J. Eble, r. G. Saute, J. Epstein, and I. Sesterhenn, *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs*. IARC Press, 2004.
- [12] G. T. MacLennan, Z. Krikkali, and L. Cheng, “Histologic grading of noninvasive papillary urothelial neoplasms,” *European Urology*, pp. 889–898, April 2007.

- [13] K. Yorukoglu, B. Tuna, E. Dikicioglu, E. Duzcan, A. Isisag, S. Sen, U. Mungan, and Z. Kirkali, "Reproducibility of the 1998 world health organization/international society of urologic pathology classification of papillary urothelial neoplasms of the urinary bladder," *Virchows Archiv*, vol. 443, no. 6, pp. 734–740, 2003.
- [14] O. Mangrud, E. Gudlaugsson, I. Dalen, I. Tasdemir, E. Janssen, and J. Baak, "Further evaluation of the reproducibility of the who grading systems for tat1 urinary bladder urothelial carcinomas: the influence of different microscopic characteristics." Manuscript, submitted to *Histopathology* March 2014.
- [15] O. Mangrud, R. Waalen, E. Gudlaugsson, I. Dalen, I. Tasdemir, E. A. M. Janssen, and J. Baak, "Reproducibility and prognostic value of who1973 and who2004 grading systems in tat1 urothelial carcinoma of the urinary bladder," *PLoS ONE*, vol. 9, pp. 1–8, 01 2014.
- [16] K. Yeong-Taeg, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE Transactions on Consumer Electronics*, vol. 43, pp. 1–8, Feb 1997.
- [17] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.
- [18] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51 – 59, 1996.
- [19] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 2037–2041, Dec 2006.
- [20] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, Jul 2002.
- [21] T. Mäenpää and M. Pietikäinen, "Multi-scale binary patterns for texture analysis. scia2003."
- [22] H. Wang, S. Viswanath, and A. Madabuhshi, "Discriminatively weighted multi-scale local binary patterns: Applications in prostate cancer diagnosis on t2w mri," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pp. 398–401, IEEE, 2013.
- [23] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 40, pp. 185–197, Jan 2010.
- [24] G. M. Weiss, "Mining with rarity: A unifying framework," *SIGKDD Explor. Newsl.*, vol. 6, pp. 7–19, June 2004.
- [25] V. B. Vaghela, A. Ganatra, and A. Thakkar, "Boost a weak learner to a strong learner using ensemble system approach," in *Advance Computing Conference, 2009. IACC 2009. IEEE International*, pp. 1432–1436, March 2009.
- [26] Y. Freund and R. Schapire, "A short introduction to boosting," 1999.
- [27] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," 1996.

- [28] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” *SIGKDD Explor. Newsl.*, vol. 6, pp. 20–29, June 2004.
- [29] C. Drummond and R. C. Holte, “C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling,” in *Workshop on learning from imbalanced datasets II*, vol. 11, Citeseer, 2003.
- [30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *J. Artif. Int. Res.*, vol. 16, pp. 321–357, June 2002.
- [31] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, “Smoteboost: improving prediction of the minority class in boosting,” in *In Proceedings of the Principles of Knowledge Discovery in Databases, PKDD-2003*, pp. 107–119, 2003.
- [32] K. Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *Philosophical Magazine Series 5*, vol. 50, no. 302, pp. 157–175, 1900.
- [33] U. of Oulu, “Local binary pattern, MATLAB implementation.” <http://www.cse.oulu.fi/CMV/Downloads/LBPmatlab>. lbp.m: ver. 0.3.3, cont.m: ver. 0.1.0.
- [34] S. Marsland, *Machine Learning: An Algorithmic Perspective*. CRC Press, 2011.
- [35] A. Depeursinge, A. Foncubierta-Rodriguez, D. Van de Ville, and H. Müller, *Lung Texture Classification Using Locally-Oriented Riesz Components*, pp. 231–238. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.
- [36] M. R. Turner, “Texture discrimination by gabor functions,” *Biological Cybernetics*, vol. 55, no. 2, pp. 71–82, 1986.

A. Matlab code

The following MATLAB files are embedded as `matlab.7z`.

runme.m:

A simplified initialization of `jstructi` and descriptors.

setupData.m:

Imports patient information from a xlsx document, and creates a `< struct >` variable for easy access to patient ID, stage and prognostic information from Matlab.

setupNormalization.m:

Inputs: `< struct >` The function imports all patient images into Matlab and stores it in the `< struct >` using the desired normalization method.

setupMasking.m:

Inputs: `< struct >` (including normalized images) The function analyse all normalized images, creates a logic mask to describe areas covered by tissue and stores the mask back to the `< struct >`.

generateLBPMMap:

Inputs: `< struct >`, number of neighbours, radius and parameter
LBP histograms for all images are returned (images are masked using the masks defined in `< struct >`). Map of LBP labels for all pixels can also be returned.

generateVARMap:

Inputs: `< struct >`, number of neighbours, radius, number of bins
Variance histograms for all images are returned (images are masked using the masks defined in `< struct >`).
Map of variance labels for all pixels can also be returned.

crossVal :

Inputs: `< struct >`, LBPMMap, VARMap, scales, number of weak learners
Designed cross validation.

cv_Chi_squared.m

Inputs: `< struct >`, histograms to use, naming of new fields, image ID X
Computes chi squared for all images when predicting prognosis of image ID X Used from `crossVal`

LBP.m

Computes the LBP descriptor, source: University of Oulo.

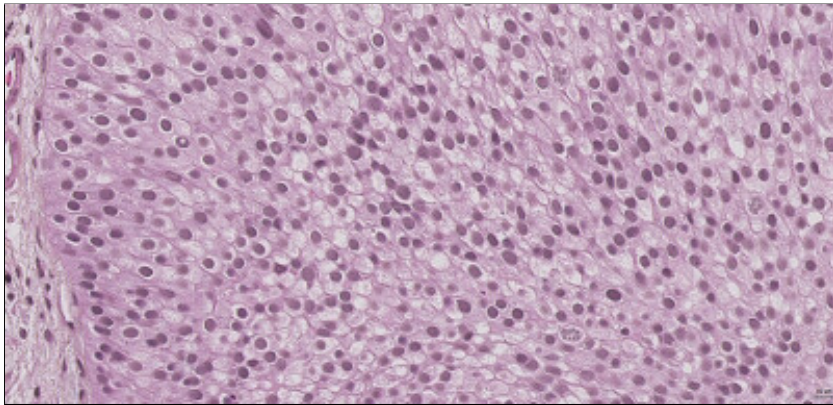
cont.m

Computes the VAR descriptor, source: University of Oulo.

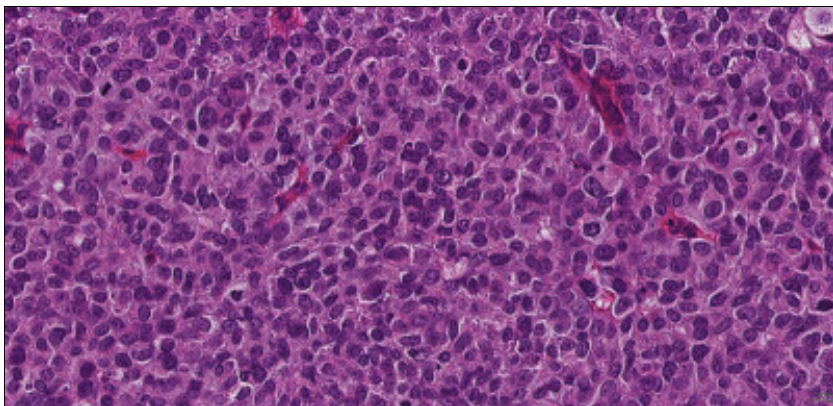
getmapping.m

Configures mapping required for LBP.m and cont.m, source: University of Oulo.

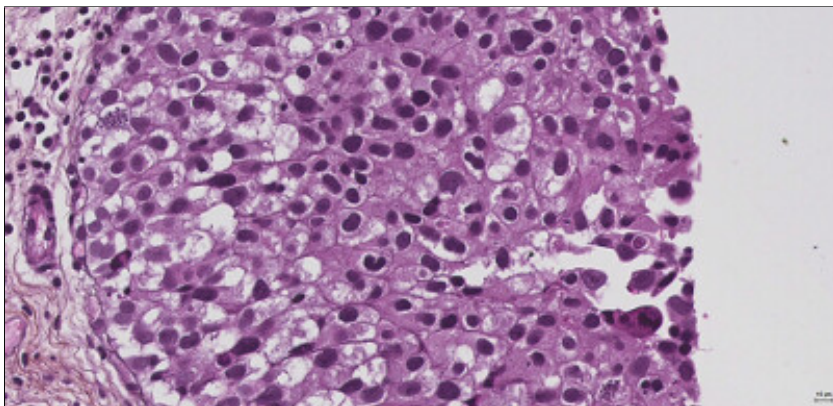
B. Data set



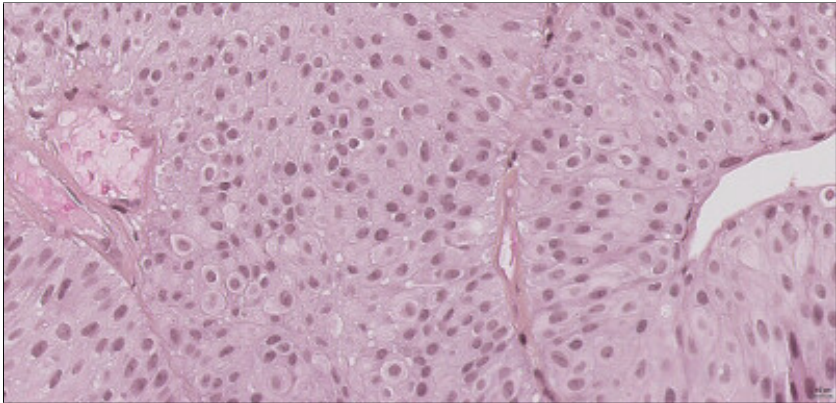
ID: 5



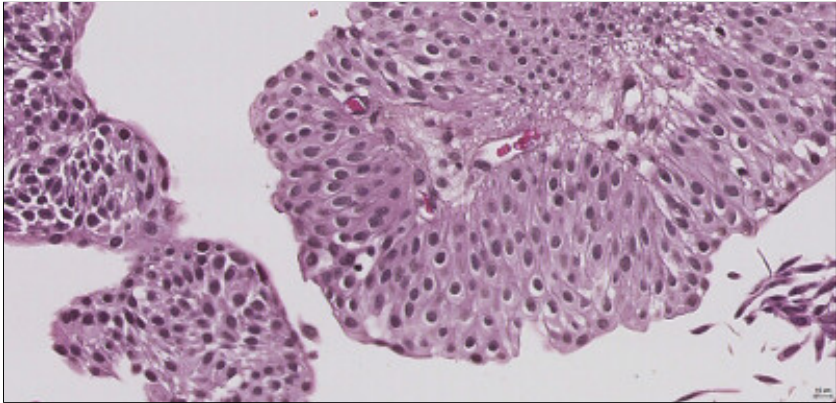
ID: 6



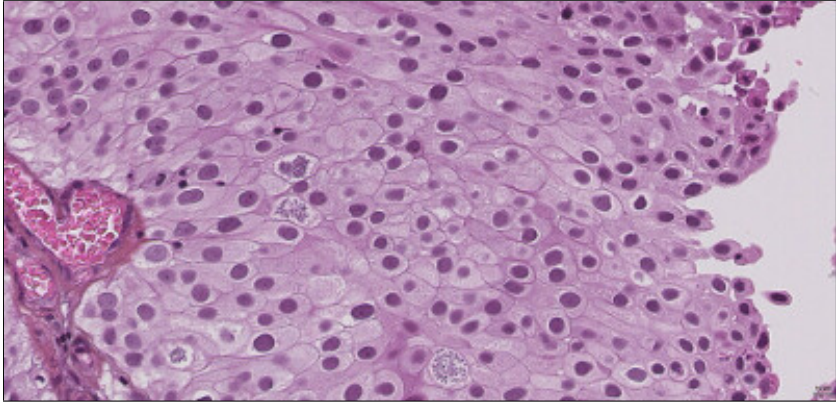
ID: 11



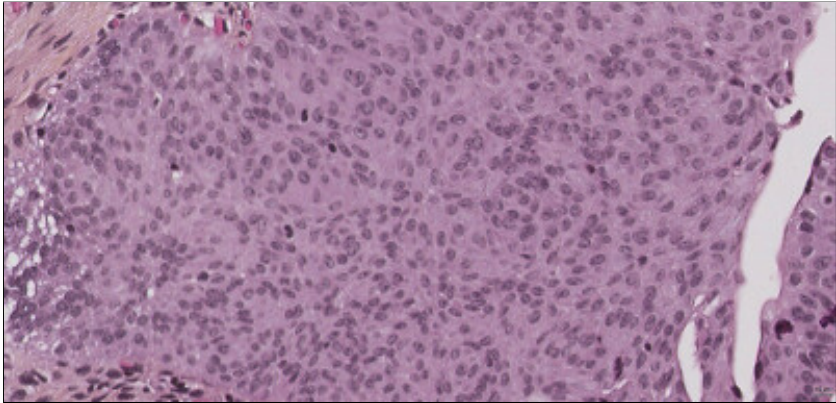
ID: 13



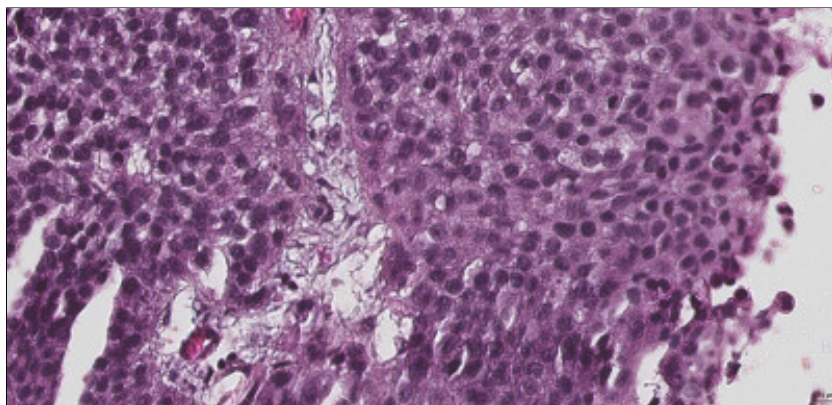
ID: 15



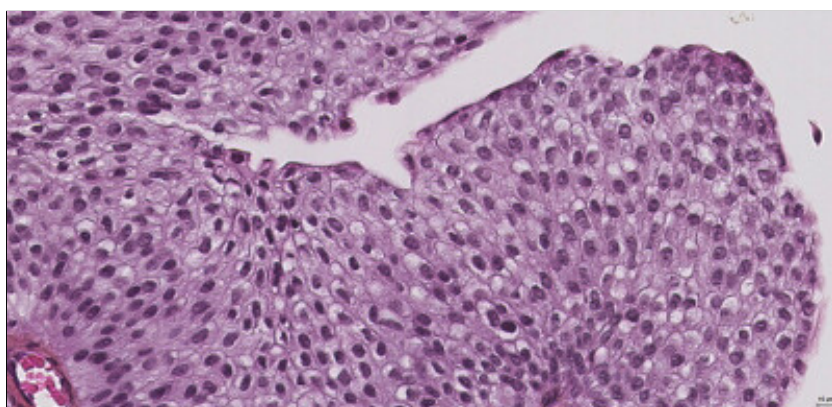
ID: 18



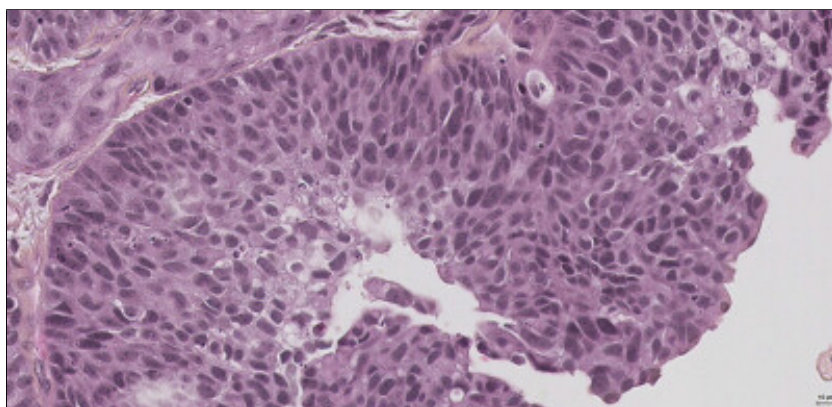
ID: 21



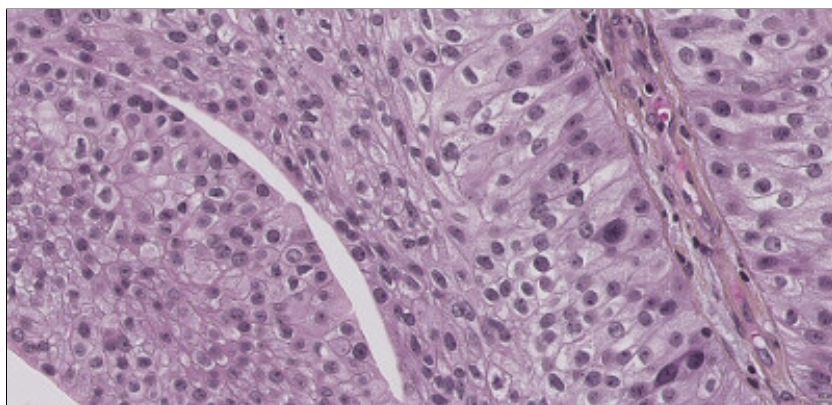
ID: 26



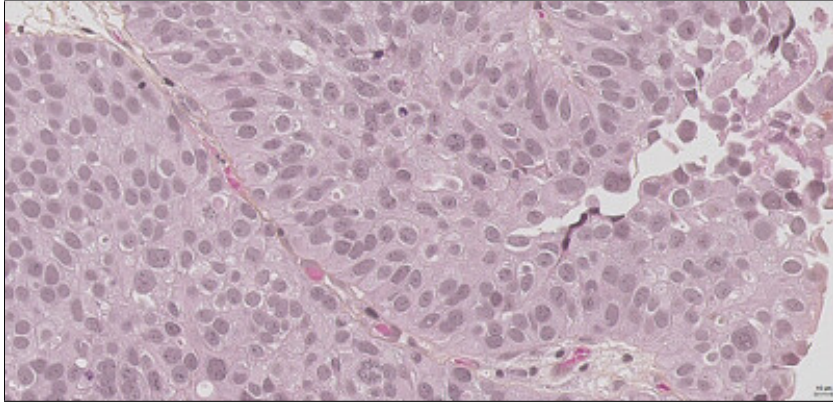
ID: 28



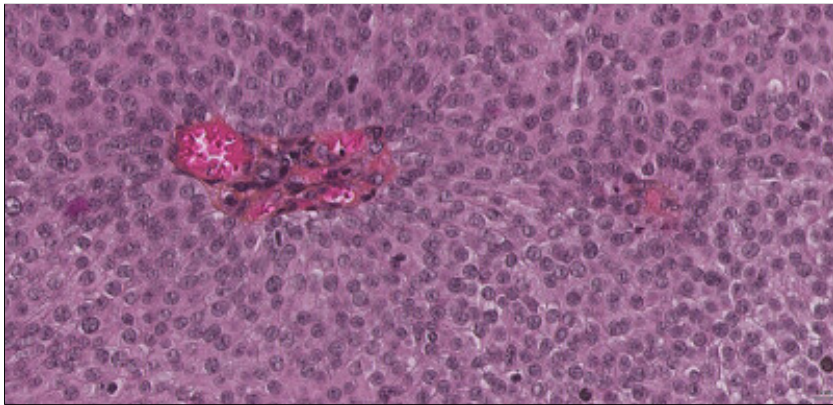
ID: 31



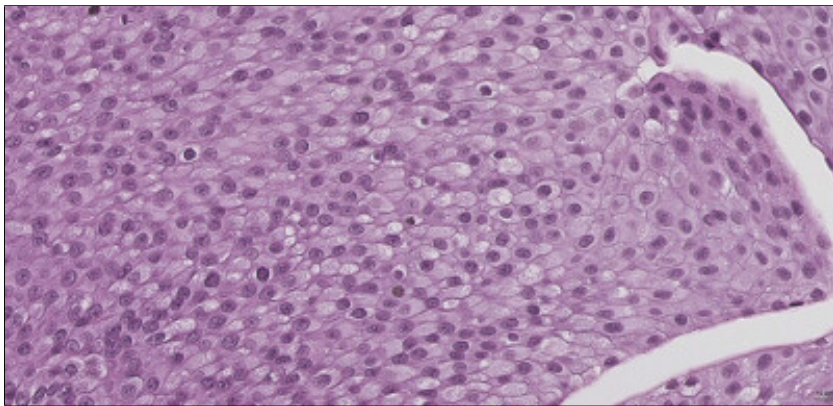
ID: 43



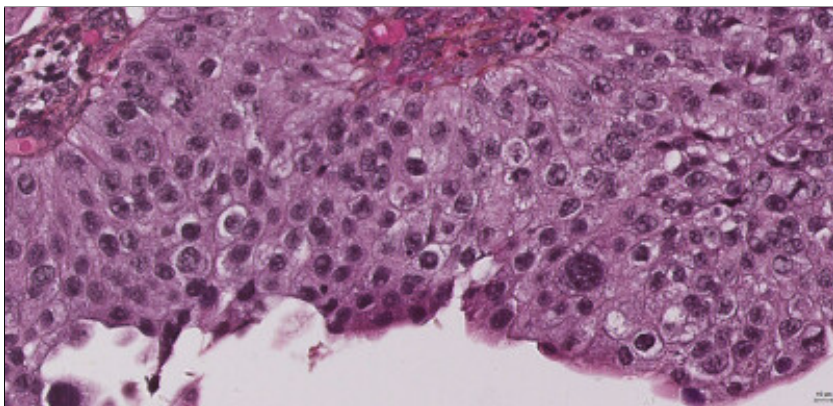
ID: 46



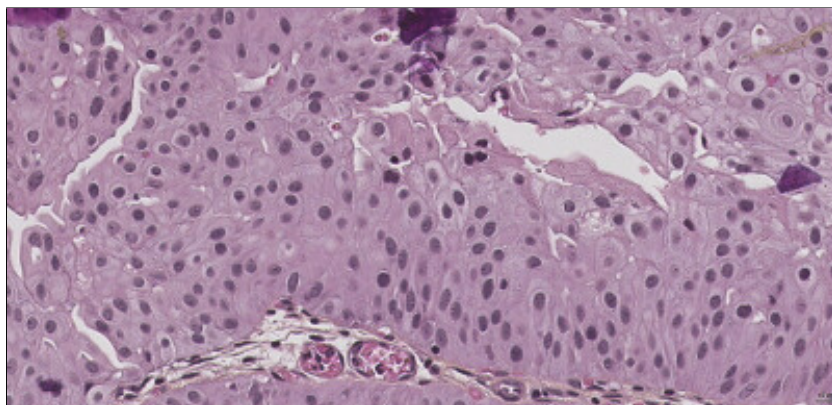
ID: 47



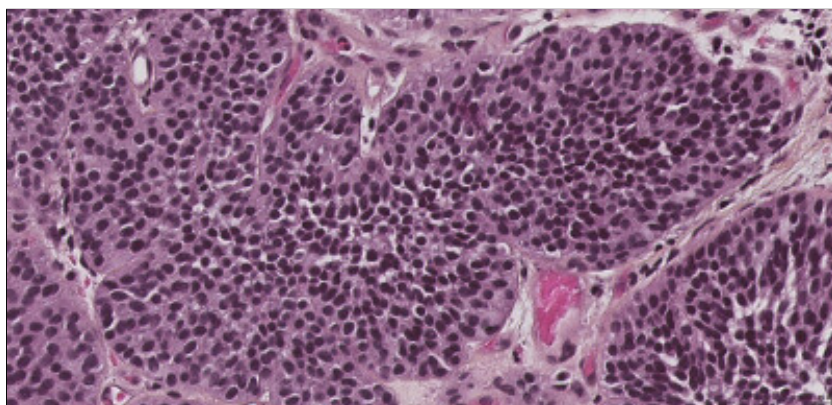
ID: 52



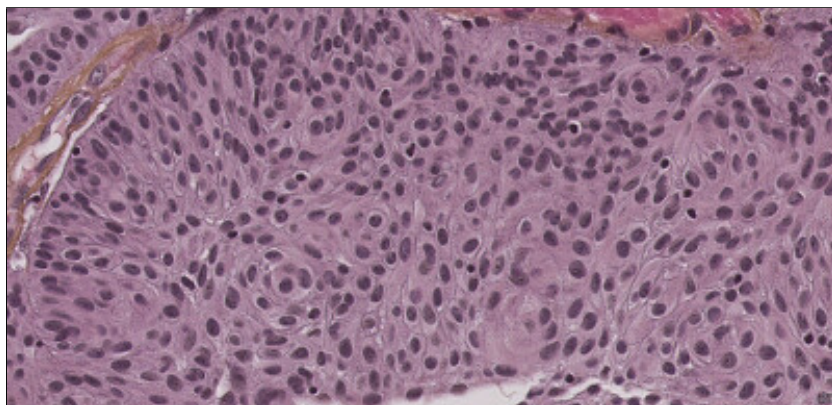
ID: 56



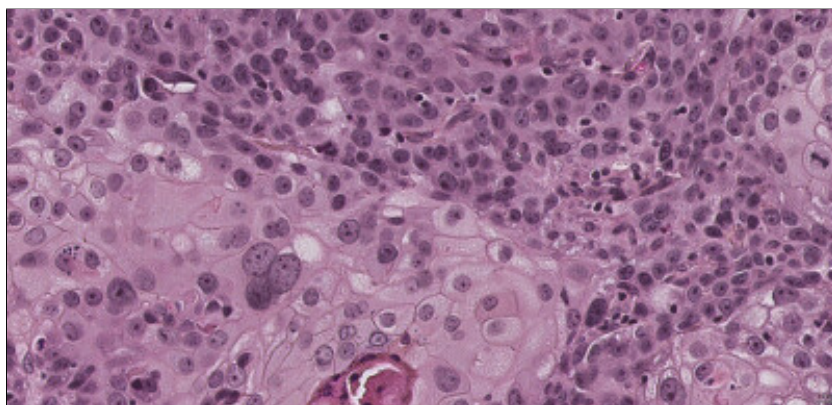
ID: 59



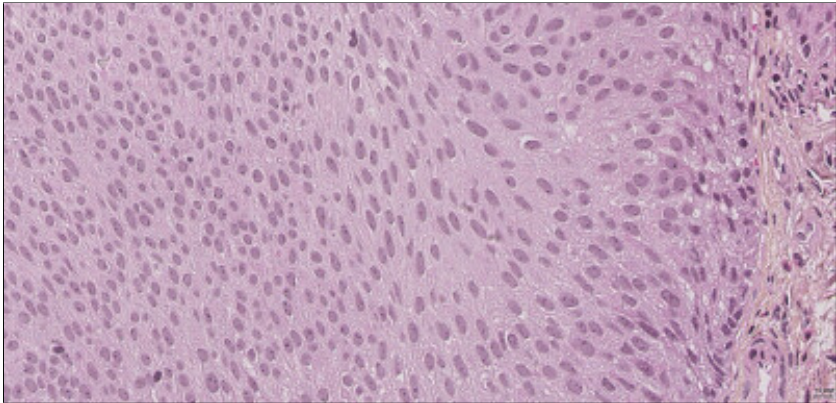
ID: 67



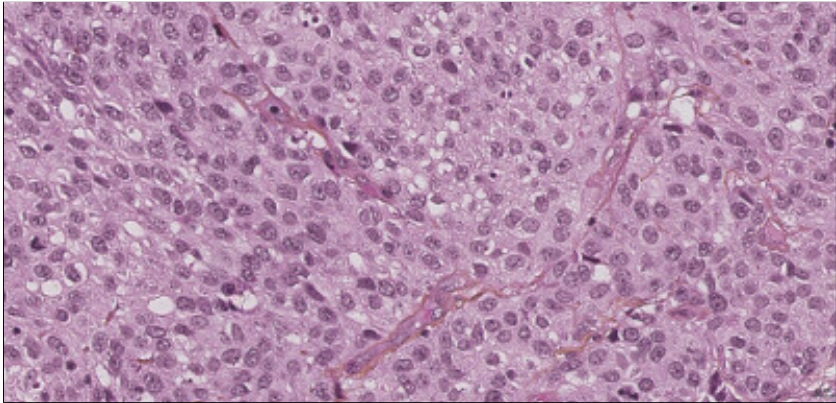
ID: 68



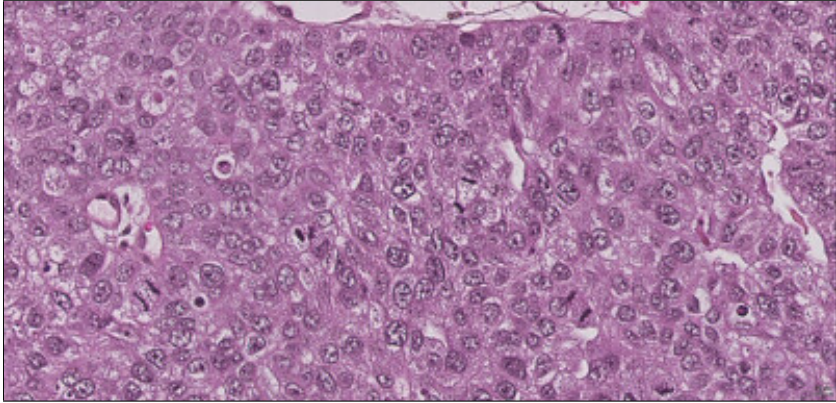
ID: 74



ID: 78



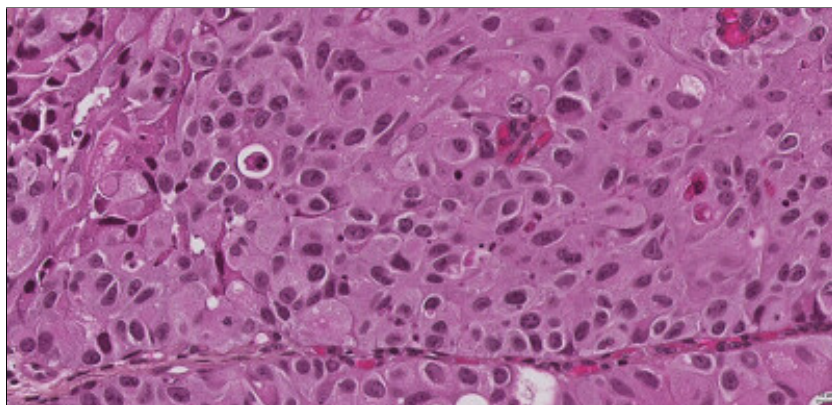
ID: 85



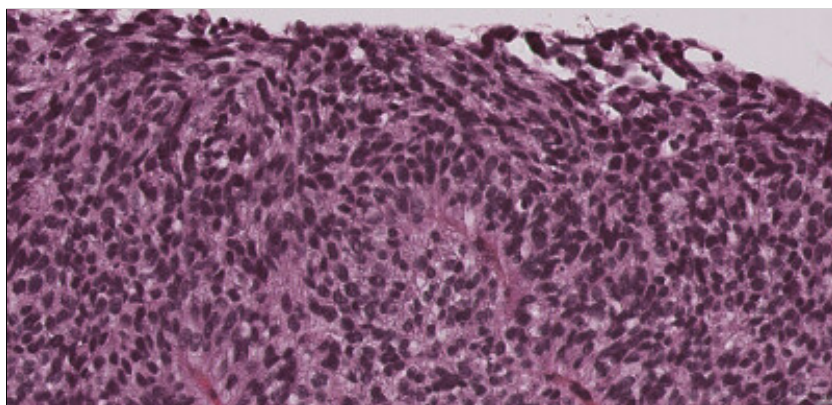
ID: 91



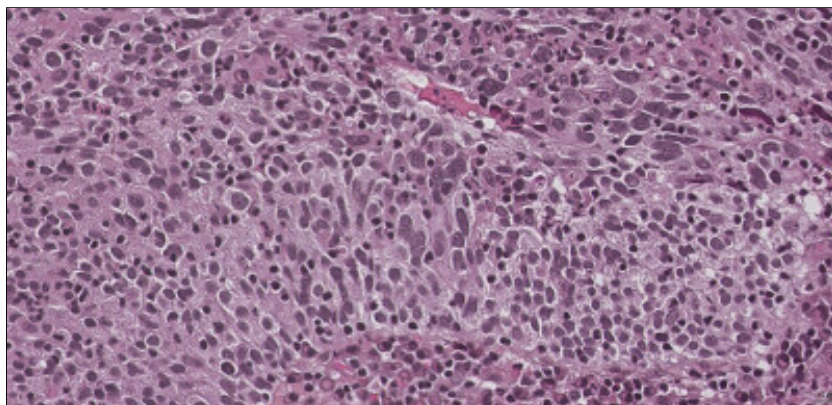
ID: 98



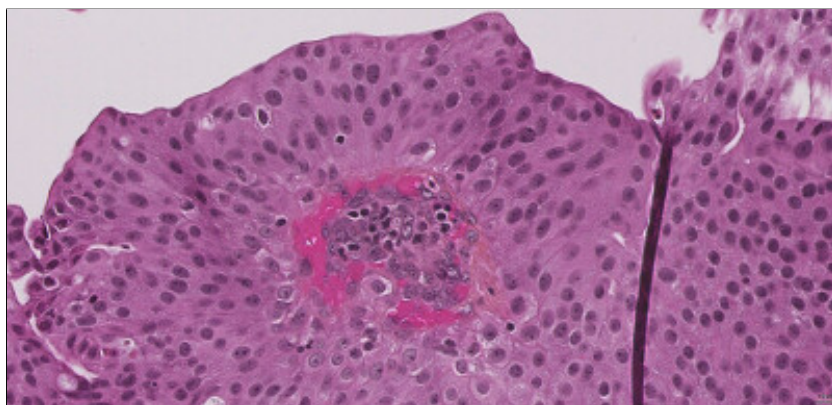
ID: 99



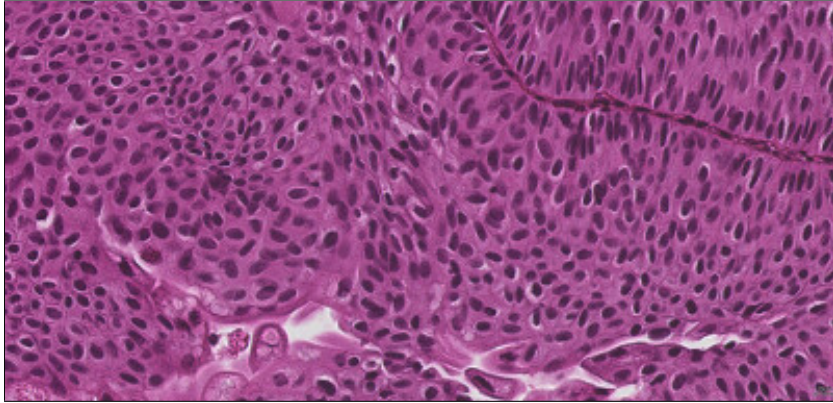
ID: 111



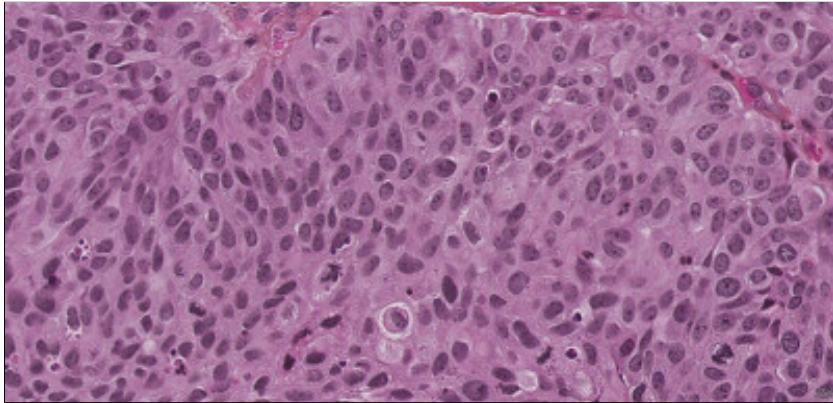
ID: 113



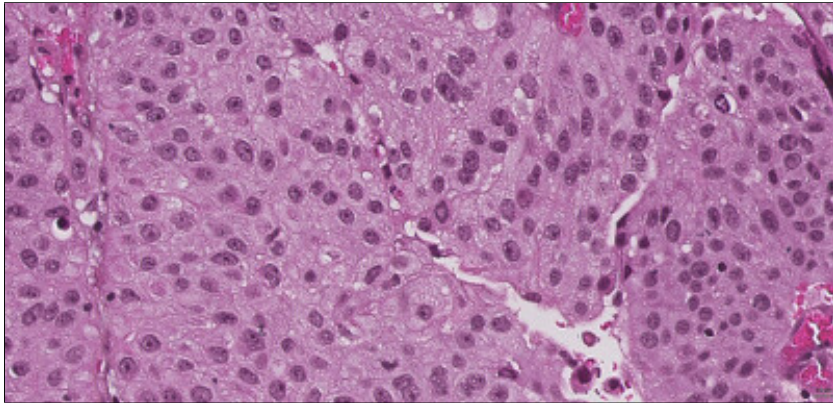
ID: 116



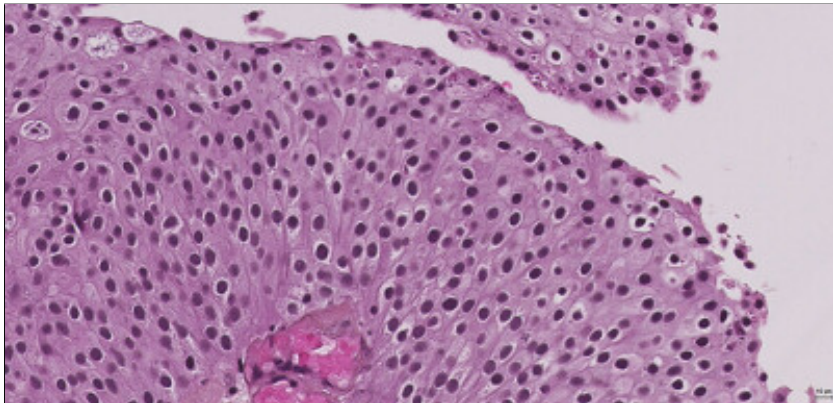
ID: 134



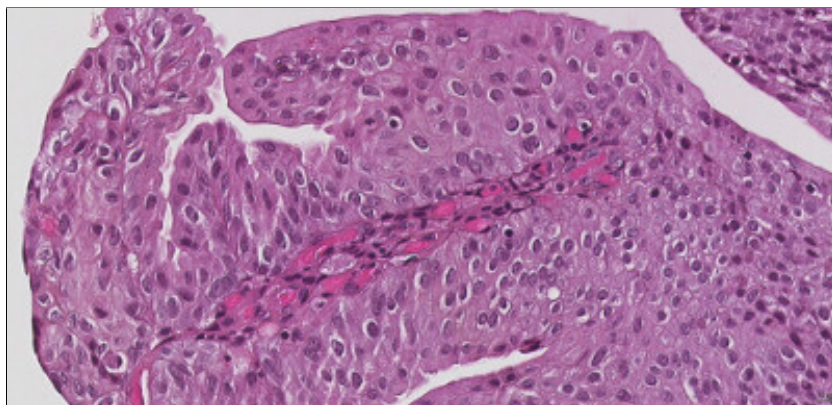
ID: 135



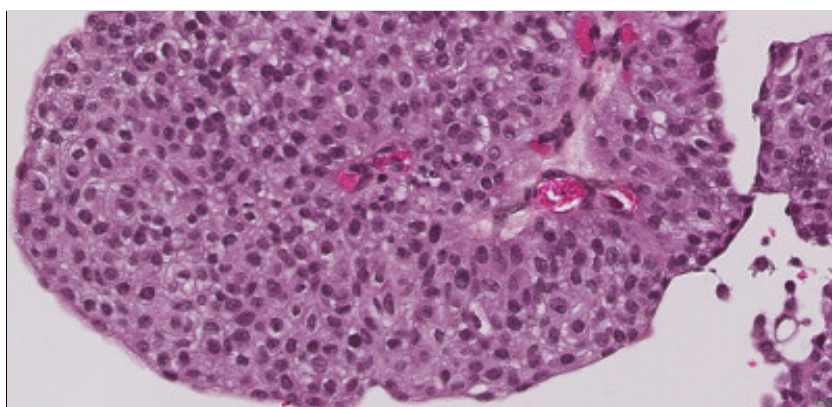
ID: 145



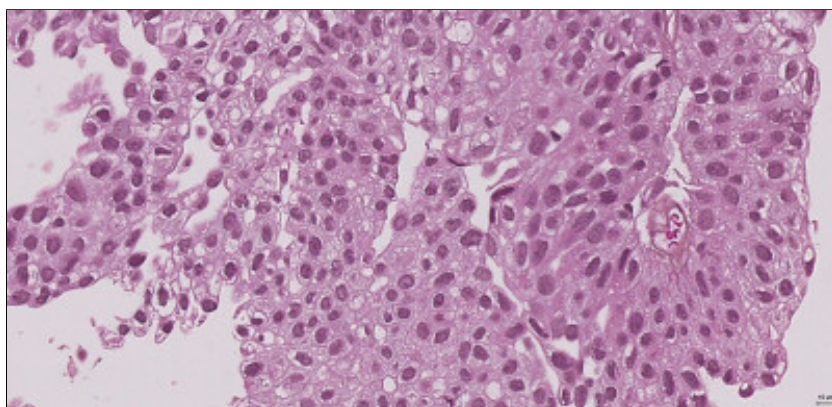
ID: 147



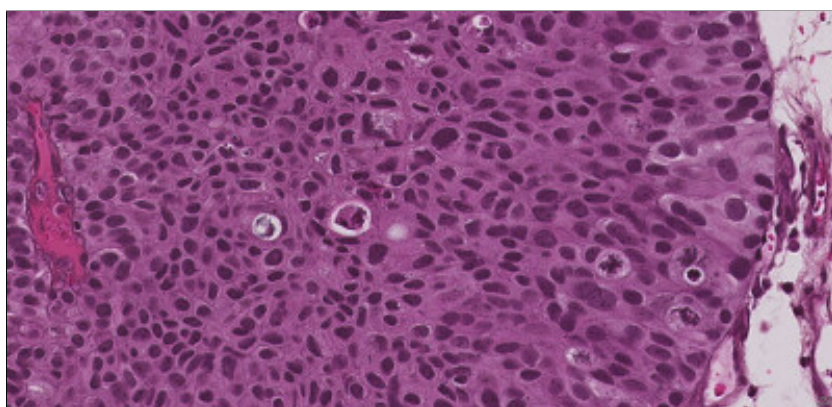
ID: 157



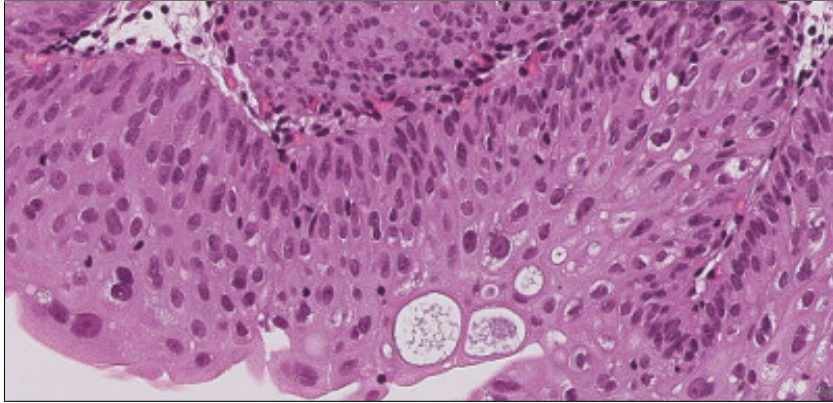
ID: 166



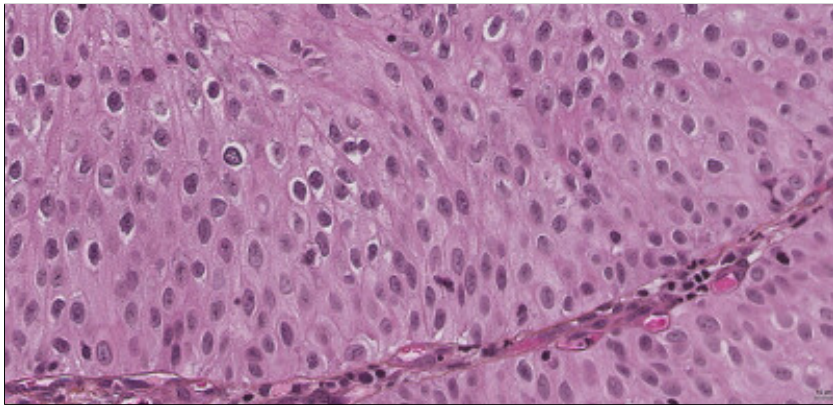
ID: 167



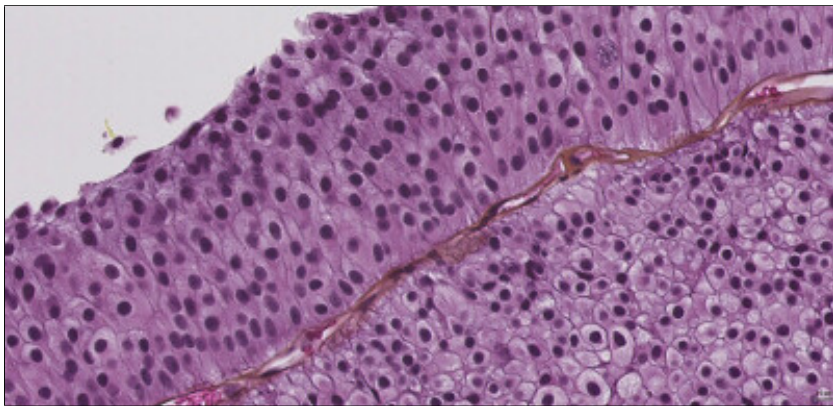
ID: 169



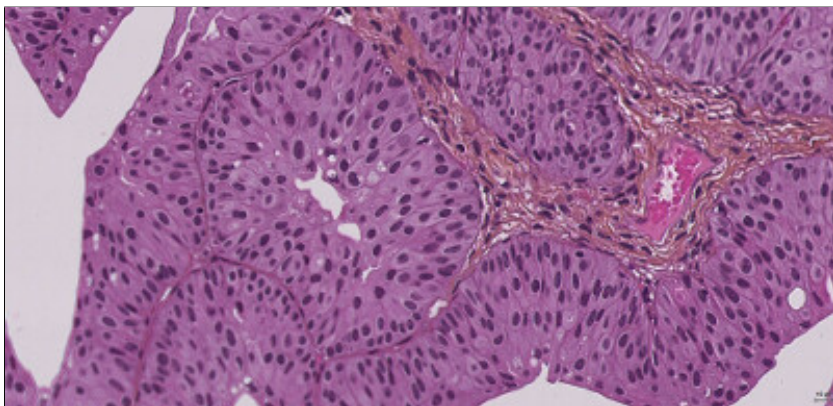
ID: 171



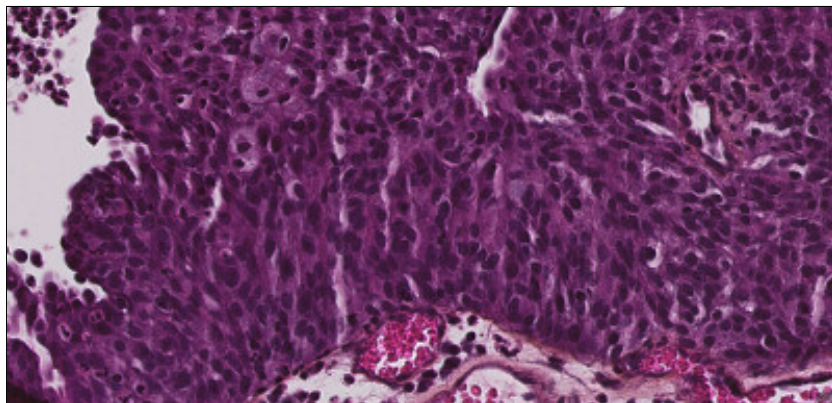
ID: 172



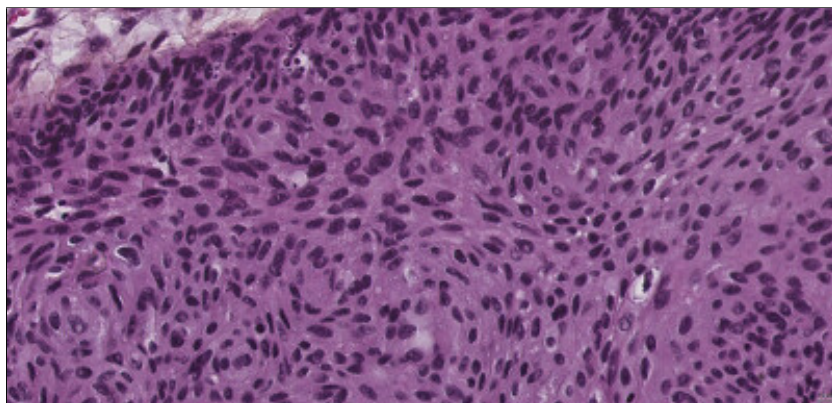
ID: 181



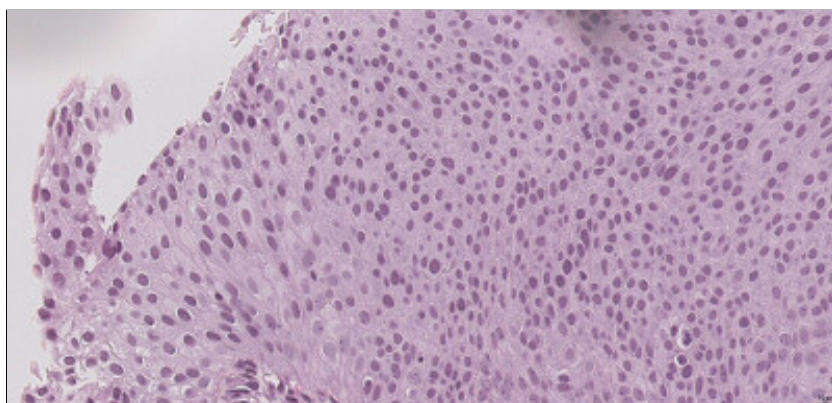
ID: 182



ID: 186



ID: 187



ID: 190

C. Patient follow-up

Case ID	Recurrence	Progression	Muscle-invasive progression	Stage
85	yes	no	No	T1
6	yes	yes	Yes	TA
15	yes	no	No	TA
134	yes	no	No	TA
181	yes	no	No	TA
182	no	no	No	TA
186	no	no	No	T1
47	yes	yes	Yes	T1
52	yes	no	No	TA
5	yes	no	No	TA
28	yes	no	No	TA
98	no	no	No	TA
167	no	no	No	TA
190	yes	no	No	TA
21	yes	no	No	TA
13	no	no	No	TA
187	yes	yes	Yes	TA
171	no	no	No	TA
67	yes	yes	Yes	TA
111	yes	no	No	TA
116	yes	yes	Yes	T1
43	yes	yes	Yes	T1
78	no	no	No	TA
166	no	no	No	TA
46	yes	yes	Yes	T1
145	yes	yes	Yes	T1
172	yes	no	No	TA
68	yes	no	No	TA
26	yes	no	No	T1
18	yes	yes	Yes	TA
157	no	no	No	TA
11	yes	yes	Yes	T1
59	no	no	No	TA
91	yes	yes	No	TA
135	yes	yes	Yes	T1
169	yes	no	No	T1
31	yes	yes	Yes	T1
56	no	no	No	TA
74	no	no	No	T1
99	yes	yes	Yes	T1
113	no	no	No	T1
147	yes	yes	No	TA