DOCTOR OF PHILOSOPHY

Analysis of breast tissue microarray spots

Amaral, Telmo

Award date:
2010

Awarding institution:
University of Dundee

Analysis of breast tissue microarray spots

Telmo Amaral

2010

University of Dundee

Analysis of breast tissue microarray spots

A thesis submitted in application for the degree of Doctor of Philosophy

Telmo Amaral

School of Computing

University of Dundee

May 2010

Contents

List of Figures

List of Tables

Acknowledgements

I would like to acknowledge several people whose contributions greatly helped me to accomplish the work reported in this thesis. My supervisor, Stephen McKenna, provided me with essential guidance throughout countless regular meetings, in which I could benefit from his deep theoretical knowledge and ability to bring attention to seemingly unrelated concepts. His thorough and rigorous comments led to major improvements in all the manuscripts (including that of this thesis), presentations, and posters that I had to prepare. My co-supervisor, Alastair Thompson, of the Division of Surgery and Molecular Oncology (DSMO), warranted access to the necessary data, facilitated important contacts, and helped to clarify many clinical concepts, in the course of several meetings at the Ninewells Hospital. Katherine Robertson, of the Division of Pathology and Neuroscience (DPN), offered indispensable insight into the procedures undertaken by pathologists and patiently assessed hundreds of tissue microarray spots, to complement the pre-existing data. Michele Sciarabba, of the Department of Human Morphology and Biomedical Sciences of the University of Milan, contributed with many useful discussions and his implementation of a nuclear segmentation method, as well as the patient manual drawing of regions of interest on many tissue spots. Phil Quinlan, of the DSMO, made arrangements for a trial of the Aperio Genie tool and granted me access to it. Lee Jordan, of the DPN, fuelled interesting discussions on potential future directions of the work carried out. My colleagues at the Computer Vision and Imaging Processing group have always ensured a friendly and stimulating work environment. Their choices of papers for our fortnightly reading group meetings often provided me with very useful ideas for this work. My family and friends offered essential support and encouragement throughout my studies, and demonstrated great understanding for my limited availability to spend time with them. I should also acknowledge some organisations, in particular the Breast Cancer Research Trust, which offered the research grant that funded my studentship. The School of Computing, the Leng Trust Fund, and the British Machine Vision Association contributed with travel funds that allowed me to participate in several events of great relevance to this work. To all, my big Thanks.

Associated publications

This work has been reported in research papers submitted to a number of events, namely: the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI) in 2008 (poster presentation); the Medical Image Understanding and Analysis conferences (MIUA) in 2008 and 2009; the International Conference on Computer Vision Theory and Applications (VISAPP) in 2009; and the Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (OPTIMHisE) MICCAI workshop in 2009 (poster presentation). These publications are listed below.

Additional dissemination material included a poster presented at the BMVA Computer Vision Summer School in 2007, an abstract and poster presented at the San Antonio Breast Cancer Symposium (SABCS) in 2008, and an extended abstract presented at the BMVA technical meeting on Microscopy Image Analysis for Biomedical Applications in 2010.

T. Amaral, S. McKenna, K. Robertson, and A. Thompson. Classification of breast tissue microarray spots using colour and local invariants. In *IEEE International Symposium on Biomedical Imaging*, pages 999–1002, 2008.

T. Amaral, S. McKenna, K. Robertson, and A. Thompson. Classification of breast tissue microarray spots using texton histograms. In *Annual Conference on Medical Image Understanding and Analysis*, pages 144–148, 2008.

T. Amaral, S. McKenna, K. Robertson, and A. Thompson. Scoring of breast tissue microarray spots through ordinal regression. In *International Conference on Computer Vision Theory and Applications*, volume 2, pages 243–248, 2009.

T. Amaral, S. McKenna, K. Robertson, and A. Thompson. Automated classification of breast tissue microarray spots. *Cancer Research*, 69(2 Supplement): 4010, 2009.

T. Amaral, S. McKenna, K. Robertson, and A. Thompson. Analysis of breast tissue microarrays using Latent Dirichlet Allocation. In *Optical Tissue Image* *analysis in Microscopy, Histopathology and Endoscopy, MICCAI workshop*, pages 112–123, 2009.

T. Amaral, M. Sciarabba, S. McKenna, K. Robertson, and A. Thompson. Scoring of breast tissue microarrays using ordinal regression: local patches vs. nuclei segmentation. *In Annual Conference on Medical Image Understanding and Analysis*, pages 77–81, 2009.

Declaration by the author

I hereby declare that I am the author of this thesis; that all references cited have been consulted by me; that the work of which this thesis is a record has been done by me, and that it has not been previously accepted for a higher degree.

Signed

Telmo Amaral

May 2010

Declaration by the supervisor

I hereby certify that Mr Telmo Amaral has satisfied all the terms and conditions of the regulations made under Ordinances 12 and 39 and has completed the required nine terms of research to qualify in submitting this thesis in application for the degree of Doctor of Philosophy.

Signed

Prof. Stephen McKenna

May 2010

Summary

Tissue microarrays (TMAs) are a high-throughput technique that facilitates the survey of very large numbers of tumours, important both in clinical and research applications. However, the assessment of stained TMA sections is laborious and still needs to be carried manually, constituting a bottleneck in the pathologist's work-flow. This process is also prone to perceptual errors and observer variability. Thus, there is strong motivation for the development of automated quantitative analysis of TMA image data. The analysis of breast TMA sections subjected to nuclear immunostaining begins with the classification of each spot as to the main type of tissue that it contains, namely tumour, normal, stroma, or fat. Tumour and normal spots are then assigned a so-called quickscore composed of a pair or integer values, the first reflecting the proportion of epithelial nuclei that are stained, and the second reflecting the strength of staining of those nuclei. In this work, an approach was developed to analyse breast TMA spots subjected to progesterone receptor immunohistochemistry. Spots were classified into their four main types through a method that combined a bag of features approach and classifiers based on either multi-layer perceptrons or latent Dirichlet allocation models. A classification accuracy of 74.6 % was achieved. Tumour and normal spots were scored via an approach that involved the computation of global features formalising the quickscore values used by pathologists, and the use of Gaussian processes for ordinal regression to predict actual quickscores based on global features. Mean absolute errors of 0.888 and 0.779 were achieved in the prediction of the first and second quickscore values, respectively. By setting thresholds on prediction confidence, it was possible to classify and score fractions of spots with substantially higher accuracies and lower mean absolute errors. A method for the segmentation of TMA spots into regions of different types was also investigated, to explore the generative nature of latent Dirichlet allocation models.

List of symbols

An effort was made to avoid ambiguities in the notation used in mathematical descriptions. Lowercase Roman letters denote scalars, as in x, except when denoting functions. Lowercase bold Roman letters denote vectors, as in $\mathbf{x}$, and uppercase Roman letters denote constants, as in D. Uppercase bold letters denote matrices, as in $\mathbf{W}$ and $\mathbf{\Sigma}$. Whenever possible, indexing variables use the same letter as the range, as in x_d, $d \in \{1, \dots, D\}$. The calligraphic $\mathcal{D}$ denotes a data set of observations.

In order to maintain the notation commonly used in some literature, a few exceptions that should be clear from the text were made. Thus, a lowercase Greek letter (such as α, β, θ, γ, and φ) denotes in most cases a set of model parameters, but μ and σ keep their common usage as single scalars for mean and standard deviation, and δ is used once to denote a random variable.

The following sections list the symbols that are shared across the discussions of different techniques, as well as those that are used in the description of specific techniques.

Shared

D	Size of a feature vector.
$\mathbf{x} = (x_1, \dots, x_D)^\mathrm{T}$	A feature vector.
x_d, $d \in \{1, \dots, D\}$	d-th element in a feature vector.
C	Number of classes or ordinal targets.
$\{T_1, \dots, T_C\}$	Unordered set of classes or ordered set of ordinal targets.
$t \in \{T_1, \dots, T_C\}$	A class or ordinal target.
N	Number of observations in a data set.
$\mathcal{D} = \{(\mathbf{x}, t_1), \dots, (\mathbf{x}_N, t_N)\}$	A data set of observations.
$\mathbf{x}_n$, $n \in \{1, \dots, N\}$	n-th feature vector in a data set.
t_n, $n \in \{1, \dots, N\}$	n-th class or ordinal target in a data set.

Classification with neural networks

a	An activation variable.
$\mathbf{W}$	Matrix of network weights.
w	A network weight.
y	A network output.
M	Number of hidden units.
z	Hidden unit output.
e	Entropy error.
t_{bin}	Element of binary vector encoding an observed class.
e_r	Regularised error.
A	Regularisation constant.

Latent Dirichlet Allocation

V	Number of unique codewords.
$\mathbf{z}$	A vector of topics.
z	A topic.
K	Number of unique topics.
θ	Topic mixture for a given feature vector.
α	Dirichlet parameters.
β	Codeword distributions per topic.
γ	First variational parameters.
φ	Second variational parameters.
q	Variational distribution.
l	Log likelihood of (α, β).

Gaussian process ordinal regression

$f(\mathbf{x})$	Random variable associated with input $\mathbf{x}$.
b	Boundary of ordinal interval.
σ^2_{noise}	Variance of Gaussian noise.
μ	Mean of distribution of random variable.
σ^2	Variance of distribution of random variable.
θ	Model hyper-parameters.
$\mathcal{N}$	Noise Gaussian distribution.
δ	Noise random variable.
κ_o	Kernel parameter.

κ_a	Kernel parameter.
Z_f	A normalisation factor.
$\mathbf{f}$	Vector of random variables.
$\boldsymbol{\Sigma}$	Covariance matrix.

Chapter 1

Introduction

This thesis reports work carried out by the author at the School of Computing of the University of Dundee in collaboration with the School of Medicine, concerning the automated classification and scoring of breast tissue microarray spots subjected to progesterone receptor immunohistochemistry. This work started in October 2006.

Section 1.1 of this introduction provides an overview of the problem in question and the motivation to address it. Section 1.2 summarises some findings from the review of related literature and explains the contributions of this work. Finally, Section 1.3 describes the structure of the remainder of this thesis.

1.1 Problem and motivation

Tissue microarrays (TMAs) are a high-throughput technique that facilitates the survey of very large numbers of tumours. At present, TMA construction, staining, and high-resolution image capture are largely automated processes. This renders the TMA technique ideal for high-volume analysis of specimens, which can play an important role both in clinical and in research applications. Clinical applications include diagnosis, identification of causes behind particular diagnoses, and targeting of treatment, whereas research applications may involve large trials that investigate associations between biological markers and disease behaviour. However, the assessment of stained TMA sections (each of which may contain hundreds of tissue specimens, called *spots*) is laborious and still needs to be carried out manually, constituting a bottleneck in the pathologist's work-flow.

The analysis of breast TMA sections subjected to some form of nuclear immunostaining (such as progesterone receptor immunohistochemistry) begins with the classification of each spot as to the main type of tissue that it contains, namely tumour, normal, stroma, or fat. Tumour and normal spots are then assigned a *quickscore*. The quickscore is a technique for semiquantitation of immunostained tissue sections that dispenses with the need to count individual cells, introduced by Detre et al. in 1995 [30]. Each score is composed of a pair or integer values, one reflecting the proportion of epithelial nuclei that are stained (therefore immunopositive), and the other reflecting the strength of staining of those nuclei. Besides being time-consuming, this process is also prone to perceptual errors and observer variability. Thus, there is strong motivation for the development of automated methods for quantitative analysis of breast TMA image data.

1.2 Contributions

The following bullet points list the main contributions of the present work and provide some context to each of them, in the light of the literature reviewed in Chapter 3.

- An approach was developed to classify breast TMA spots into their four main types, namely tumour, normal, stroma and fat, with the purpose of identifying tumour and normal spots that needed to be subsequently scored, while discarding spots containing only stroma and fatty tissue. The reviewed literature reflects the existence of research work dealing with the classification of tissue sections into distinct types. The reviewed methods, however, focus mainly on the distinction between tumour and normal tissue, and do not deal with the detection of sections containing only connective or fatty tissue.

- The developed method was applied to spots subjected to a form of nuclear immunostaining. Most of the reviewed methods on classification of tissue into types deal with sections stained only with haematoxylin and eosin, as opposed to sections subjected to some form of immunohistochemistry. It should be noted that the presence of immunostaining does not necessarily help to distinguish normal tissue from tumour, given that both can exhibit staining (as explained in Section 2.2, the staining simply expresses an antigen that can be present in any cells). Thus, there is a possibility that the classification of immunostained tissue spots into different types constitutes a more demanding task than the classification of spots stained solely with haematoxylin and eosin, from the point of view of automated analysis (this, however, is speculative and would need to be tested experimentally).

- The implemented method was based on the technique introduced by Varma and Zisserman in 2005 [98], so that a histogram of *texton* frequencies was computed to characterise each spot and a classifier was trained to classify spots based on their texton histograms. This classification approach based on bags of textural features has been applied to the analysis of histological images of breast tissue, but only very recently [17].

- The classification performance of a multi-layer perceptron (MLP) was compared with that of a classifier based on latent Dirichlet allocation (LDAL) models. Unlike the MLP, LDAL is a generative approach that tries to explain the modelled data, thus lending itself to other interesting applications besides classification. By associating distinct types of tissue with latent variables of the LDAL model, a method for the segmentation of TMA spots into regions of different types was also explored. The models underlying the reviewed classification methods are typically discriminative, as opposed to generative. In particular, to the best of the author's knowledge, the LDAL model has not been used in the classification or segmentation of histopathology images.

- In this work, an approach was developed to predict the quickscores of tumour and normal breast TMA spots subjected to a form of nuclear immunostaining. No literature was found on the prediction of TMA quickscores, as defined by Detre et al. [30]. Several other reviewed methods are concerned with the "ranking" of tissue sections. However, in all cases this corresponds to the prediction of Bloom-Richardson grades (or variants thereof) for sections stained solely with haematoxylin and eosin, as opposed to the prediction of quickscores for immunostained sections.

- Both quickscore integer values were predicted, to reflect the proportion of epithelial nuclei that were stained as well as the strength of their staining. Methods have been reported that estimate the proportion of epithelial nuclei that are immunopositive within tissue sections subjected to some form of nuclear immunostaining. In the case of breast tissue, this type of result could in principle be used to predict the first quickscore integer value, but most of the reviewed approaches do not deal with the estimation of staining strength.

- Given the difficulties inherent to the accurate segmentation of individual cells or nuclei in images of tissue (such as the complexity of tissue structure, the variability of cell appearance, cell overlapping, and the presence of debris), the developed method was based on the hypothesis that the prediction of scores would not need to rely on a segmentation technique. Thus, the basis for the computation of global features was the labelling of pixels as to the probability of their belonging to each of three classes, namely background, immunopositive nucleus, and immunonegative nucleus. From this labelling, features formalising the quickscore values used by pathologists were computed. Existing methods that estimate the proportion of immunopositive epithelial nuclei normally rely on the accurate segmentation of individual nuclei.

- The models trained to predict the quickscores of spots based on their global features were not classifiers, but rather Gaussian processes for ordinal regression. This type of model was expected to perform better than a classifier, given that it incorporates knowledge about the relative order between categories. Existing methods concerned with "ranking" tissue sections make use of classifiers and not ordinal regression algorithms.

- The posterior probabilities output by the MLP classifier (for classification) and by the ordinal regression algorithm (for scoring) were used to compute a simple measure of prediction confidence. This allowed to set confidence thresholds that helped to distinguish the "easier" spots that could be processed automatically with high confidence from the more "difficult" spots that should be referred for manual assessment. Several of the methods reported in the literature make use of classifiers that provide a probabilistic output. In general, however, this output is used merely to decide the predicted type or grade of a tissue section, by choosing the class associated with the highest probability.

The diagram in Figure 1.1 summarises the proposed methods.

1.3 Structure of thesis

The remainder of this thesis is structured as follows. Chapter 2 provides an overview of clinical concepts closely related with this work, including those of breast cancer, immunohistochemistry, scoring of tissue sections, and TMAs. The used data are also introduced in this chapter. Chapter 3 presents a review of literature relevant to this work and a brief overview of existing commercial systems. Chapter 4 begins with an overview of classification with neural networks

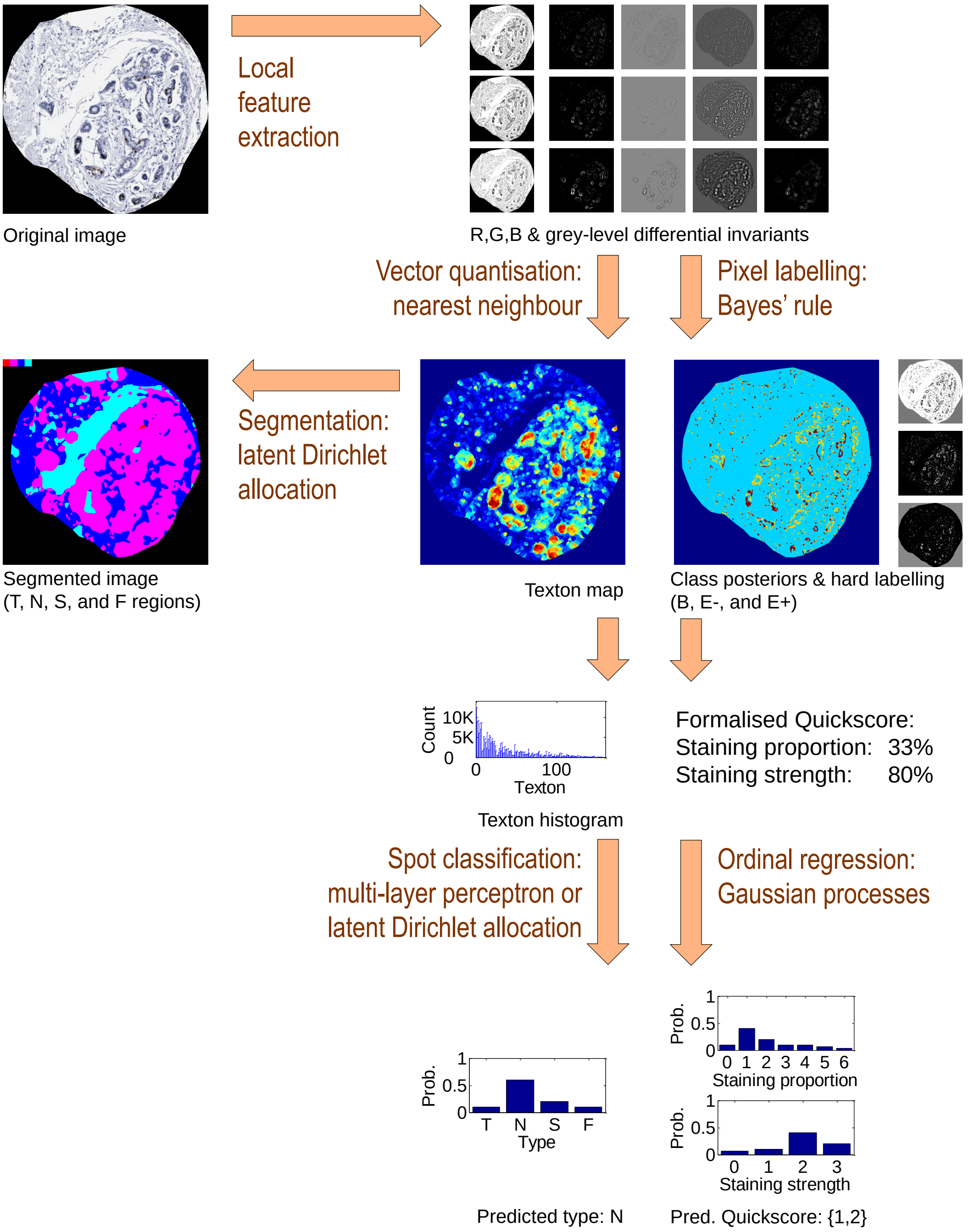

Local feature extraction
Original image
R,G,B & grey-level differential invariants
Vector quantisation: nearest neighbour
Pixel labelling: Bayes' rule
Segmentation: latent Dirichlet allocation
Segmented image (T, N, S, and F regions)
Texton map
Class posteriors & hard labelling (B, E-, and E+)
Count
10K
5K
0
0
100
Texton
Texton histogram
Formalised Quickscore:
Staining proportion: 33%
Staining strength: 80%
Spot classification: multi-layer perceptron or latent Dirichlet allocation
Ordinal regression: Gaussian processes
Prob.
1
0.5
0
0 1 2 3 4 5 6
Staining proportion
Prob.
1
0.5
0
0 1 2 3
Staining strength
Prob.
1
0.5
0
T N S F
Type
Predicted type: N
Pred. Quickscore: {1,2}

Figure 1.1: Overview of the proposed methods. (Pixels classes: B for background, E- for immunonegative, and E+ for immunopositive. Spot and region types: T for tumour, N for normal, S for stroma, and F for fat.)

and of latent Dirichlet allocation, and then presents the experiments and results related with the classification of TMA spots into types. Chapter 5 reports some experiments and results related with the segmentation of spots into regions, and describes a trial of the Genie commercial tool from Aperio, Inc. Chapter 6 begins with an overview of Gaussian processes applied to ordinal regression, and then presents the experiments and results related with the scoring of spots. Finally, Chapter 7 summarises the conclusions drawn from the accomplished experiments and discusses possible future directions of work.

Chapter 2

Breast cancer and immunohistochemistry

2.1 Breast cancer

Tumours in the breast can be formed as a result of the ungoverned development of cells. Tumours can be benign (posing no threat to health) or malignant, and the expression “breast cancer” usually refers to a malignant breast tumour. The cells in which breast cancer commonly originates are either those forming the glands (or “lobules”) that produce milk, or those making up the ducts (or “tubules”) that convey milk to the nipple from the producing glands. A less usual form of breast cancer can develop in the breast’s fibrous connective tissues and adipose tissues [14].

Breast cancer is the most frequent form of cancer in the UK, even though it is uncommon in men. In 2006, 45,822 new cases were diagnosed, over 99% of which occurred in women. There is a clear association between the risk of breast cancer and age. In the UK, breast cancer is rarely diagnosed in teenage girls or women in their early 20s, but it is the most frequent form of cancer in women under the age of 35. Each year, 1,400 women are diagnosed in the 35-39 age group. Most of the cases occur in the 50-69 age group, and cases in women over the age of 50 account for 81% of the total [20].

Like in other developed countries, the incidence of breast cancer in the UK has been growing for the last decades. Between 1977 and 2006, the age-standardised rates of incidence per 100,000 women increased from 75 to 122. Given that the survival rate for 5 years is 80%, this high incidence means that a large number of women with breast cancer are alive. This number is estimated at 550,000. In Europe, the northern and western countries report the highest incidence rates of breast cancer, whereas Romania and Latvia have the lowest rates. The highest rates worldwide occur in the developed world, and the lowest rates are reported in the African and Asian continents. Each year, over one million cases of breast cancer are diagnosed in women, which represents over a fifth of all female cancers and a tenth of all cancers regardless of gender [20].

The last two decades have witnessed advances in the targeting of treatment for patients with breast cancers that respond to therapy. Nevertheless, treatments remain unsuccessful for a percentage of patients who experience disease recurrence. In addition, adjuvant therapy is still unnecessarily administered to a fraction of patients who turn out to be disease free. These issues highlight the current importance of conducting large breast cancer trials involving the construction of tissue banks, to facilitate extensive research of associations between disease behaviour and biological markers.

2.2 Immunohistochemistry

Immunohistochemistry (IHC) emerged in 1941, when Coons et al. [25] reported on a technique for the detection of antigens labelled with fluorescent dyes in cells present in histological sections [80]. Antigen identification in tissue specimens plays a crucial role not only in tumour diagnosis, but also in prognosis, evaluation of response to specific treatments, and selection of adequate therapies for patients. The effectiveness of IHC for identification of cellular antigens is today superior to that of other methods, both in cytology and in histology applications. IHC therefore has established itself as a very important, if not fundamental, instrument in diagnostic, research, and surgical pathology. As regards light microscopy, IHC is currently the leading technique for the examination of different antigens in histological sections fixed with formaldehyde and embedded in paraffin [44].

As the name itself suggests, immunohistochemistry encompasses the disciplines of immunology, histology, and chemistry. The simple principle underlying IHC is the manifestation of particular antigens present in the tissue by taking advantage of the antigens' ability to bind with particular antibodies. The binding between antigen and antibody constitutes an immune reaction, which can be visualised if the antibody is attached to a *label* (or reporter) molecule. Most commonly, enzymes such as peroxidase, alkaline phosphatase, and glucose oxidase are used

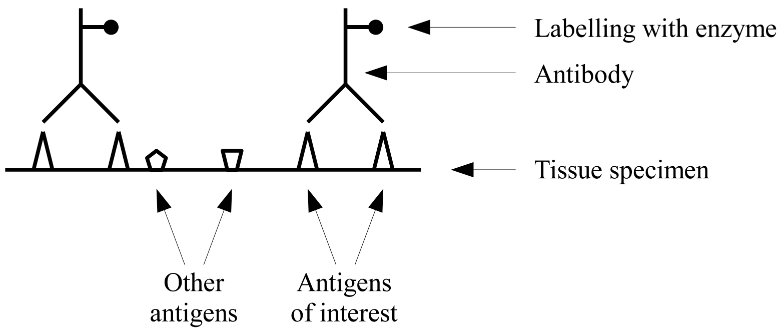


Figure 2.1: Diagrammatic illustration of direct immunohistochemistry. (Based on Coligan et al. [23].)

as labels. A histochemical reaction between the enzyme, a substrate, and a chromogen results in a precipitate that can be observed through light microscopy, revealing the location of the antigen-antibody binding. Fluorescent compounds can be used for labelling, too, resulting in fluorochromes that are visible under ultraviolet light [80]. Figure 2.1 illustrates a basic immunohistochemical method, so-called direct method.

Immunohistological techniques comprise a typical sequence of tasks. Once the specimens have been collected, their morphology and antigens must be preserved. This can be achieved via chemical fixation (involving tissue processing, embedding in paraffin, and slicing) or via frozen section processing. In some cases, the specimens need to be subjected to antigen retrieval (also called unmasking), a technique that increases the exposure of the antigen and its ability to bind with the antibody. The preserved tissue should then be incubated in an antibody or series of antibodies, and stained (to trigger the labelling reaction). Following the observation of antigen-antibody reaction, the results must be interpreted [44, 70].

2.3 Scoring of tissue sections

Oestrogens are hormones that play a very important role in health as well as in disease. Vital functions, including the differentiation, development, and behaviour of a large variety of tissues, are brought about and controlled by oestrogens. In the breast, they promote not only the normal, but also the abnormal (neoplastic) proliferation of epithelial cells, affecting to a great extent the metastisation of cancer cells. The evolution of breast cancer is particularly influenced by the oestrogen receptor and progesterone receptor antigens. IHC is one of the main techniques through which the oestrogen and progesterone receptor statuses can be assessed in tissue specimens, proving to be a valuable instrument in the prognosis of breast cancer, as well as in the evaluation of therapy [44].

Within a tissue section subjected to IHC designed to express an oestrogen, nuclei that manifest the antigen-antibody reaction are said to be immunopositive and appear stained, whereas immunonegative nuclei maintain an unstained appearance and are visible only as the result of a counter-stain such as haematoxylin and eosin.

The evaluation of molecular biomarkers (such as antigens) for prognosis of breast cancer often benefits from semiquantitation. However, many of the available methods are burdensome. The proportion of nuclei that are immunopositive and the strength of their staining are taken into account in one such method, the H-score [30, 64]. Similar semiquantitative methods include those introduced by Remmele and Stegner [83] and by Reiner et al. [82], any of which require 100 cells to be assessed within each of a minimum of three fields on a microscope slide. The scoring should be carried out blindly by two or more pathologists, so that a a consensus score may be achieved between them [44].

Detre et al. [30] developed a so-called *quickscore* that makes it unnecessary to evaluate hundreds of individual nuclei. In that study, the expression of oestrogen receptor was semiquantitavely assessed via an assay incorporating the proposed quickscore, and the results were compared with those obtained via two commonly used assays that involve the more time-consuming H-score. The quickscore was found to be a trustworthy technique. The prognostic significance of the expression of p27 (a protein thought to play a role in breast tumour suppression), as semiquantitatively determined via the quickscore, has been studied by Barnes et al. [5]. A similar study has been published by Bejar et al. [7], relating to the expression of oestrogen receptor.

A quickscore is composed of two integer values. The proportion of immunopositive (stained) nuclei within the tissue section is given a score between 1 and 6 (1 for 0 to 4%, 2 for 5 to 19%, 3 for 20 to 39%, 4 for 40 to 59%, 5 for 60 to 79%, and 6 for 80 to 100%), whereas the average strength of staining is assigned a score between 0 and 3 (0 for negative, 1 for weak, 2 for intermediate, and 3 for strong staining) [30]. Similar quick-scoring methods used in breast cancer clinical IHC include the Histoscore [49] and Allred [2] systems.

The scoring methods described above should not be confused with cancer histological grading, which does not aim at scoring the reaction of breast tissue sections to IHC, but rather at grading invasive breast cancer based on sections stained only with haematoxylin and eosin. The Bloom-Richardson staging system introduced in 1957 [13] can be used to assess the histological grade of tissue sections, by combining three individual scores associated with: the degree of structural differentiation as shown by the presence of tubular arrangements of epithelial cells; the variation in size, shape, and staining of epithelial nuclei; and the frequency of hyperchromatic and mitotic figures (dividing cells). More recent variations of this system exist, such as the Nottingham grade, which, according to Rakha et al. [79], can reliably predict the clinical outcome in patients diagnosed with invasive breast carcinoma. Although these grading systems have been historically criticised as being observer dependent, more recently they have achieved wide acceptance in routine clinical practice.

2.4 Tissue microarrays

A high-throughput technique that allows high volume analysis of tissue samples involving multiple immunohistochemical markers is the construction of tissue microarrays (TMAs), proposed by Kononen et al. in 1998 [54]. TMAs are now extensively utilised in the study of cancers.

To create a TMA, a pathologist identifies six or more sites on a *donor* block of formalin-fixed, wax-embedded cancer tissue. Sites of interest are regions of tumour and normal epithelial tissue, which may be identified with the help of a whole section of tissue cut from the donor block, mounted on a microscope slide, and stained with haematoxylin and eosin.

Cylindrical biopsies, named cores, are then extracted from the identified sites in the donor block and inserted into a *recipient* wax block, which constitutes the actual TMA. The extraction and transference of cores can be done with a microarrayer device that incorporates a biopsy punch with stylet. This process is repeated for multiple donor tissue blocks, in such a way that cores of known provenance are placed alongside each other. The result is a grid arrangement of cores in the (single) recipient TMA block. Typical cores range from 2 to 4 mm in length and have a diameter of 0.6 mm.

Sections of the TMA block, 4 to 8 μm in thickness, are then cut and mounted on microscope slides. Thus, each cylindrical core of tissue from the TMA block originates a disk of tissue on each slide. These disks of tissue are named spots. Figure 2.2 illustrates the process of constructing a TMA, while Figures 2.3(a) and (b) show examples of a breast TMA slide and an individual spot, respectively.

IHC can be carried to detect protein expression in tissue spots, by staining each

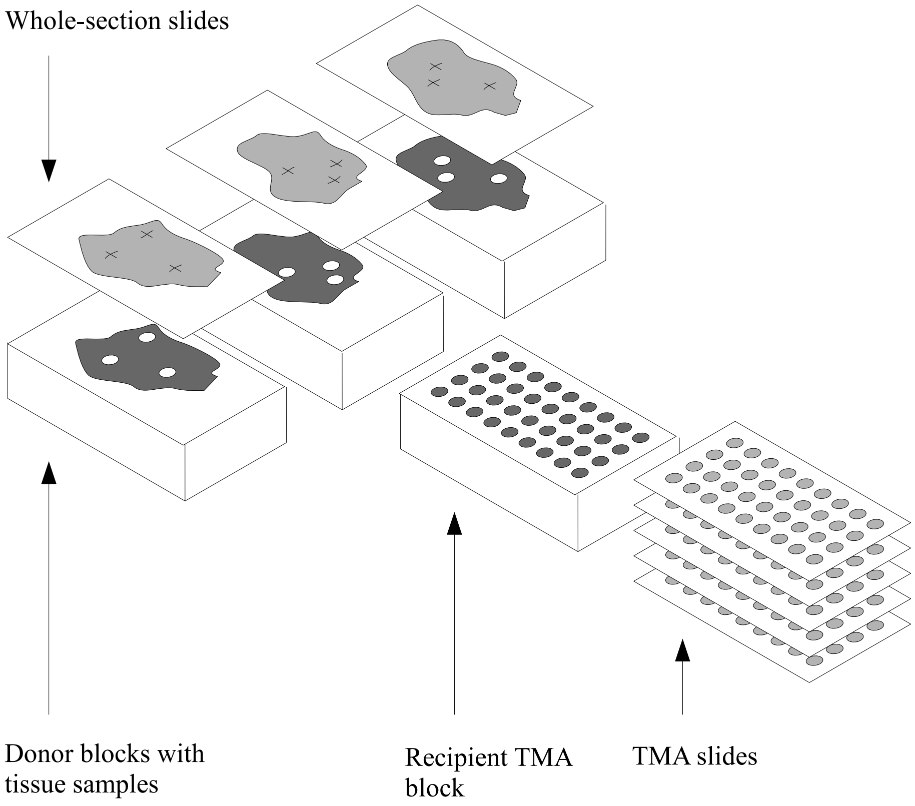


Figure 2.2: Construction of a tissue microarray. Sites of interest are identified via whole-section slides. Tissue cores are extracted from multiple donor blocks and transferred into a single recipient block. Multiple sections can be cut from the recipient block and mounted on microscope slides.

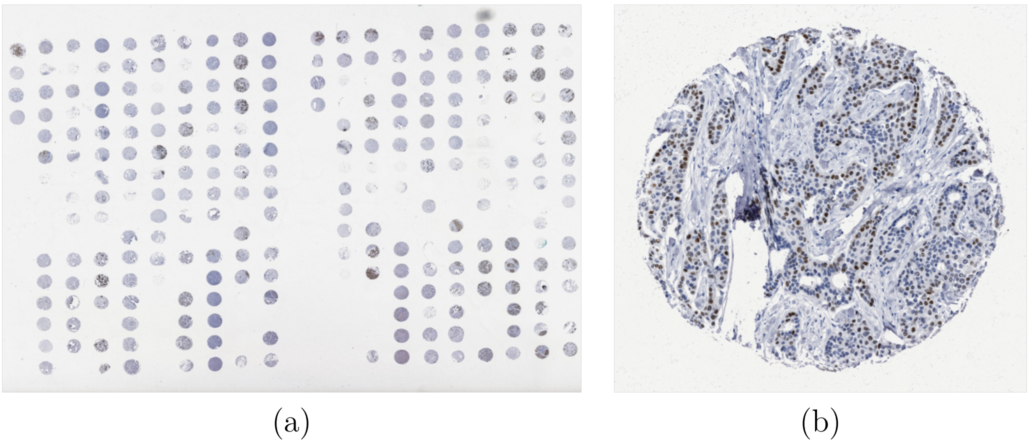


Figure 2.3: (a) A breast tissue microarray slide and (b) an individual spot.

TMA section with a small aliquot of antibody. For example, antibodies directed against progesterone receptor can be used to detect nuclear expression of that antigen in epithelial cells of breast tumours. TMA sections should also be counter-stained (for example with haematoxylin), to allow immunonegative structures to be visible. However, this counter-stain should be sufficiently light not to mask immunopositive structures.

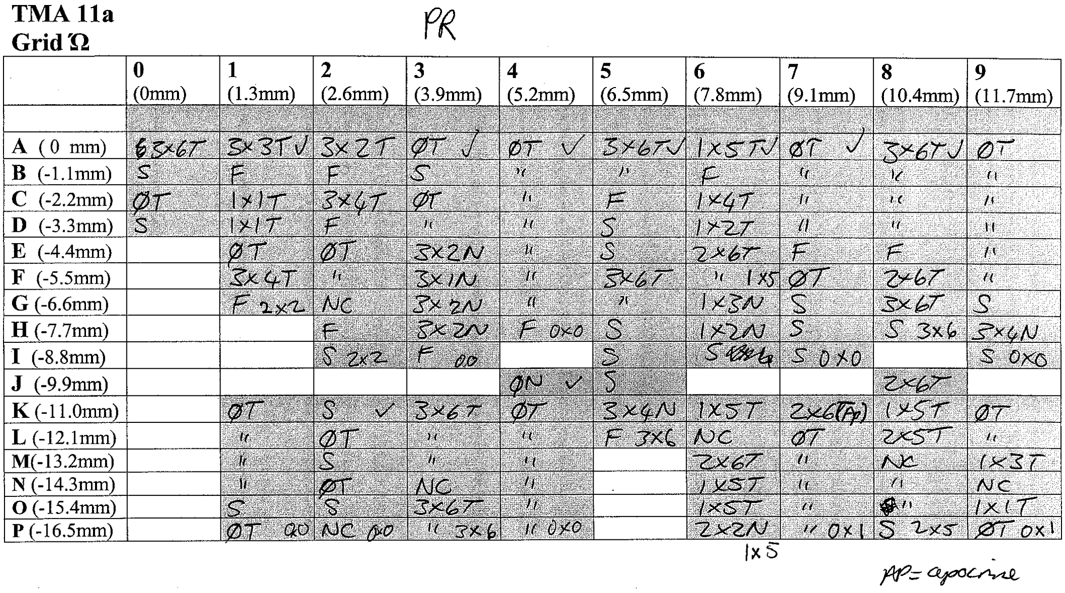
A single TMA section can therefore be used to test a given biological marker on hundreds of cancer specimens from multiple patients, whereas multiple sections cut from the same TMA block can be used to test multiple markers on the same set of specimens. Camp et al. [18] have concluded that two TMA cores per patient are sufficient to assess the expression of oestrogen receptor and progesterone receptor in specimens of invasive breast carcinoma.

Once IHC has been carried out, the assessment by pathologists of the stained breast TMA sections starts with the classification of each tissue spot. This initial step must be carried out prior to assessing the immunostaining, given that the cylindrical cores extracted from donor blocks and embedded into the TMA block are not always homogeneous throughout their depth (for example, there may be a region of tumour in the top third of a core, while its remainder contains only stroma). Therefore, to ensure correct analysis, each tissue spot on the TMA section should first be classified as to the type of tissue present.

In the experience of the pathologists working at the Ninewells Hospital in Dundee, spots usually belong to one of several types, namely tumour, normal, stroma, fat, blood, and invalid (spot not present or not assessable). The first four of these types are the most frequent. Both tumour and normal spots contain at least some epithelial tissue, but in tumour spots at least some tumour tissue is present, whereas in normal spots all epithelial tissue is healthy. Stroma and fat spots do not contain any epithelial tissue, but only connective and/or fatty tissue.

The aim of the classification process is to discard stroma and fat spots (as well as infrequent blood and invalid spots). For those spots classified as belonging to either the tumour or the normal type, the degree of immunostaining (that is, the level of expression of the protein of interest, such as progesterone receptor) is then assessed and assigned a quickscore. Once all of the spots have been scored, their scores can be compared. Figure 2.4 shows a typical manual scoring sheet, used by a pathologist to register the classes and scores of one half of a TMA slide containing 18 columns of spots.

Clearly, as the number of tissue spots on TMA slides increases and the complexity



TMA 11a
Grid Ω

PR

	0 (0mm)	1 (1.3mm)	2 (2.6mm)	3 (3.9mm)	4 (5.2mm)	5 (6.5mm)	6 (7.8mm)	7 (9.1mm)	8 (10.4mm)	9 (11.7mm)
A (0 mm)	~~6~~3x6T	3x3T ✓	3x2T	ØT ✓	ØT ✓	3x6T✓	1x5T✓	ØT ✓	3x6T✓	ØT
B (-1.1mm)	S	F	F	S	"	"	F	"	"	"
C (-2.2mm)	ØT	1x1T	3x4T	ØT	"	F	1x4T	"	"	"
D (-3.3mm)	S	1x1T	F	"	"	S	1x2T	"	"	"
E (-4.4mm)		ØT	ØT	3x2N	"	S	2x6T	F	F	"
F (-5.5mm)		3x4T	"	3x1N	"	3x6T	" 1x5	ØT	2x6T	"
G (-6.6mm)		F 2x2	NC	3x2N	"	"	1x3N	S	3x6T	S
H (-7.7mm)			F	3x2N	F 0x0	S	1x2N	S	S 3x6	3x4N
I (-8.8mm)			S 2x2	F 00		S	S [illegible]	S 0x0		S 0x0
J (-9.9mm)					ØN ✓	S			2x6T	
K (-11.0mm)		ØT	S ✓	3x6T	ØT	3x4N	1x5T	2x6(Ap)	1x5T	ØT
L (-12.1mm)		"	ØT	"	"	F 3x6	NC	ØT	2x5T	"
M(-13.2mm)		"	S	"	"		2x6T	"	NC	1x3T
N (-14.3mm)		"	ØT	NC	"		1x5T	"	"	NC
O (-15.4mm)		S	S	3x6T	"		1x5T	"	[illegible]"	1x1T
P (-16.5mm)		ØT 00	NC 00	" 3x6	" 0x0		2x2N	" 0x1	S 2x5	ØT 0x1

1x5

AP= apocrine

Figure 2.4: A typical manual scoring sheet for one half of a tissue microarray slide.

of the staining (aimed at different cell types and different cellular compartments) also increases, there is potential for variations in the quality of data collected. Given that TMA sections may contain hundreds of tissues spots, the classification and scoring exercise is time-consuming and has the potential for perceptual errors, inter- and intra-observer variability, and severe quantisation that leads to the loss of potentially valuable information. For these reasons, there is strong clinical and research-related motivation for the development of automated methods for quantitative analysis of breast TMA image data. Chapter 3 discusses current research on automated analysis of cancer tissue sections in the context of bright field microscopy, and identifies some existing commercial systems.

2.5 Data

The data used in this work originated from the National Cancer Research Institute's Adjuvant Breast Cancer (ABC) chemotherapy trial, carried out between 1992 and 2000 [1]. Pathologists at the Ninewells Hospital in Dundee have constructed a TMA bank for the 112 patients entered into the ABC trial from Dundee. TMA slides created from this bank were subjected to various forms of IHC. In particular, four of those slides were subjected to progesterone receptor IHC. The type and the quickscore of each spot in these four slides, as assessed by a pathologist, were known.

A total of 364 spots were randomly selected from those four slides, with the restriction that approximately equal numbers of spots were picked for each of the four main types, namely tumour, normal, stroma, and fat. In this work, the provenance of the selected spots (that is, which patients they originated from) was not taken into account. From digitised images of the four slides, colour images of the 364 selected spots were then manually extracted. These images of TMA spots and the expert-assigned types and scores associated with them constituted the data used in this work.

The spot types and quickscores for the four TMA slides immunostained for progesterone receptor were assessed by the same pathologist on two separate sessions. Thus, data on manual classification was available not only for the 364 spots selected for this work, but for a wider total of 935 spots. For those spots classified as either tumour or normal, the quickscore was also known. In particular, of the 935 spots, 686 were assigned to either the tumour or the normal type on *both* sessions. As reported in Sections 4.5 and 6.4, these data made it possible to obtain estimates of intra-observer variability. However, it should be noted that approximately five years elapsed between the two scoring sessions and, on both occasions, the assessment was carried out "on glass" and not "on screen" (that is, by observing the physical TMA slides under a microscope, as opposed to observing digitised slide images on a computer monitor). The detected variability may therefore have been partly due to degradation of the tissue samples. Unfortunately, no inter-observer data was available.

It is worth pointing out that the scoring carried out by the pathologist included a minor deviation from the quickscoring method proposed by Detre et al. [30], in that the proportion of epithelial nuclei that were immunopositive was not scored between 1 and 6, but rather between 0 and 6 (0 corresponding to a total absence of immunopositive nuclei, and 1 corresponding to less than 5%).

The IHC to which the four source slides were subjected involved the progesterone receptor mouse monoclonal antibody from Novocastra Laboratories Ltd (catalogue number NCL-PGR-312/2) at a dilution of 1/800, as well as microwave pressure antigen retrieval. This type of IHC results in nuclear staining, observable with a light microscope. As shown in Figure 2.5, the nuclei of immunopositive epithelial cells display a brown colour, caused by the labelling of the used antibody, while the colour of immunonegative nuclei remains blue, due to a light haematoxylin counter-staining.

The images of the 364 spots selected for this work were manually extracted from digitised TMA slide images acquired using a ScanScope digital slide scanner from Aperio Technologies, Inc. These images had a resolution of 0.23 μm/pixel and a

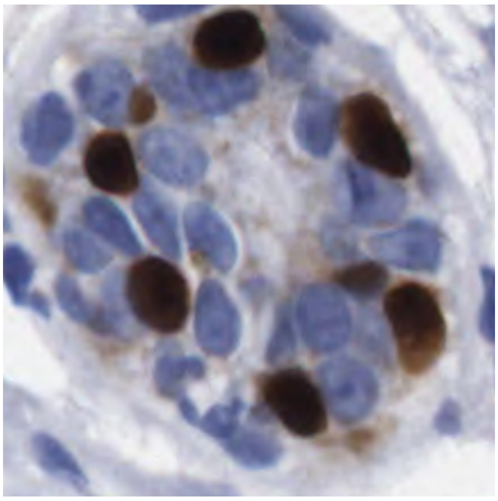


Figure 2.5: Detail of breast tissue section, showing the effect of progesterone receptor immunohistochemistry, with immunopositive nuclei stained brown in contrast to blue immunonegative nuclei.

typical spot had a diameter of 700 μm (that is, about 3000 pixels). Each image contained three colour channels, namely red, green, and blue, and each channel had a depth of eight bits. These images had JPEG 2000 compression with a compression quality of 70%. Figure 2.3(b) previously showed an example of a spot image.

***

This chapter provided an overview of clinical concepts closely related with the present work, namely those of breast cancer, immunohistochemistry, scoring of immunostained histological sections, and tissue microarray technology. The data used in this work was also introduced.

The next chapter will present a review of literature relevant to this work, as well as a brief overview of existing commercial systems.

Chapter 3

Review of literature on histological image analysis

This chapter deals mainly with recent literature on automated analysis of histological images of breast tissue obtained through bright field microscopy, although other literature is also cited when relevant (mostly focusing on other types of tissue). The literature was reviewed while keeping in mind that all image analysis techniques incorporate underlying models, even if such models are often not made explicit in method descriptions. The reviewed methods are organised (and, when appropriate, broken down into distinct stages) into the three first sections in a way that reflects the *scale* of modelling involved: cells or nuclei in Section 3.1; multi-cellular structures or distinct regions in the tissue in Section 3.2; and whole sections of tissue in Section 3.3.

Section 3.4 identifies a number of existing commercial tools whose functionality can assist the scoring of tissue microarray spots. Finally, Section 3.5 presents some conclusions and gives an overview of the work reported in this thesis in the light of the reviewed literature.

The author is aware of existing literature reviews that are pertinent to the subject of this thesis. A relatively old survey by Materka and Strzelecki [63] focuses specifically on texture analysis. Although this survey is not concerned with tissue images in particular, several of the reviewed systems deal with such images. In contrast, two more recent surveys focus specifically on histological images. Loukas and Linney [60] present an overview of analysis methodologies with emphasis on the assessment of certain biological factors that influence the outcome of radiotherapy, while the survey by Demir and Yener [29] deals more generically with automated cancer diagnosis. A review by Zhu et al. [104] focuses on computer-aided approaches to the diagnosis and staging of prostate carcinoma, and addresses some systems that deal with images of tissue. Very recently, Gurcan et al. [41] have reviewed the state of the art in computer-assisted diagnosis technology for digitised histopathology.

3.1 Modelling of cells or nuclei

There is an immense body of literature on automated cytology, one of the earliest real-world applications of computer vision, dating back to the early 1970s. This section, however, deals with literature on automated histology, focusing mainly on recent work.

In the following two sections, the reviewed methods are divided into those that do not involve learning, as far as the modelling of cells or nuclei is concerned, and those that do rely on some form of learning.

3.1.1 Methods not involving learning

In some cases, objects are modelled merely in terms of the intensity or colour of the pixels that belong to them, without taking into account any spatial relationships between pixels. For example, in their analysis of sections of lung tissue stained with r-H2AX and PX-DAB antibodies, McKee and Land [65] assume that pixels belonging to nuclei, cytoplasm, and background form clusters in the RGB colour space. A kernel-based extension of fuzzy C-means clustering is then used to achieve segmentation. Similarly, Arif and Rajpoot [4] assume that all pixels belonging to nuclei in sections of prostate tissue stained with haematoxylin and eosin (H&E) form a cluster in the grey-level intensity space. K-means clustering is then used to segment nuclei. The quality of the segmentation achieved by this type of methods is not good; cells that are close together tend not to correspond to individual segments, but rather become merged into larger segments, and the boundaries of segments are not smooth. In fact, in both cases cited above, clustering is employed merely to achieve an initial coarse segmentation that is later enhanced.

Several methods, like those cited in the following paragraphs, take into account prior knowledge about the spatial coherence of the objects being modelled. This can take the form of very simple assumptions, such as that the pixels belonging to certain objects are locally similar in intensity or colour.

Gurcan et al. [40], for example, detect and segment nuclei in H&E-stained sections of neuroblastoma, employing so-called morphological top-hat reconstruction (which involves a number of morphological operations) and hysteresis thresholding (a form of adaptive thresholding). McKee and Land [65] incorporate assumptions about spatial coherence into heuristically based mechanisms, used to improve a previous coarse segmentation of nuclei and cytoplasm. Both these methods successfully separate nuclei of cells that slightly touch, but have difficulty in separating nuclei of cells that either press against each other or overlap. Considerably better results are achieved by Jones et al. [47]. In their work, drosophila cells stained for DNA and actin are modelled by assuming that pixels belonging to cells are close under a metric that takes into account pixel position and edge information. Given an image where seed regions have been pre-segmented, cells are accurately segmented by assigning each pixel to the closest seed.

Other methods incorporate more sophisticated prior knowledge about shape. For example, Mouroutis et al. [66] model nuclei in sections of H&E-stained laryngeal tissue as being approximately circular, with distances from boundary points to centroid that follow a normal distribution. The compact Hough transform is used to localise nuclear centroids. An initial set of boundary points is obtained by finding the maximum edge magnitude along a number of radial directions, and likelihood maximisation is used to improve the boundary. Although this method yields very good results, its use is limited to relatively small images, due to the reliance on the computationally intensive Hough transform. Arif and Rajpoot [4] model candidate nuclei as corresponding to points in a low-dimensional manifold, embedded in a high-dimensional space of boundary information. From a previous coarse segmentation of candidate nuclei, large vectors of centroidal distances are extracted and transformed via Fast Fourier Transform. A diffusion map-based framework is then employed to obtain the positions of objects of interest in a 2-dimensional manifold. Most nuclei and non-nuclei fall into distinct areas of this manifold and can thus be identified. This technique of unsupervised learning of shape manifolds is comprehensively discussed by Rajpoot et al. [78].

3.1.2 Methods involving learning

Most of the methods that involve learning are also supervised, in the sense that the training of models relies on data annotated by experts.

Models used in supervised methods, too, may ignore spacial relationships between the pixels that belong to the modelled objects, therefore incorporating no information about texture or shape. Dalle et al. [28], for example, use training images to obtain Gaussian colour models for three types of epithelial cells and for candidate mitotic cells, in sections of H&E-stained breast tissue. These models are first used to segment indistinctly all cells present in neoplasm regions identified in whole-slide images. Then, measures of similarity between the colour distribution within the detected segments and the colour models are used to classify segments into epithelial cells (of each of the three considered types) and candidate mitotic cells.

In turn, several models, like those referred to in the following paragraphs, incorporate basic learned information about texture or shape.

Spyridonos et al. [93] simply extract local textural features and train a neural network to classify pixels as belonging or not to nuclei, in H&E-stained sections of urinary bladder carcinoma. This work has been applied by Glotsos et al. [38] to isolate nuclei in H&E-stained sections of astrocytomas (neoplasms of the brain).

Often, information on shape or texture takes the form of object-level, summary features that characterise pre-segmented cells or nuclei. In the work of Dalle et al. [28], training images are used to obtain Gaussian models for mitotic and non-mitotic cells, based on the mean and variance of the cell's intensity, as well as on roundness, eccentricity, and area. These models are then used to classify candidate mitotic cells. McKee and Land [65] extract features such as mean and variance of nuclear intensity, nucleo-cytoplasmic ratio, cell size, and average nuclear radius, from previously segmented cells. A support vector machine is then trained to classify cells into normal and cancerous, based on the extracted features.

Other models incorporate more complex information about the shape or spatial coherence of objects of interest. The analysis of prostate tissue sections stained with H&E reported by Begelman et al. [6] involves an unsupervised stage, in which the colour of pixels that belong to nuclei, glands, and stroma is modelled as a Gaussian mixture with three components. The parameters of this mixture are determined by likelihood maximisation. Given a new image, this colour model allows to obtain three class probability maps. A supervised stage is also implemented, in which training images containing manually selected nuclei are averaged together to obtain a mean intensity model of the nucleus. Given a new image, this model is used to obtain a correlation map. A set of fuzzy rules is designed to classify pixels based on the obtained probability and correlation maps, so that a fuzzy logic engine may segment the nuclei. Similarly, Petushi et al. [72] manually pick small windows from images of H&E-stained breast tissue sections, to learn mean intensity models for different types of cell (inflammatory, normal, tumour, and stroma) and fat. Optimal adaptive thresholding is used in combination with morphological operations that make use of the learned models, to segment candidate blobs in larger images. A binary decision tree is trained to classify blobs into different cell types, based on summary features such as mean and variance of blob intensity and area. Lee and Street [58] assume that the boundaries of nuclei in breast tissue sections are approximately elliptical. A set of elliptical templates are tested through the iterative generalised Hough transform (IGHT), to determine which templates match benign and malignant nuclei, and how often. These templates are used to initialise snakes that run to convergence, yielding a set of flexible templates with an averaged shape and an uncertainty region. Given a new image, the flexible templates are used by IGHT to detect and classify nuclei, based on the majority class count of the matched template. Lee and Street [59] extend this work, to incorporate a neural resource allocating network that learns to cluster shapes and to classify nuclei.

3.2 Modelling of multi-cellular structures and tissue regions

In the following sections, the reviewed methods are divided into those that involve the modelling of multi-cellular structures such as mammary glands or tubules, for segmentation or detection purposes, and those that model regions in the tissue that do not constitute multi-cellular objects, such as regions of tumour, stroma, or inflammatory cells.

3.2.1 Multi-cellular structures

Petushi et al. [73] rely on the previous identification of cells, to detect tubules in sections of breast tissue stained with H&E. This is done simply by modelling tubules as regions of high pixel intensity surrounded by a string of cells. An equally straightforward approach is taken by Dalle et al. [28] on the same type of specimens, but now relying on the previous segmentation of certain regions in the tissue. Tubular formations are detected by applying a morphological closing operator and filling to previously segmented regions of neoplasm, and by identifying blob structures that contain fat or lumen. Both these approaches are somewhat simplistic and in principle would need to be re-tuned in order to yield satisfactory results with images of other types of tissue. A more flexible (if less recent) approach is proposed by Fernandez-Gonzalez and de Solorzano [35], who analyse mammary gland tissue sections immunostained for human epidermal growth factor receptor 2 (HER2) to detect structures such as ducts, end buds, and alveoli. Delaunay triangulation is used to obtain a graph whose nodes are previously identified cells (classified as either immunopositive or immunonegative), and a relative neighbourhood graph is built on top of the triangulation. To detect structures of interest, cells are clustered based on the distance between cells that maximises a certain measure of their neighbourhood relationships.

In turn, Naik et al. [68] propose a method for the detection and segmentation of glands (as well as individual nuclei) in sections of prostate or breast tissue. A Bayesian classifier is trained to generate the likelihood of each pixel belonging to an object of interest. Level sets, as well as template matching using shape models, are then applied to the likelihood scenes to achieve object segmentation. Objects are validated based on structural constraints imposed by domain knowledge. Whereas the above-cited method by Fernandez-Gonzalez and de Solorzano [35] outputs simply a sub-graph formed by the centroids of nuclei that belong to each gland, this method delineates the actual boundaries of detected glands, which may be a more interesting visual output from the pathologists' point of view.

3.2.2 Tissue regions

Dalle et al. [28] begin their analysis of breast tissue sections stained with H&E by applying Otsu colour thresholding to a low-resolution version of each image. Morphological opening and closing operations then allow the localisation of regions of neoplasm. Petushi et al. [73] resort to similar techniques to analyse the same type of specimens, although relying on the previous segmentation of nuclei (classified as inflammatory, normal, or tumour). By scanning the segmented image with a small window, density maps are obtained that reflect the concentration of nuclei belonging to each of the three considered types. Adaptive optimal thresholding, standard morphological filling, and edge smoothing are then applied to these maps, to segment high nuclei density areas of the three types.

Karaçali and Tözeren [48], too, analyse breast tissue sections stained with H&E, to segment certain regions of interest. K-means clustering is applied to grey-level information to find candidate regions (with k=3, for chromatin-rich, stromal, and unstained regions). A 2-component Weibull mixture is fit to the luminance information within the candidate regions. The maximum likelihood threshold between components permits the distinction of foreground regions (chromatin-rich and stromal) from background (unstained). Modified k-means clustering is then applied to the *a* and *b* colour information within foreground regions converted to an *Lab* colour space, to distinguish between chromatin-rich and stromal regions. Kong et al. [53] employ a segmentation method that uses an expectation maximisation (EM) algorithm with the Fisher-Rao criterion as its kernel. This method is used to segment not only sub-cellular compartments (nuclei and cytoplasm), but also regions of neuropil, in sections of peripheral neuroblastic tumours. Although more complex and computationally intensive, these two approaches are more principled than the methods cited in the previous paragraph, and would be more likely to produce good results if applied to images of different types of tissue.

3.3 Modelling of tissue sections

In the following two sections, the reviewed methods are divided into those that do not involve learning, as far as the modelling of whole tissue sections is concerned, and those that do rely on some form of learning.

3.3.1 Methods not involving learning

Typically, methods that dispense with learning rely on the previous modelling and identification of structures within the tissue section. Features that characterise those pre-detected structures are then summarised by means of simple formulas, which constitute the actual model of the tissue section. Spatial relationships between pre-identified structures are disregarded by the model.

Dalle et al. [28] analyse breast tissue sections stained only with H&E, to predict Nottingham grades. Three individual scores are computed, namely: a tubule formation score, from the amount of previously detected tubules; a nuclear pleomorphism score, from the proportions of previously segmented epithelial cells of three types; and a mitotic score, from the amount of previously detected mitotic cells. These scores are combined to obtain the global Nottingham grade of the tissue section.

Several methods deal with images of tissue sections subjected to some form of nuclear immunostaining, to predict measures that are closely related to the scores of tissue microarray spots. Elie et al. [32], for example, analyse sections of ovarian adenocarcinomas immunostained against cyclin A (for the characterisation of cellular proliferation), to predict two measures that characterise each tissue section. The ratio between stained area and the surface of the epithelium is computed from the previous segmentation and classification of epithelial cells, whereas the number of so-called hot spots (regions of high concentration of stain) per unit area of epithelium is computed from the previous segmentation of hot spots. Weaver and Au [101] deal with sections of human solid tumour specimens (head, neck, and bladder) immunostained either for proliferating cell nuclear antigen or with bromodeoxyuridine. A labelling index is computed as the proportion of cells that are labelled (stained), based on the previous segmentation and labelling of nuclei. Similarly, Kostopoulos et al. [55] analyse breast carcinoma sections immunostained with diaminobenzidine, to determine the percentage of epithelial nuclei that are stained. This value is computed from the previous segmentation and classification of epithelial nuclei, and allows predicting the oestrogen receptor status of the tissue section (positive if the percentage is above 20%). Sont et al. [92] assess inflammatory cell counts and cytokine expression in immunostained sections of bronchial tissue. Brown-red staining is separated from blue counter-staining through a segmentation procedure that essentially manipulates the RGB colour components. The cell count is then estimated from the area and morphometric characteristics of stained regions, whereas the cytokine expression density is estimated by the average grey level of those regions.

Methods that model tissue sections without resorting to any form of learning are in general less reliable and flexible than methods that involve learning. For example, in the above-cited work of Dalle et al. [28], three individual scores are directly derived from previous detection results, such as the number of tubules and the proportions of cells of different types. However, this assumes a possibly unrealistic linear relationship between scores and detection counts. This assumption might be avoided by using some of the available data to train a classifier, to predict scores based on detection counts. Similarly, Kostopoulos et al. set a 20% threshold of stained nuclei for immunopositivity, which might have to be adjusted for different types of tissue or nuclear stain. A binary classifier trained to predict immunopositivity based on the proportion of stained nuclei would obviate the need for a free parameter and almost certainly achieve better results. The use of classifiers could also adequately extend to methods reported by Elie et al. [32] and by Sont et al. [92], so as to convert the predicted measures (proportion of stain, cell count, or expression strength) into actual discrete scores used by pathologists.

3.3.2 Methods involving learning

Most of the methods that involve learning are also supervised, in the sense that model training relies on data annotated by experts. Typically, the trained model consists of a data structure that is used for classification purposes, such as a binary decision tree or a neural network.

There is an extensive body of literature on supervised methods for automated grading of cancer, including the methods cited in the following paragraphs.

Petushi et al. [73] predict the Nottingham grade of breast tissue sections subjected to H&E staining. The previous segmentation of cells (classified into several types), tubules, and areas of high density of nuclei allows the computation of four measures, namely the numbers of inflammatory nuclei, normal nuclei, tumour nuclei, and tubules per unit high density area. In contrast with the work of Dalle et al. [28] cited above, a classifier (such as a neural network or a decision tree) is then trained to predict the tissue section's grade, based on the computed measures.

Weyn et al. [102] train a nearest-neighbour classifier to predict the Bloom-Richardson grade of invasive breast cancer sections, based on wavelet features that reflect chromatin texture and are computed from images of nuclei previously isolated within each tissue section. In the work of van de Wouwer et al. [97], a classifier of the same type is trained to assess the Bloom-Richardson grade of breast tissue sections subjected to Feulgen staining (used to identify chromosomal material in cells), based on chromatin texture. The employed textural descriptors include wavelet energy parameters, as well as statistics of the image's grey levels. Wavelet features are adopted also by Jafari-Khouzani and Soltanian-Zadeh [46], who use nearest-neighbour classification to predict the Gleason grade of H&E-stained sections of prostate tissue, from energy and entropy features of multi-wavelet coefficients computed for each whole-section image. Simulated annealing is employed to select the most discriminative features.

Other methods concerned with the prediction of Gleason grades include that proposed by Doyle et al. [31], in which a support vector machine is trained to classify H&E-stained slides of prostate tissue as either benign epithelium, benign stroma, Gleason grade 3 adenocarcinoma, or Gleason grade 4 adenocarcinoma. Classification is based on a large number of graph-based, morphological, and textural features that capture the arrangement of nuclei and glandular structures within each section. In turn, Tabesh et al. [94] combine the use of colour, texture, and morphometric features computed both at the global level and at the histological object level, in the prediction of high and low grades of prostate tumour sections stained with H&E. The sequential forward feature selection algorithm is employed, in conjunction with different types of classifier. In contrast with the methods cited in the previous paragraph, these two approaches make use of not only textural information but also features that reflect spatial relationships between pre-detected nuclei, illustrating a tendency of the more recent automated grading strategies.

Methods that deal with the grading of other types of cancer include those proposed by Glotsos et al. [38] for malignancy grading of astrocytomas (neoplasms of the brain) and Spyridonos et al. [93] for grading of urinary bladder carcinoma, both employing morphological and textural nuclear features. In turn, Keenan et al. [50] address the grading of cervical intra-epithelial neoplasia, making use of Delaunay triangulation and triangle features that include the mean area, the mean edge length, and the number of triangles per unit area.

Methods such as those cited in the following paragraphs have goals that are not far from the classification of tissue microarray spots into types, although focusing in most cases on sections stained only with H&E.

Karaçali and Tözeren [48] rely on the previous segmentation of epithelial and stromal regions in sections of breast tissue, to compute a so-called texture profile for each image. Each profile consists of the percent areas covered by chromatin-rich nuclei and by collagen-rich stroma, along with a measure of spatial heterogeneity. The probability densities of the texture profiles are learned for three classes, namely: specific to normal tissue; specific to cancerous tissue; and not specific to either. Given a new image, log-likelihood ratios are employed to achieve its classification. Brook et al. [15], too, analyse breast tissue sections stained with H&E to identify healthy and tumour sections, but, in their case, two types of tumour are considered. The images are converted to grey-level through principal component analysis, and level sets are obtained for them. A histogram of connected component sizes is computed for each level set, so that each image is characterised by a set of histograms. A support vector machine is trained to classify sections into healthy, tumour in situ, and invasive carcinoma, based on histogram sets.

Like Brook et al., Komosinski and Krawiec [52] extract for each image a relatively small, fixed-size global feature vector in the form of a histogram, although following a different approach. Images of neuroepithelial (astrocytic and glial) tumour sections are processed using a region growing technique, and four summary features are extracted for each segmented region (area and mean hue, saturation, and intensity). Hierarchical cluster analysis is used on training data to obtain a dictionary of cluster centres in the four-dimensional feature space. Given a new image segmented by region growing, a histogram of region counts per cluster centre is determined (each region being assigned to the closest centre). Evolutionary feature weighting is used on further training data to learn the relative importance of each histogram component. Images are classified into six types of astrocytic tumour or into six types of glial tumour, through nearest-neighbour classification. It is interesting to note that simply computing a set of global summary statistics (such as the means of the area, hue, saturation, and intensity over all segmented regions) would equally achieve the goal of reducing each image to a fixed-size global feature vector; however, this vector would have much less descriptive power than that obtained via the above-described “bag-of-features” approach.

Esgiar et al. [33] deal with tissue sections subjected to a form of immunohistochemistry, employing fractal analysis to differentiate between normal and cancerous sections of colon tissue. Specifically, nearest-neighbour classification is used to classify tissue sections immunostained for cytokeratins, based on fractal dimension features.

Some methods model different types of tissue with the aid of graphs whose nodes are associated to pre-detected cells or nuclei. Gunduz et al. [39], for instance, resort to topological information, to classify samples of glioma (a type of brain cancer) into cancerous, healthy, and non-neoplastic inflamed. Graph edges are computed via a decaying exponential function of the Euclidean distance between every pair of cells. A neural network is then trained to classify tissue samples based on graph metrics that include the clustering coefficient, eccentricity, and closeness for each cell. For some applications, however, this type of simple cell-graph approach may be insufficient, as the composition of the extracellular matrix (ECM) surrounding each cell is ignored. Bilgin et al. [9] address this limitation by proposing a method called ECM-aware cell graph mining, to classify H&E-stained bone tissue samples as healthy, fractured, or cancerous. A colour code is assigned to each node based on the composition of the ECM that surrounds the associated cell, and graph edges are established between nodes when the membranes of the corresponding cells are in contact and have similar colour. Support vector machines are then trained to classify tissue sections based on topological and spectral graph features.

In very recent work, Masood and Rajpoot [62] employ spatial analysis of hyperspectral image data to achieve classification of colon tissue patterns into benign and malignant. From the available spectral bands, a single band is selected and analysed using circular local binary pattern (CLBP) features. A feature selection method is proposed, allowing to determine the best feature set before classification. This method is based on measures of clustering quality. Classification using support vector machines with a Gaussian kernel yielded an accuracy of 90%. In turn, Qureshi and Rajpoot [77] address the problem of classifying images of meningioma into four subtypes. The Adaptive Discriminant Wavelet Packet Transform (ADWPT) is compared with three popular texture analysis feature sets, namely Local Binary Patterns, Grey Level Co-occurrence Matrices, and Gabor Transforms. Classification is achieved by support vector machines with a Gaussian kernel. This study concludes that ADWPT is a superior technique for meningioma classification, achieving accuracies above 90%. This technique is comprehensively discussed in the thesis authored by Qureshi [76].

Sanders et al. [87] developed a system to score tissue microarray spots of various types and immunostained for each of several antibodies, by quantifying staining strength and fraction of cells stained. Sub-cellular compartments (membranes, cytoplasm, and nuclei) are previously segmented and classified into stained and non-stained. Four staining strength scores are considered, namely negative, weak, moderate, and strong. For both training and test images, a number of normalised global features are computed, such as the number of stained nuclei and the average stained nucleus intensity. For each global feature, the midpoints between the means of adjacent strength scores are learned from the annotated training images (these midpoints thus constitute a learned model of staining strength). Given a test image, each of its global features can thus be assigned to a strength score, and these partial scores are combined into a final staining strength score. Interestingly, this simple scoring approach was found to perform at the same level as an alternative method based on support vector machines. In turn, percent staining is modelled simply as the percentage of nuclei that were stained (or non-stained with surrounding stained cytoplasm or stained membrane).

3.4 Commercial systems

Several commercial tools exist that assist in analysing images of tissue sections subjected to nuclear staining immunohistochemistry, for example against oestrogen receptor or progesterone receptor. These systems include the Immunostaining Quantification module of S.CO LifeSciences, the Tissue Image Analysis product of SlidePath, and the Digital IHC Solution of Aperio Technologies, as well as Ariol of Genetix / Applied Imaging, the Ventana Image Analysis System of TriPath

Imaging, the Automated Cellular Imaging System of Chromavision / Clarient, the Quantitative Cellular Assessment system of Cell Analysis, and IHCscore of Bacus Labs. Analysis results normally consist in counts of immunopositive and immunonegative nuclei, but some systems also assist in determining the strength of staining of immunopositive nuclei.

Some academic literature reports work that is based on commercial systems. For example, Turbin et al. [96] trained the Ariol software to analyse oestrogen receptor (ER) expression in breast carcinoma tissue microarrays. It was concluded that the prognostic significance of ER positivity determined by automated quantitation did not differ from that determined by human scoring. Thus, this study did not directly compare scores obtained automatically with scores assessed by histopathologists; rather, automated and human scores were dichotomised between ER positive and ER negative and their prognostic significance compared. In turn, the system commercialised by Chromavision has been used Weaver et al. [100] in the detection of micro-metastases in sentinel lymph nodes in breast cancer.

Of the systems mentioned above, five have received pre-market approval from the U.S. Food and Drug Administration (FDA), for quantification of biomarker expression as an aid in diagnosis. These systems are commercialised by Aperio, Genetix / Applied Imaging, TriPath, Chromavision / Clarient, and Cell Analysis. In order to obtain FDA approval, the level of concordance between manual and automated image analysis must be assessed [24]. In practice, however, the available methods require high levels of user interaction: analysis remains time-consuming. In particular, in systems such as those of Aperio and Bacus Labs, the pathologist is required to manually outline regions of interest in the tissue sections [26].

Aperio Technologies also commercialise the Genie histology pattern recognition tool, which can be trained from annotated data to segment tissue sections into regions, for example of normal and tumour epithelial cells, stroma, or fat. Clearly, this could assist in automating the classification of tissue microarray spots into types. However, trials carried out by the author (as reported in section 5.6) and by an experienced pathologist reveal that the achieved segmentation results are still very unreliable.

Camp et al. [19] have developed a set of algorithms called AQUA (Automated QUantitative Analysis) for the automated analysis of tissue microarrays. These algorithms, however, require the use of fluorescent tags as well as the acquisition of out-of-focus images. Fluorescent tags are used by a pixel-based locale assignment algorithm, to distinguish tumour from stroma and to detect subcellular compartments; the expression of a target antigen can then be quantitatively assessed from its co-localisation with these compartments. In addition, a rapid exponential subtraction algorithm is employed to overcome difficulties associated with overlapping subcellular structures, by subtracting an out-of-focus image from an in-focus image; this improves the assignment of pixels to particular subcellular compartments. In the analysis of breast tissue microarray spots immunostained for oestrogen receptor, a high degree of correlation was found between automatic scoring and pathologist-based H-scoring. Moreover, automated analysis was shown to have slightly better reproducibility than manual analysis. In this same study, the expression of beta-catenin in colon cancer was also analysed. The AQUA software is commercialised by HistoRx, Inc., together with dedicated instrumentation for image acquisition. A considerable fraction of recent work on automated quantitation of antigen expression in tissue microarray data reports the use of this technology, with applications to a wide variety of tissue types, including prostate cancer [86], breast cancer [43], melanoma [8], epithelial ovarian cancer [74], oropharyngeal squamous cell cancer [75], and small cell lung carcinoma [37].

3.5 Conclusions

For many of the reviewed methods involving models of cells or nuclei, the detection, segmentation, or classification of objects of interest constitutes the ultimate goal, and not an intermediate step in the assessment of whole tissue sections. Methods such as those reported by Dalle et al. [28] and Petushi et al. [72, 73] represent preliminary steps in the grading of tissue sections, involving only H&E staining. It is interesting to note that many of the detection and segmentation techniques reported in recent literature are not more complex than those discussed in earlier literature.

Most of the reviewed methods that model multi-cellular structure or tissue regions do not involve any form of learning from expert-annotated data, as far as the modelling of multi-cellular structures or regions in the tissue is concerned. It is also worth noting that, often, the modelling and detection of tubular formations ultimately plays a role in the prediction of cancer grades, as in the methods reported by Petushi et al. [73] and Dalle et al. [28].

Some of the reviewed methods deal with the classification of whole sections into distinct types, but they typically focus on the distinction between tumour and normal tissue, and do not involve the identification of sections containing only connective or fatty tissue. Most of these methods deal with sections stained only with H&E, as opposed to sections subjected to some form of immunohistochemistry. The models underlying classification methods are typically discriminative, as opposed to generative.

Other reviewed methods estimate the proportion of epithelial nuclei that are immunopositive within tissue sections subjected to some form of nuclear immunostaining, an output that is akin to the first quickscore integer value used in the assessment of breast tissue. Nevertheless, no literature was found dealing the prediction of TMA quickscores, as defined by Detre et al. in 1995 [30]. Most of the reviewed approaches do not deal with the estimation of staining strength and normally rely on the accurate segmentation of individual nuclei. Several other reviewed methods are concerned with the "ranking" of tissue sections, but, in all cases, this corresponds to the prediction of Bloom-Richardson grades or variants thereof, for sections stained solely with haematoxylin and eosin. These methods make use of classifiers and not ordinal regression algorithms.

Several of the methods reported in the literature make use of classifiers that provide a probabilistic output. In general, however, this output is used merely to decide the predicted type or grade of a tissue section, by choosing the class associated with the highest probability.

In the present work, an approach was developed to classify breast TMA spots into their four main types, namely tumour, normal, stroma and fat, with the purpose of identifying tumour and normal spots that needed to be subsequently scored, while discarding spots containing only stroma and fatty tissue. The method was applied to spots subjected to a form of nuclear immunostaining, in addition to haematoxylin counter-staining. The classification performance of a multi-layer perceptron (MLP) was compared with that of a classifier based on LDAL models. Unlike the MLP, LDAL is a generative approach that tries to explain the modelled data. By associating distinct types of tissue with latent variables of the LDAL model, a method for the segmentation of TMA spots into regions of different types was also explored.

In addition, an approach was developed to predict the quickscores of tumour and normal breast TMA spots subjected to a form of nuclear immunostaining. Both quickscore integer values were predicted, to reflect the proportion of epithelial nuclei that were stained as well as the strength of their staining. The models trained to predict the quickscores of spots based on their global features were not classifiers, but rather Gaussian processes for ordinal regression. This type of model was expected to perform better than a classifier, given that it incorporates knowledge about the relative order between categories.

The posterior probabilities output by the MLP classifier (for classification) and by the ordinal regression algorithm (for scoring) were used to compute a simple measure of prediction confidence. This allowed to set confidence thresholds that helped to distinguish the "easier" spots that could be processed automatically with high confidence from the more "difficult" spots that should be referred for manual assessment.

***

This chapter presented a review of recent literature on automated analysis of histological images, focusing mainly on images of breast tissue obtained through bright field microscopy. The reviewed methods were organised in a way that reflected the scale of the modelled objects or regions of interest, namely cells or nuclei, multi-cellular structures or distinct regions in the tissue, and whole sections of tissue. Some conclusions were presented and an overview of the present work was given, in the light of the reviewed literature. In addition, a number of existing commercial tools whose functionality can assist the scoring of tissue microarray spots were identified.

The next chapter will address the classification of breast tissue microarray spots subjected to progesterone receptor immunohistochemistry into four main types.

Chapter 4

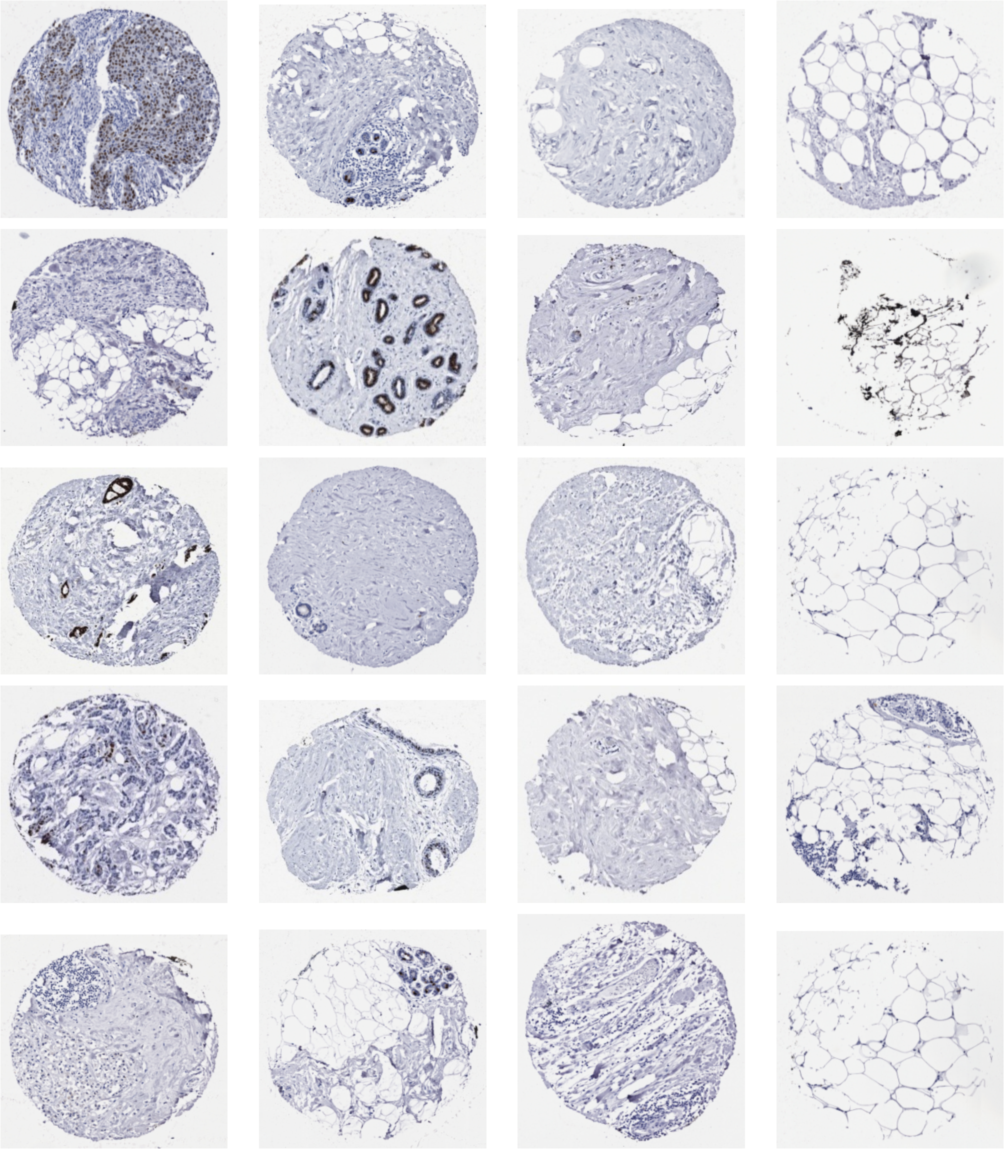
Classification of spots

4.1 Introduction

This chapter addresses the classification of breast tissue microarray (TMA) spots subjected to progesterone receptor immunohistochemistry into four main types, namely tumour, normal, stroma, and fat. Figure 4.1 shows examples of spots belonging to each of these types, illustrating the large inter- and intra-class variability.

The spatial orientation of cells, sub-cellular compartments, and arrangements of such structures in breast tissue sections varies greatly. This suggests that the use of local features invariant to rotation could lead to a better use of the training data, in the sense that features computed for each training pixel based on its neighbourhood help to teach the system to deal with test pixels with a similar neighbourhood, regardless of orientation. In this work, local texture was characterised through differential invariants computed as combinations of derivatives of two-dimensional Gaussians, which have been discussed by Schmid and Mohr [90] and theoretically studied by Koenderink and van Doorn [51]. Differential invariants are only one of a wide variety of local texture descriptors, including the more recent Gabor filters and wavelet transforms. However, the focus of this work was not on exploring different local descriptors, and differential invariants were found to perform reasonably well both in classification and in scoring experiments.

In contrast with cytology applications, the accurate segmentation of cells and intra-cellular compartments in histological data can be especially problematic, for reasons that include complex tissue structure, variable appearance, cell over-



(a) Tumour. (b) Normal. (c) Stroma. (d) Fat.

Figure 4.1: Examples of breast tissue microarray spots stained for progesterone receptor, for each of the four main types, illustrating inter- and intra-class variability.

lapping, and the presence of debris. This work explored the hypothesis that the classification of whole tissue spots would not need to rely on the detection or segmentation of individual nuclei. Rather, a computationally efficient system was implemented that approximated the joint probability distribution of local features by clusters in the feature space, and then characterised the appearance of each spot via a histogram of cluster frequencies. Spots were classified into the four main types based on their histograms. Thus, this work followed an approach similar to that reported by Varma and Zisserman [98] for statistical texture classification, although making use of texture features and a classifier of different types. A simpler alternative to the use of texton histograms would have been to compute a set of global statistical features (such as the popular Haralick features [42]) directly from the results of differential invariant filtering, and classify spots based on those statistical features. It was felt, however, that texton histograms would be able to characterise the textural content of the images better than a set of global summary features.

Classifiers are often based on discriminative models, which are trained purely for classification purposes, through the learning of discriminant functions. Given a test input $\mathbf{x}$, a classifier with probabilistic output determines posterior probabilities $P(t|\mathbf{x})$ for each of the C involved classes, $t \in \{T_1, ..., T_C\}$. But generative models, too, can be used for classification purposes. Such models aim at explaining the data, by modelling class-conditional distributions $P(\mathbf{x}|t)$ that can be sampled to synthesise (generate) random data. These distributions can also be used to form class posterior probabilities $P(t|\mathbf{x})$ via Bayes' rule, thus enabling the use of a generative model in classification tasks.

In this work, both a discriminative approach and a generative approach to classification were implemented, the former based on the generalised linear model (GLM) and the multi-layer perceptron (MLP) [10], and the latter based on latent Dirichlet allocation (LDAL) models [12]. GLMs are very quick to train and unlikely to overfit the data, due to their simplicity. However, they are also likely to underfit the data, hence the motivation to alternatively employ the MLP, whose complexity could be more appropriate to the data being modelled. In turn, the main motivation for the use of a generative model was the expectation that, given the power of such a model to explain the data, it could later be adapted to other interesting problems besides classification (such as the segmentation task addressed in Chapter 5).

The remainder of this chapter is structured as follows. Section 4.2 describes the techniques used to extract both local and global features from TMA image data.

Sections 4.3 and 4.4 provide an overview of neural network-based classification and of LDAL, respectively, and explain how these techniques were used to classify TMA spots into types. Section 4.5 presents the data and provides details on the experiments carried out and their results. Finally, Section 4.6 discusses the results and presents some conclusions.

4.2 Feature extraction

4.2.1 Extraction of local feature jets

Each pixel location was associated with a set of local features, namely the r, g, and b colour values and a set of grey-scale differential invariants. The latter were computed as explained in the following paragraphs.

Each original image was converted to grey-scale and down-sampled to both $^1/_4$ and $^1/_{16}$ of its original size through REDUCE operations, in order to build a three-level Gaussian pyramid with three levels. A REDUCE operation corresponds to the two-dimensional convolution of the input image with a $[\ 1\quad 4\quad 6\quad 4\quad 1\ ]/16$ pattern of weights, followed by down-sampling to half the input size on both dimensions [16]. The one-dimensional pattern of weights could be used to achieve two-dimensional convolutions, given the separability of two-dimensional Gaussian functions.

Six Gaussian derivative kernels were computed with a standard deviation $\sigma = 8$ pixels, as shown in Equations (4.1), where $G(\sigma)$ represents a two-dimensional Gaussian kernel and x and y denote the horizontal and vertical directions, respectively. It was assumed that $G_{yx}(\sigma) = G_{xy}(\sigma)$. These kernels had a radius of three standard deviations, and therefore were 49×49 pixels in size.

$$\begin{array}{rclrcl} G_x(\sigma) & = & \frac{\partial}{\partial_x} G(\sigma) & G_y(\sigma) & = & \frac{\partial}{\partial_y} G(\sigma) \\ G_{xx}(\sigma) & = & \frac{\partial^2}{\partial_x \partial_x} G(\sigma) & G_{yy}(\sigma) & = & \frac{\partial^2}{\partial_y \partial_y} G(\sigma) \\ G_{xy}(\sigma) & = & \frac{\partial^2}{\partial_x \partial_y} G(\sigma) & G_{yx}(\sigma) & = & G_{xy}(\sigma) \end{array} \tag{4.1}$$

The derivative kernels were then convolved with the whole Gaussian pyramid. In the case of spots containing epithelial nuclei (that is, spots of tumour or normal type), and given that the average nuclear radius was about 16 pixels, these kernels were expected to encompass parts of nuclei at the pyramid's base level, whole nuclei at the middle level, and nuclei and their immediate surroundings at the top level. The convolution results were finally used to compute the four differential invariants defined in Equations (4.2), where all multiplications are scalar products and each L term denotes the convolution of an image with a Gaussian derivative (for example, L_{xy} denotes the result of convolving an image with G_{xy}) [90]. In this case, the images involved in the convolutions are the individual levels of the Gaussian pyramid.

$$\begin{aligned} d_1 &= L_x L_x + L_y L_y \\ d_2 &= L_{xx} L_x L_x + 2L_{xy} L_x L_y + L_{yy} L_y L_y \\ d_3 &= L_{xx} + L_{yy} \\ d_4 &= L_{xx} L_{xx} + 2L_{xy} L_{yx} + L_{yy} L_{yy} \end{aligned} \tag{4.2}$$

Zero-order invariants were not used, as they represent merely the Gaussian smoothing of the grey-level image. This would in principle constitute redundant information, given that colour values were already included as local features.

The middle and top levels of the four resulting differential invariant pyramids were subjected to the necessary EXPAND operations. The effect of an EXPAND operation is to expand an $(M+1)$-by-$(N+1)$ array into a $(2M+1)$-by-$(2N+1)$ array by interpolating new node values between the given values, using the same pattern of weights as in REDUCE [16]. A total of 15 local features were thus associated with each pixel in the image, namely the r, g, and b colour values, and the four grey-level invariants for each of the three considered scales. The twelve invariants can be generically denoted as $d_{k,\sigma}$, $k \in \{1, 2, 3, 4\}$, $\sigma \in \{8, 16, 32\}$, and the set of 15 local features can be called a *jet*.

4.2.2 Extraction of global features

A proportion of local feature jets was randomly sampled from all spots in the data set (both training and test spots). The mean and variance of each individual local feature were computed over all training samples, to be used as normalisation constants. Using these constants, the sample jets of all spots were normalised to zero mean and unit variance.

A proportion of normalised sample jets was sub-sampled from the training spots.

K-means clustering was applied to these sub-samples, in order to determine the centres of a number of clusters in the multi-dimensional normalised local feature space. Essentially, the obtained set of cluster centres aimed to capture the range of colours and local textures characteristic of spot images, and could be called a *texton dictionary*. This procedure could also be seen as an attempt to approximate the density of local features by dividing their multi-dimensional space into irregular bins.

Nearest-neighbour classification was applied to the normalised sample jets of all spots (training and test), so as to assign each jet to the nearest texton in the dictionary. This constituted a vector quantisation step, in that each sample jet was quantised into a sample texton. A histogram of texton frequencies was then computed for each spot, based on its sample textons. Thus, the extraction of global features followed an approach similar to that used by Varma and Zisserman [98] in statistical texture classification.

4.3 Classification with neural networks

This section presents a brief overview of the types of neural networks used for classification in this work, based on Chapters 2, 4, and 5 from the book *NETLAB: Algorithms for Pattern Recognition* by Ian T. Nabney [67]. For a comprehensive discussion, that reference should be consulted, as well as the book *Pattern Recognition and Machine Learning* by Chistopher M. Bishop [10].

It should be pointed out that neural networks such as the GLM and the MLP belong to a range of available methods for computing and learning a discriminant function in classification problems. Other methods include, for example, the popular support vector machines (SVMs).

4.3.1 Single-layer networks

Single-layer networks can be referred to simply as generalised linear models (GLMs), although they constitute also an implementation of the statistical technique of linear regression. These models can be trained very quickly and, applied to a test data set, provide a useful baseline for comparison with more complex techniques. Due to their simplicity, GLMs are unlikely to overfit the data, but underfitting may represent a problem.

Considering a system with D inputs and C outputs, the GLM defines for each output an intermediate activation variable a_c ($c \in \{1, ..., C\}$) as a linear combination of the input variables x_d ($d \in \{1, ..., D\}$), as shown in Equation (4.3). The parameters of the model are the coefficients $w_{cd}^{(1)}$, which represent the elements of a weight matrix $\mathbf{W}$, and the bias parameters $w_{c0}^{(1)}$. The superscript $^{(1)}$ denotes the fact that the model contains a single layer of weights, which establish full connectivity between its inputs and outputs.

$$a_c = \sum_{d=1}^{D} w_{cd}^{(1)} x_d + w_{c0}^{(1)} \tag{4.3}$$

The actual outputs y_c ($c \in \{1, ..., C\}$) of the model are obtained by transforming the intermediate variables a_c through an activation function that should be adequate to the problem and data being handled. In regression problems, a linear function $y_c = a_c$ is appropriate. However, in this work the model was applied to classification problems with C mutually exclusive classes, hence the *softmax* activation function (which generalises the logistic sigmoid function) was used, as shown in Equation (4.4).

$$y_c = \frac{\exp(a_c)}{\sum_{c'} \exp(a_{c'})} \tag{4.4}$$

An advantage of using this kind of activation function is that the network outputs meet the requirements of probabilities and can therefore be seen as estimates of class posterior probabilities $P(t|\mathbf{x}, \mathbf{W})$, where t belongs to the set of classes $\{T_1, \ldots, T_C\}$, $\mathbf{x}$ is a vector of input variables, and $\mathbf{W}$ is the matrix of network weights (that is, the model's parameters). In a classification problem, given a test input $\mathbf{x}$, the predicted class t can be chosen as that associated with the highest class posterior probability. In this work, the actual probabilistic output of the network proved to be useful too, as discussed later in Sections 4.5 and 4.6.

4.3.2 Two-layer networks

Most practical applications of neural networks employ the multi-layer perceptron (MLP) architecture. Typically, this corresponds to a network with two layers of adaptive weights, each of which establishes a full connectivity: the first layer between inputs and a set of *hidden units*, and the second layer between hidden units and outputs. Ideally, the number of hidden units should be determined through cross-validation on training data. The generic architecture of a two-

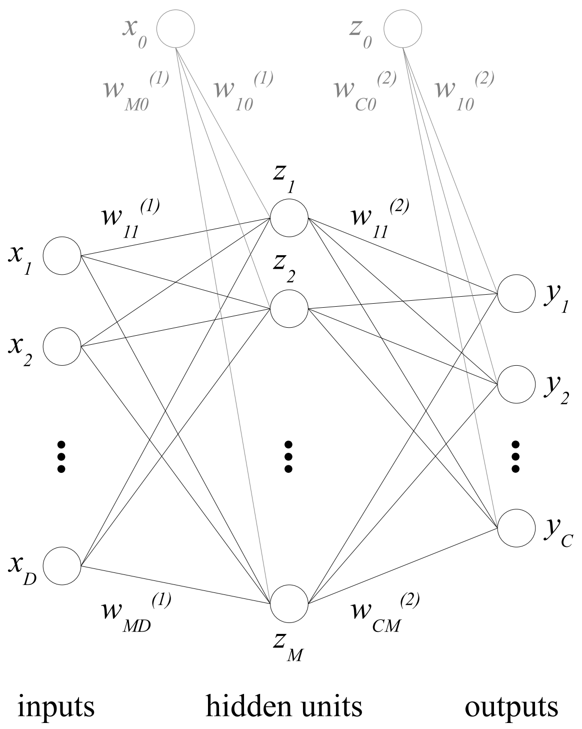


Figure 4.2: Generic architecture of a multi-layer perceptron with two layers of adaptive weights. (Based on Bishop [10].)

layer network is shown in Figure 4.2, whose notation is made clear in the following paragraphs.

Considering a system with D inputs, C outputs, and M hidden units, the first layer of the model defines for each hidden unit an intermediate activation variable $a_m^{(1)}$ ($m \in \{1, ..., M\}$) as a linear combination of the input variables x_d ($d \in \{1, ..., D\}$), as shown in Equation (4.5). The bias parameter associated with the hidden unit is denoted by $w_{m0}^{(1)}$.

$$a_m^{(1)} = \sum_{d=1}^{D} w_{md}^{(1)} x_d + w_{m0}^{(1)} \tag{4.5}$$

The outputs z_m ($m \in \{1, ..., M\}$) of the hidden units are obtained by transforming the intermediate activation variables $a_m^{(1)}$ through a non-linear activation function. This function is generally chosen to be sigmoidal, such as the logistic sigmoid function defined in Equation 4.6. Alternatively, the ‘tanh’ function shown in Equation (4.7) can be used (as it is bears a linear relation to the logistic sigmoid, so that a linear combination of ‘tanh’ functions is equivalent to a linear combination of logistic sigmoids).

$$z_m = \frac{1}{1 + \exp(-a_m^{(1)})} \tag{4.6}$$

$$z_m = \tanh(a_m^{(1)}) \tag{4.7}$$

In turn, for each of the C outputs of the network, the second layer of weights defines an activation variable $a_c^{(2)}$ ($c \in \{1, ..., C\}$) as a linear combination of the hidden layer outputs, taking into account the bias parameter $w_{c0}^{(2)}$, as shown in Equation (4.8). (In Figure 4.2 the bias parameters for both layers are associated with links that originate in additional input and hidden variables x_0 and z_0. These variables equate to 1.)

$$a_c^{(2)} = \sum_{m=1}^{M} w_{cm}^{(2)} z_m + w_{c0}^{(2)} \tag{4.8}$$

Finally, the actual outputs y_c ($c \in \{1, ..., C\}$) of the network are obtained by transforming the intermediate variables $a_c^{(2)}$ through the softmax activation function defined in Equation (4.4).

4.3.3 Parameter learning

Given a training set of N input vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and corresponding observed classes $\{t_1, \ldots, t_N\}$, the GLM or MLP model should be trained to determine the parameters that minimise an error function. In the problem considered in this work, with $C > 2$ mutually exclusive classes, this takes the form of the *entropy* error function defined in Equation (4.9), where $t_{bin.c}^n \in \{0, 1\}$ is the c-th element of a binary target vector containing a 1-of-C encoding of the observed class $t_n \in \{T_1, \ldots, T_C\}$ associated with the n-th input vector $\mathbf{x}_n$ (that is, this target vector has C bits; the bit corresponding to the observed class is 1 and all other bits are 0). In addition, y_c^n is the value y_c output by the activation function for that same input.

$$e = -\sum_{n=1}^{N} \sum_{c=1}^{C} t_{bin.c}^n \ln y_c^n \tag{4.9}$$

To avoid situations in which some of the weights in the network would become too large, it is advisable to *regularise* the model by adding a weight decay penalty to the error function, as shown in Equation (4.10), where e_r is the regularised error, e is the non-regularised entropy error previously defined in Equation (4.9), and A is the regularisation constant. The summation in this equation denotes the sum of *all* network weights, regardless of what layer they belong to. The value of A can be determined through cross-validation on training data.

$$e_r = e + A \sum w^2 \tag{4.10}$$

Based on the error function (and its gradient), it is possible to train the GLM via the same general purpose non-linear optimisation algorithms used with other neural networks. However, the linear (or near-linear) structure of the network permits the use of a special purpose algorithm called iterated re-weighted least squares training, with considerable advantage in terms of efficiency. The optimisation algorithm used in this work for training of the MLP was that of scaled conjugate gradients optimisation.

It is worth noting that, in probabilistic terms, minimising the entropy error function in Equation (4.9) is equivalent to maximising the likelihood of the model given the data (in other words, choosing the network weights that are most probable given the data), assuming independent and identically distributed data. In turn, minimising the regularised error in Equation (4.10) is equivalent to setting a multi-variate Gaussian prior on the network weights and maximising their posterior probability given the data. In fact, optimising an error function is not the only way to pose the parameter learning problem. An alternative is to attempt Bayesian inference, as discussed by MacKay [61] and Neal [69].

4.3.4 Classification of spots

On a discriminative approach to the problem of classifying TMA spots into the four main types, the MLP was compared in terms of performance both with the GLM and with a nearest-neighbour classifier. The classifiers were trained to classify spots from their normalised histograms of texton frequencies, computed as described in Section 4.2.2.

4.4 Latent Dirichlet allocation

This section presents a brief overview of latent Dirichlet allocation (LDAL) based on the 2003 paper by Blei et al. [12]. For a comprehensive discussion, that reference should be consulted.

4.4.1 The LDAL model

LDAL is a generative probabilistic model for collections of discrete data. The basic idea behind LDAL is that feature vectors can be represented as random mixtures over latent *topics,* where each topic is characterised by a distribution over *codewords*. Although commonly associated with the modelling of text collections, the LDAL model is not necessarily tied to text and has applications to other types of data.

Considering a feature vector set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, in LDAL terminology the n-th feature vector $\mathbf{x}_n$ in the set is said to contain D_n codewords and can be denoted by $\mathbf{x}_n = (x_1, ..., x_{D_n})^{\mathrm{T}}$. The d-th codeword x_d in a feature vector is the basic unit of discrete data and is defined to be an item from a *vocabulary* of V unique codewords. LDAL assumes that each codeword x_d is associated with a latent topic z_d, defined to be an item from a set of K unique latent topics. The topic mixture for a particular feature vector (that is, the frequencies of occurrence of the K unique topics in the vector) is denoted by θ and is endowed with a Dirichlet distribution. This distribution is assumed because it has properties that facilitate the development of inference and parameter estimation algorithms for LDAL.

Given Dirichlet parameters α (corresponding to K non-negative scalars) and codeword distributions for each unique topic parameterised by $\beta = P(x|z)$ (which could be pictured as a $K \times V$ matrix), LDAL assumes the following generative process for a feature vector $\mathbf{x}$ containing D codewords :

1. Sample a topic mixture θ from the Dirichlet distribution $\mathrm{Dir}(\alpha)$.
2. For each of the D codewords x_d:
   (a) Sample a topic z_d from Multinomial(θ).
   (b) Sample a codeword x_d from $P(x_d|z, \beta)$.

The joint distribution of the topic mixing weights θ, the vector $\mathbf{z}$ containing D topics, and the feature vector $\mathbf{x}$ containing D codewords is given in Equation (4.11).

$$P(\theta, \mathbf{z}, \mathbf{x}|\alpha, \beta) = P(\theta|\alpha) \prod_{d=1}^{D} P(z_d|\theta) P(x_d|z_d, \beta) \tag{4.11}$$

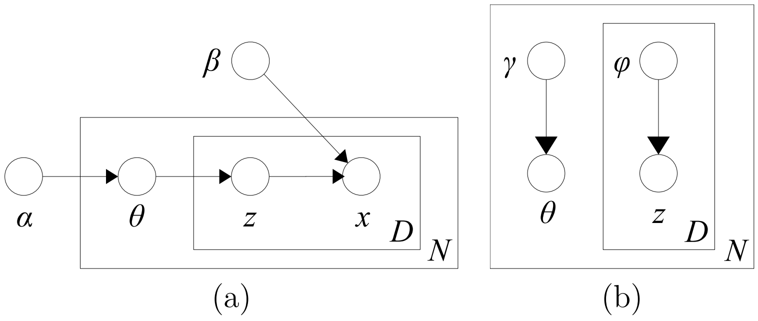


Figure 4.3: (a) Graphical model representation of latent Dirichlet allocation (LDAL). The boxes are *plates* representing replicates: the outer plate represents the feature vectors in a data set, while the inner plate represents the codewords and associated topics within each feature vector. (b) Graphical model representation of the variational distribution used to approximate the posterior distribution in LDAL. (Based on Blei et al. [12].)

It can be seen from Equation (4.11) that LDAL assumes the order of codewords in a feature vector to be irrelevant. This means that, as far as the LDAL model is concerned, a feature vector $\mathbf{x}$ is effectively a *bag* of codewords and could be fully represented by a histogram of codeword frequencies.

Considering a set of N feature vectors, Equation (4.11) corresponds to the probabilistic graphical model shown in Figure 4.3(a). This graphical representation makes clear that the model in question is a three-level hierarchical Bayesian model: α and β represent data set-level parameters, assumed to be sampled only once in the process of generating a set of N feature vectors; θ are feature vector-level variables sampled once per vector; and z and x are codeword-level variables, sampled once for each codeword in each feature vector. (Thus, the previously used notation could be extended, θ_n denoting the topic mixture associated with the n-th feature vector, x_d^n denoting the d-th codeword in the n-th feature vector, and z_d^n denoting the topic associated with that codeword.)

4.4.2 Inference

The main inference problem in LDAL is the computation of the posterior distribution of the hidden variables (latent topics) given a feature vector $\mathbf{x}$, as shown in Equation (4.12).

$$P(\theta, \mathbf{z}|\mathbf{x}, \alpha, \beta) = \frac{P(\theta, \mathbf{z}, \mathbf{x}|\alpha, \beta)}{P(\mathbf{x}|\alpha, \beta)} \tag{4.12}$$

Unfortunately, the quantity $P(\mathbf{x}|\alpha, \beta)$ is intractable to compute in general. A variety of approximate inference algorithms can, however, be considered for LDAL. Blei et al. [12] describe a variational inference algorithm based on a simple modification of the original graphical model, in which some of the edges and nodes are removed, as shown in Figure 4.3(b). It can be demonstrated that the goal of finding a tight lower bound on $\log(P(\mathbf{x}|\alpha,\beta))$ translates directly into the problem of finding the optimising values of the variational parameters (γ,φ). These values are found by minimising the Kullback-Leibler (KL) divergence between the variational distribution $q(\theta,\mathbf{z}|\gamma,\varphi)$ and the true posterior $P(\theta,\mathbf{z}|\mathbf{x},\alpha,\beta)$, which can be achieved via an iterative fixed-point method. It should be noted that the optimising variational parameters (γ,φ) are a function of $\mathbf{x}$, that is, they are feature vector-specific.

The problem of classifying feature vectors into C mutually-exclusive classes can be dealt with by building one LDAL model for each class, based on a training set of feature vectors known to belong to that class. For a given test vector $\mathbf{x}$, variational inference can then be used on each model, to obtain a lower bound on $\log(P(\mathbf{x}|\alpha,\beta))$ for each class $t \in \{T_1,\ldots,T_C\}$. It should be kept in mind that these lower bounds cannot be interpreted as estimates of true class-conditional probabilities, because, in this particular generative approach, latent topics are estimated separately for each model and not shared across classes. Therefore, the lower bounds cannot be used to obtain a probabilistic output in the form of class posterior probabilities, via Bayes' rule. Nevertheless, a classification decision can be made, simply by choosing the class associated with the highest lower bound.

4.4.3 Parameter estimation

Given a training set of N feature vectors $\{\mathbf{x}_1,...,\mathbf{x}_N\}$, an LDAL model can be estimated by finding the parameters α and β that maximise the (marginal) log likelihood shown in Equation (4.13).

$$l(\alpha,\beta)=\sum_{n=1}^{N}\log P(\mathbf{x}_n|\alpha,\beta) \tag{4.13}$$

Approximate empirical Bayes estimates for the LDAL model can be found via the alternating variational expectation maximisation (EM) procedure proposed by Blei et al. [12]. This procedure corresponds to the following iterative algorithm, where (γ_n,φ_n) denotes the variational parameters associated with the n-th feature vector:

1. (E-step) For each feature vector$\mathbf{x}_n$ ($n \in \{1, ..., N\}$), find the values of the variational parameters (γ_n, φ_n) that maximise the lower bound on the log likelihood, via the inference algorithm outlined in Section 4.4.2.
2. (M-step) For fixed the values of the variational parameters, find the model's parameters α and β that maximise the lower bound on the log likelihood. This corresponds to estimating the maximum likelihood of each document under the posterior obtained in the E-step.
3. Repeat these steps until convergence of the lower bound on the log likelihood.

4.4.4 Classification of spots

Classification using the MLP was compared with a generative approach based on LDAL models. The classifiers were trained to classify spots from their normalised histograms of texton frequencies, computed as described in Section 4.2.2.

In LDAL terms, the sample textons computed for a given spot can be seen as codewords, whereas the dictionary of unique textons corresponds to a vocabulary of unique codewords. As explained in Section 4.4.1, LDAL assumes the order of codewords in a feature vector $\mathbf{x}$ to be irrelevant, so that feature vectors are effectively bags of codewords. The histograms of texton frequencies computed for all spots were therefore suitable for use in LDAL modelling, as they constituted sufficient representations of bags of sample textons.

The generative approach to the classification problem was that described in Section 4.4.2, in this instance involving four LDAL models, one for each spot type. The parameters α of each model were initialised as 0.5, while the parameters β were initialised randomly. All parameters (α and β) were then estimated via the variational expectation maximisation procedure referred to in Section 4.4.3.

4.5 Experiments and results

4.5.1 Data and code

The data used in classification experiments consisted of colour images of 364 breast TMA spots subjected to progesterone receptor immunohistochemistry. The type assigned by a pathologist to each spot was known (classification data was available for two separate manual assessment sessions, but only the first session was considered here). The spots belonged to the tumour (T), normal (N), stroma (S), and fat (F) types, and included approximately the same number of spots of each type. Figure 4.1 (on page 50) showed examples of spots belonging to each type.

The code for extraction of local and global features, as well as the code used to control the classification experiments (that is, to prepare the data, call the algorithms, and process their results), was implemented in Matlab. Local feature extraction made use of the Sussex convolution function [103]. The Netlab [67] implementations of k-means, k-nearest-neighbour (with Euclidean distance metric), the GLM, and the MLP were used. The C implementation of LDAL made available by David M. Blei [11, 12] was used as well.

4.5.2 Comparison of MLP with simpler classifiers

A first experiment involved 344 spots (86 of each type) and only the d_1 and d_3 invariants, so that each jet contained nine local features: the r, g, and b colour values and the two invariants for each of the three considered scales. The data set was randomly divided into two halves of 172 spots each, suitable for running leave-half-out experiments.

For each half of the data set, a texton dictionary of 160 centres was computed, based on a random sample of 610,000 normalised jets over all the involved spots (equivalent to 0.06% of all jets). These values were chosen to reach a reasonable compromise between computational memory limitations, the size of the dictionary, and the number of samples used to obtain it. Each dictionary was used along with a nearest-neighbour classifier to obtain a texton histogram for all 344 spots, based on a random sample of at most 610,000 normalised jets per spot (specifically, 3.95% of each spot's jets).

For each of the two passes of a leave-half-out experiment, 10 runs of training and testing with the MLP were executed, to filter the effect of the network's random initialisation. The number of hidden units in the MLP was fixed at three, and different values for the regularisation constant A were tested. Table 4.1(a) shows the resulting average classification accuracy and associated standard deviation over the 20 runs, for four different choices of A. Table 4.1(b) shows the confusion matrix for the best case ($A = 0.1$). For comparison, Table 4.1(c) shows the average classification rates obtained using a nearest-neighbour classifier (with Euclidean and χ^2 distance metrics) and a GLM trained through scaled conjugate

Table 4.1: (a) Classification accuracies obtained using multi-layer perceptrons (MLPs) with different choices of A. (b) Confusion matrix for best results with MLPs ($A = 0.1$). (c) Correct classification rates using nearest-neighbour classifiers and generalised linear models.

(a)

A	Min. (%)	Avg. (%)	Max. (%)
0.00	68.6	71.4±1.8	75.0
0.01	70.9	74.3±2.0	77.3
0.10	73.3	**74.6±0.9**	75.6
0.20	69.2	69.9±0.3	70.3

(b)

Truth	Predicted			
(%)	T	N	S	F
T	74.3	14.1	11.6	0.0
N	20.9	55.8	23.0	0.2
S	7.3	11.2	73.0	8.5
F	0.0	0.2	4.4	95.3

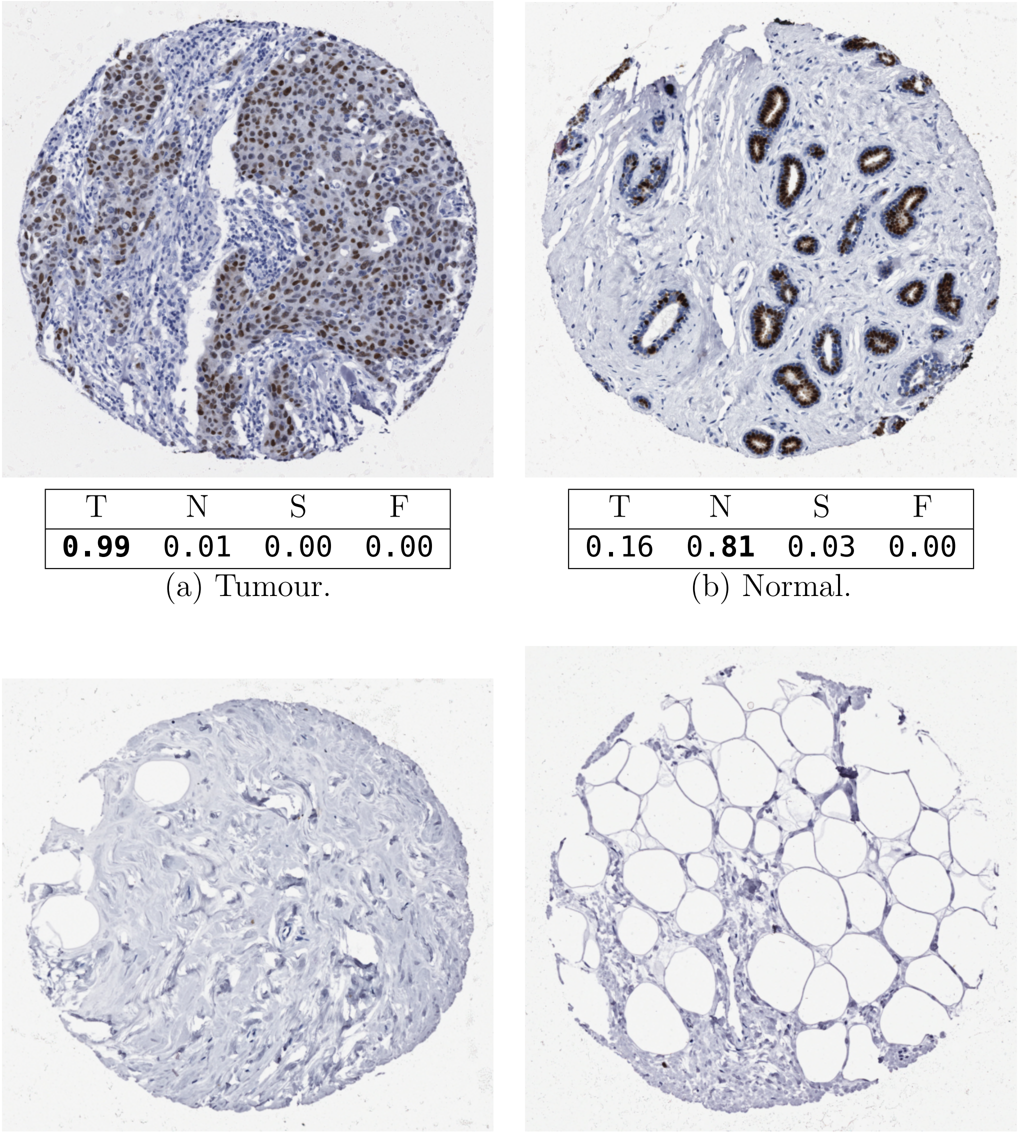
(c)

Method	Avg. (%)
Nearest-neighbour (Euclidean)	59.6
Nearest-neighbour (χ^2)	58.4
GLM	65.4±1.5
MLP	74.6±0.9

gradients optimisation with $A = 0.1$.

Figure 4.4 shows typical examples of correctly classified spots. Below each image, the values of the softmax activation function (that is, the output class posterior probabilities) are presented. In turn, Figure 4.5 shows four examples of misclassified spots.

The entropy of the posteriors distribution for each spot can be used as a simple measure of classification confidence (the lower the entropy, the higher the confidence). Figure 4.6 shows the fractions of test spots that can be classified below different entropy thresholds, averaged over the experiment's 20 runs. Also shown are the mean classification accuracy and the mean rate of misclassified tumour spots. It can be seen, for example, that an entropy threshold of 0.69 allowed to classify about one third of the spots with an accuracy of 95%.



T	N	S	F
0.99	0.01	0.00	0.00

(a) Tumour.

T	N	S	F
0.16	**0.81**	0.03	0.00

(b) Normal.

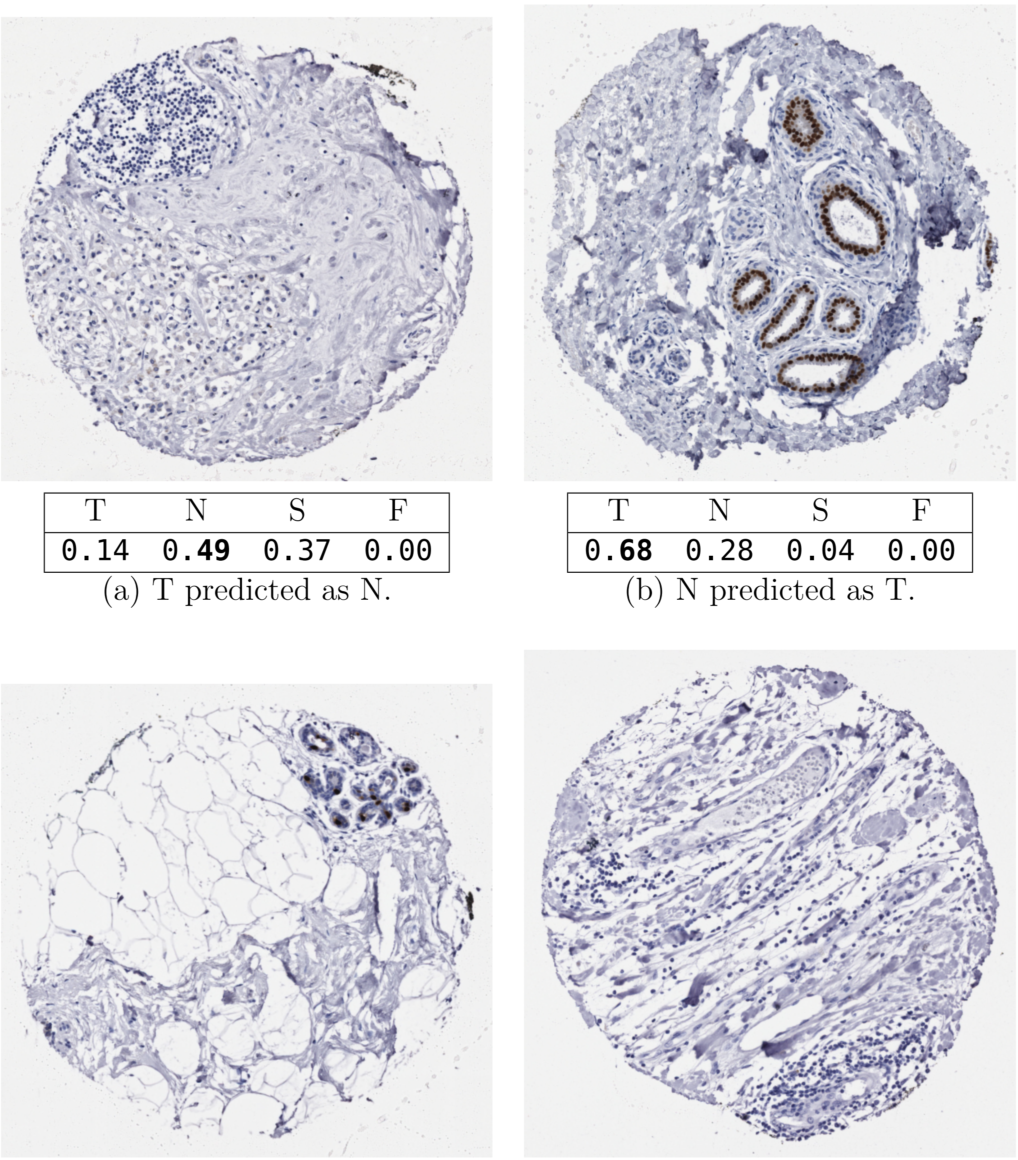
T	N	S	F
0.01	0.10	**0.89**	0.00

(c) Stroma.

T	N	S	F
0.01	0.02	0.13	**0.84**

(d) Fat.

Figure 4.4: Examples of correctly classified spots and corresponding softmax values for the four classes.



T	N	S	F
0.14	**0.49**	0.37	0.00

(a) T predicted as N.

T	N	S	F
0.68	0.28	0.04	0.00

(b) N predicted as T.

T	N	S	F
0.05	0.11	**0.45**	**0.40**

(c) N predicted as S.

T	N	S	F
0.34	**0.49**	0.17	0.00

(d) S predicted as N.

Figure 4.5: Examples of misclassified spots and corresponding softmax values for the four classes.

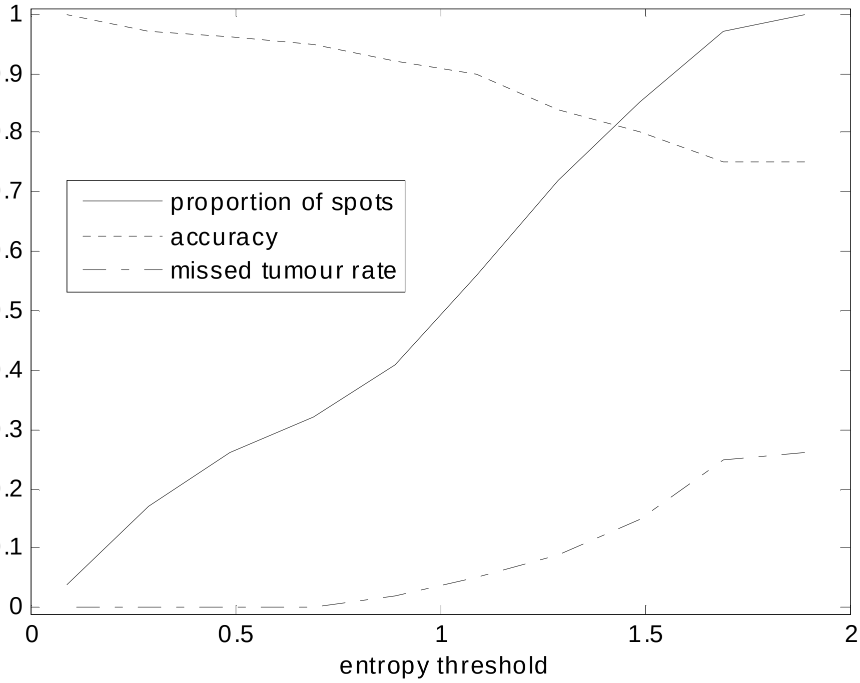


Figure 4.6: Fraction of classified spots, correct classification rate, and rate of missed tumour spots, for different entropy thresholds. (Lower entropy means higher confidence.)

4.5.3 Comparison of MLP with LDAL

A second experiment involved all the available 364 spots and all four differential invariants, so that each jet contained 15 local features. A training set of 40 randomly selected spots (15 tumour, 15 normal, 5 stroma, and 5 fat) was used exclusively to obtain the texton dictionary with 160 centres, based on a random sample of 400,000 normalised jets over the 40 spots. The dictionary was then used along with a nearest-neighbour classifier to obtain a texton histogram for each of the remaining 324 spots, based on a random sample of up to 400,000 jets per spot.

A leave-one-tenth-out experiment was carried out over the subset of 324 spots, using the classification approach based on LDAL models as outlined in Section 4.4.4. The number of latent topics was fixed at 60 and the α parameters of the model were initialised at 0.5. For each of the 10 passes of the experiment, nine runs of training and testing were executed, to filter the effect of randomly initialising the LDAL models. In addition, the whole experiment was repeated while varying the amount of data effectively used to train the models, between 10% and 100% of the 290 spots available for training at each pass. Figure 4.7 shows the classification accuracy results averaged over the nine runs of each

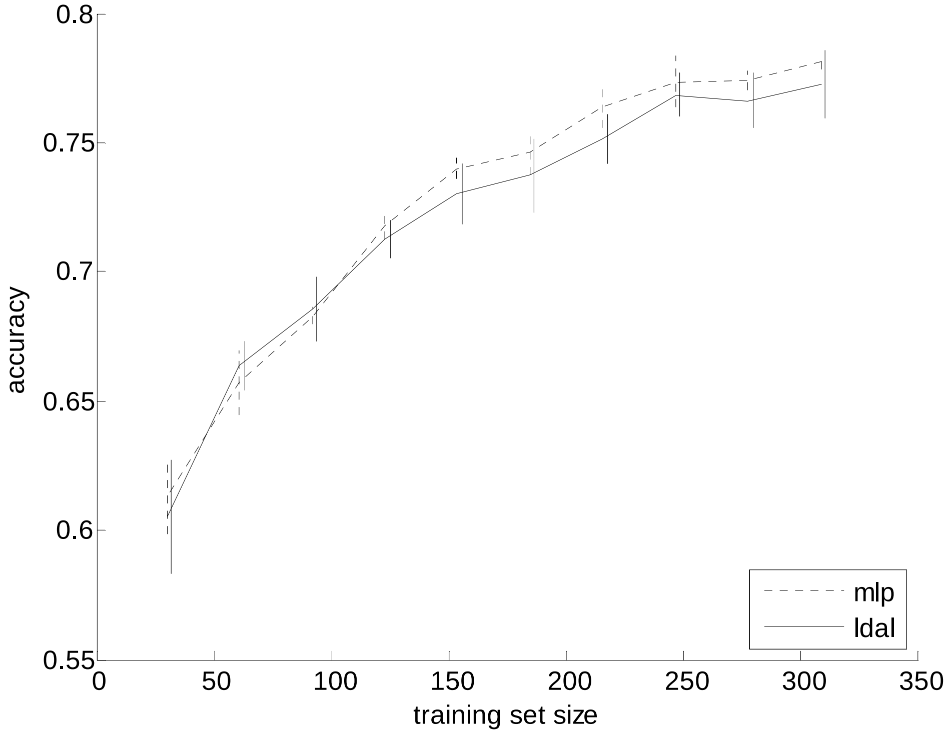


Figure 4.7: Accuracy of classification using either multi-layer perceptrons or latent Dirichlet allocation models, for different training set sizes (in number of spots).

experiment (the error bars corresponding to ± one standard deviation). Also shown for comparison are the results obtained employing not LDAL models but an MLP with three hidden units and regularisation constant $A = 0.1$.

As mentioned in Section 2.5, a trial was conducted in which a set of 935 spots were classified by the same pathologist on two separate sessions (this larger set included all 324 spots used in the previously described classification experiments). This resulted in an intra-observer agreement of 94.0%, corresponding to an unweighted Cohen's kappa coefficient of 0.885. Table 4.2 shows the contingency table for the two sessions. It can be seen that the proportions of spots of each type were not uniform, tumour spots representing approximately two thirds of the whole.

Table 4.2: Contingency table of the intra-observer classification trial.

Session 1	Session 2				
	T	N	S	F	
T	600	8	14	2	624
N	11	67	4	1	83
S	5	0	122	6	133
F	1	1	3	90	95
	617	76	143	99	

4.6 Discussion and conclusions

4.6.1 Comparison of MLP with simpler classifiers

The accuracy of 74.6±0.9% achieved in the first classification experiment is far from the intra-observer agreement of 94.0%. It should be kept in mind, however, that the *inter*-observer agreement would presumably be lower and therefore constitute a more realistic criterion for comparison.

As Figure 4.7 shows, higher accuracies are attainable for fractions of the data, by setting classification confidence thresholds. Very low or zero misclassified tumour rates can also be achieved, as is desirable in this application. This suggests that the system could be used to automatically classify the more unequivocal spots, while pointing out to the pathologist the more difficult spots in need of manual assessment.

Some instances of misclassification are discussed in the following paragraphs.

The epithelial cells in the bottom-left region of spot 4.5(a) are unusually far apart, which may explain the low posterior probability of true class T and suggest that a model solely based on local features is incapable of capturing some of the relevant information contained in the spot.

Spot 4.5(b), when compared with the correctly classified spot 4.4(b), shows similar ring-like arrangements of epithelial cells, but a substantially larger quantity of scattered epithelials, which seems to have caused the posterior probability of class T to be much higher than that of true class N.

In spot 4.5(c), large regions of stroma and fat boosted the posterior probabilities of classes S and F, when what truly counted for the pathologist was the small portion of normal tissue in the top-right region of the spot. This indicates a difficulty in dealing with heterogeneous spots that contain large proportions of different types of tissue. In theory, if enough relevant examples are made available for training, the system should be able to learn to ignore large regions of stroma or fat when small portions of tumour or normal tissue are clearly present in the same spot.

The scattered but non-epithelial (inflammatory) cells in spot 4.5(d) seem to have been misperceived as epithelials, leading to a very low posterior for the true class S. This may indicate that the local filters used do not provide enough detail.

4.6.2 Comparison of MLP with LDAL

In Figure 4.7, it can be seen that the error bars associated with the average classification accuracies achieved via MLPs and LDAL models overlap for all training set sizes. Therefore, it can only be concluded that these two types of classifier performed similarly. This is a promising result as regards the LDAL approach, since it was achieved via the relatively simple use of separate LDAL models that did not share topics. In addition, LDAL models are generative, therefore potentially more interesting than neural networks.

***

This chapter addressed the classification of breast TMA spots subjected to progesterone receptor immunohistochemistry into four main types, namely tumour, normal, stroma, and fat. The techniques used to extract local and global features from TMA image data were described, and an overview of the principles of neural network-based classification, as well as LDAL, was given. A description of the carried out experiments was given and their results reported. Finally, these results were discussed and some conclusions drawn.

The next chapter will focus on the segmentation of breast TMA spots into regions of four types, namely tumour, normal tissue, stroma, and fat. In addition, the trial of a commercial tool used for segmentation will be reported.

Chapter 5

Segmentation of tissue regions

5.1 Introduction

This chapter addresses the segmentation of breast tissue microarray (TMA) spots subjected to progesterone receptor immunohistochemistry into regions of four types, namely tumour, normal tissue, stroma, and fat. It should be noted that, even though *spot types* (dealt with in Chapter 4) and *region types* share the same number and names, they correspond to different concepts. In fact, spots belonging to the tumour and normal spot types can (within certain limits) contain regions of tissue belonging to all four region types.

The segmentation experiments reported in this chapter did not actually play a role towards the main objectives of this work, namely the classification of TMA spots into types and their scoring. However, it is fair to assume that the successful segmentation of spots into regions of different types could help to improve the method used to classify spots, reported in Chapter 4. For example, the classifiers involved could be trained to use as input not only a histogram of texton frequencies for each spot, but also a profile reflecting the proportions of the spot's area covered by different types of tissue. Moreover, a segmentation procedure capable of identifying not only regions of tumour, normal tissue, stroma, and fat but also immunonegative and immunopositive areas within the tumour and normal regions could, in principle, serve as a basis for both classification of spots and their scoring.

There was, therefore, motivation to carry out some preliminary experiments involving at least the segmentation of spots into the four main types of tissue present in them (without distinguishing, at this stage, between immunonegative and immunopositive areas). The main idea behind these experiments was to take advantage of the generative nature of the latent Dirichlet allocation (LDAL) model and the fact that regions of different types represent an intermediate level between the textons into which a spot is quantised and the spot's type. Region types might therefore be successfully modelled as LDAL latent topics.

As mentioned in Section 3.4, Aperio Technologies, Inc. commercialise the Genie tool, which is based on genetic algorithms and can be trained from manual annotations to segment images of tissue sections into regions of different types [3, 71]. The author had the opportunity to test the Genie tool, over a limited period of time, in the segmentation of breast TMA spots. Some qualitative results from that trial are reported in this chapter as well.

The remainder of this chapter is organised as follows. Section specifies how local and global features were computed from TMA image data. Section 5.3 describes the segmentation methods that were tested, whereas Section 5.4 provides details on the experiments carried out and their results. Section 5.5 discusses the obtained results and presents some conclusions. Finally, Section 5.6 focuses on the trial of the Aperio Genie tool, presenting and discussing some of its results.

5.2 Feature extraction

A jet of local features was obtained for each pixel in each image, using the method described in Section 4.2.1. These jets contained 15 features, namely the r, g, and b colour values, and four grey-level invariants computed for three different scales.

The method described in Section 4.2.2 was used to obtain a texton dictionary, from a random sample of local feature jets over the whole data set. This dictionary was then used along with a nearest-neighbour classifier to determine the textons associated with a denser random sample of jets within each individual spot. This allowed to obtain for every spot a feature vector (or bag of sample textons) $\mathbf{x} = (x_1, ..., x_D)^\mathrm{T}$ suitable to be used in LDAL-based experiments. However, given that the ultimate goal of these experiments was that of segmenting spots into regions, the image positions associated with each sample texton in each image were stored as well.

5.3 Segmentation of spots

Three different segmentation methods were tested. These methods are referred to as LDAL, LDAL (fixed β), and texton frequencies or TF, and are described in the following sections.

5.3.1 LDAL

The main idea behind segmentation based on LDAL was that of associating region types with latent topics. A training set of spot images was manually annotated to identify regions of different types. The previously stored information on sample texton positions within those images was then used together with the annotations to obtain the joint distribution of region types and textons, $P(z, x)$, simply by determining the frequencies of textons belonging to each region type.

A distribution over textons given each region type was obtained by normalising the joint distribution: $P(x|z) = P(z, x)/P(z)$. This conditional distribution was used to initialise the parameters β of an LDAL model. The variational expectation maximisation (EM) procedure outlined in Section 4.4.3 was then used to obtain the final values of both the parameters α and the parameters β of the model. This parameter estimation step was based on the same training spots that yielded the joint distribution $P(z, x)$.

Finally, the variational inference algorithm described in Section 4.4.2 was applied to each test spot $\mathbf{x}$, to obtain the optimised parameters (γ, φ) of the model. It should be reminded that these parameters are a function of $\mathbf{x}$, that is, they are spot-specific. Of particular importance in this application were the parameters $\varphi(\mathbf{x})$, as they corresponded to the posterior distribution over region types for each sample texton in a test spot. For the purpose of segmentation, each texton could then be assigned to the region type with the highest posterior.

5.3.2 LDAL (fixed β)

This method constituted a variation of the procedure described in Section 5.3.1. As before, the estimated distribution $P(x|z)$ was used to initialise the parameters β of an LDAL model. However, in this instance the variational EM procedure was employed only to estimate the final values of the parameters α of the model, while keeping the parameters β fixed at their initial values.

5.3.3 TF

This method, though making no use of LDAL, provided a useful baseline for the comparison of results. A posterior distribution over region types given each texton in the dictionary was obtained by normalising the previously obtained joint distribution across region types: $P(z|x) = P(z, x)/P(x)$. For each sample texton in a test image, a posterior distribution could therefore be looked up and the texton assigned to the region type with the highest posterior.

5.4 Experiments and results

5.4.1 Data and code

The data used in the segmentation experiments consisted of colour images of 40 breast TMA spots (15 tumour, 15 normal, 5 stroma, and 5 fat) subjected to progesterone receptor immunohistochemistry. These spots were manually annotated to identify regions belonging to four different types, namely tumour, normal tissue, stroma, and fat or glass. So-called glass regions are empty regions without any tissue that often appear in TMA spots. The annotation of these regions was combined with that of regions of fat, because the two types differ little in appearance. In fact, regions of fat in TMA spots do not contain any fatty tissue, but rather empty space that used to be occupied by fat, meanwhile dissolved as a result of the preparation process. The used annotations were created by the author and were not confirmed by a pathologist, therefore they may not be entirely reliable. Figures 5.1(b), 5.2(b), and 5.3(b) show manual annotations for three example spots.

The code for extraction of local and global features, as well as the code used to control the segmentation experiments (that is, to prepare the data, call the algorithms, and process their results), was implemented in Matlab. The C code for LDAL made available by David M. Blei [11, 12] was modified to allow the implementation of the desired LDAL models.

5.4.2 Experiments

The 40 manually annotated spots were used to compute a texton dictionary of 160 textons, based on a random sample of 400,000 local jets over the 40 spots. This

Table 5.1: Agreement between manual annotations and the three tested segmentation methods, based on labelled sample points, along with precision (p) and recall (r) for tumour and normal regions.

	LDAL	LDAL (fixed β)	TF
Agreement	0.695	0.658	0.464
p_{tumour}	0.681	0.777	0.306
r_{tumour}	0.678	0.497	0.173
p_{normal}	0.544	0.821	0.462
r_{normal}	0.793	0.434	0.027

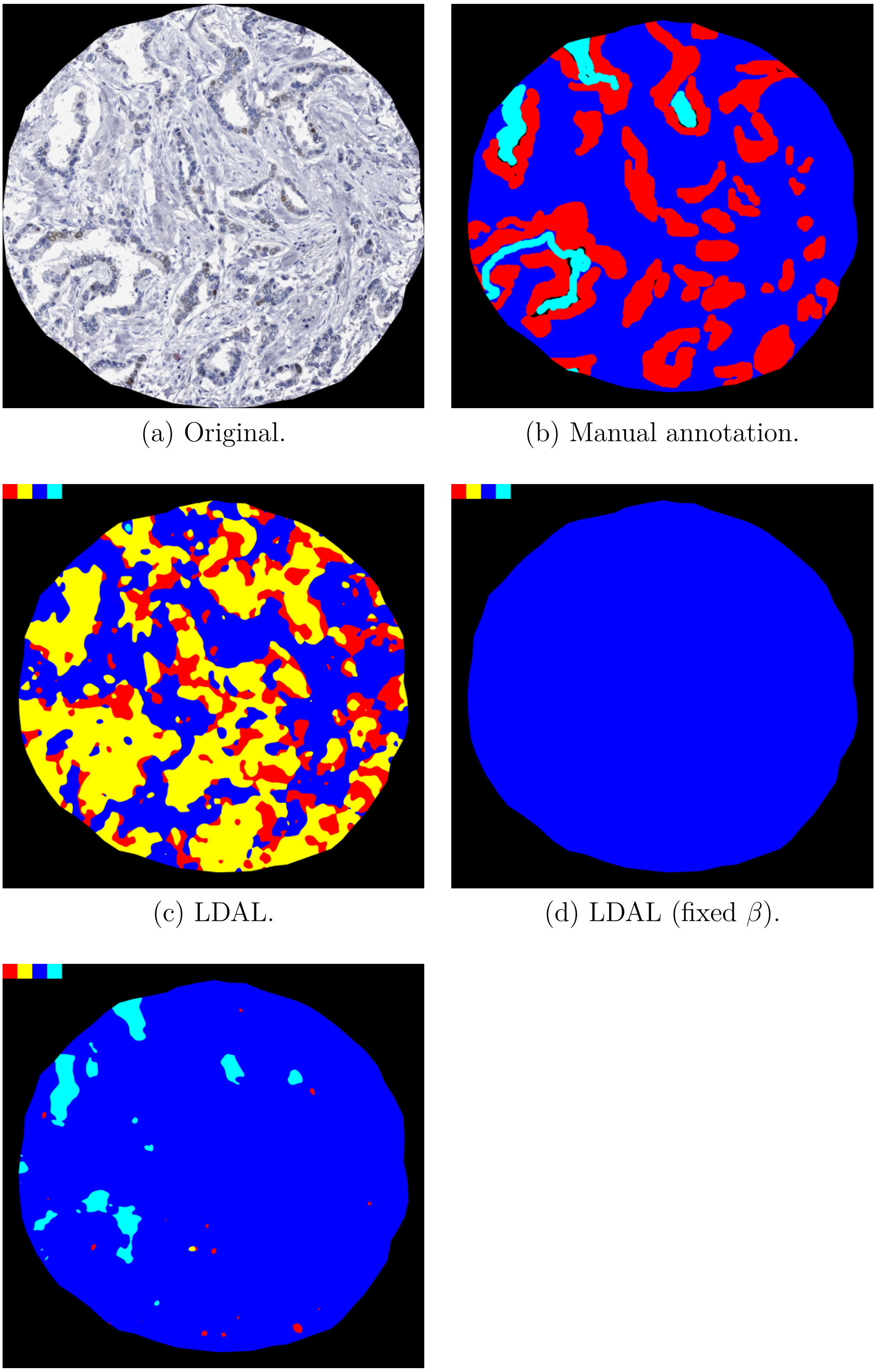
dictionary was then used along with a nearest-neighbour classifier to compute a histogram of texton frequencies for each spot, based on a random sample of up to 400,000 jets from each individual spot.

For each of the segmentation methods described in Section 5.3, a leave-one-out experiment was carried out over the 40 spots. At each pass of each experiment, the annotations of the 39 spots selected for training were used to obtain the joint distribution of region types and textons, $P(z, x)$.

In the LDAL and LDAL (fixed β) methods, the α parameters of the model were initialised as 0.5 and the joint distribution $P(z, x)$ provided the initial values for β. The 39 training spots were then used in variational EM parameter estimation. Finally, variational inference on the remaining test spot $\mathbf{x}$ yielded a posterior distribution over region types given each texton, in the form of the optimised variational parameters $\varphi(\mathbf{x})$. In the TF method, a posterior over region types given each texton, $P(z|x)$, was computed directly from the joint distribution.

In order to approximate a full segmentation of each test spot (as opposed to the mere labelling of its randomly sampled textons), a set of four "posterior" images was built, one for each region type. Each sample texton contributed to each "posterior" image with a two-dimensional Gaussian distribution (with standard deviation $\sigma = 16$ pixels) centred at the location of the sample and weighted by the texton's posterior probability for the region type in question. The full segmentation was then obtained by assigning each pixel in the test image to the region type with the highest "posterior" value. Figures 5.1(c), (d), and (e), 5.2(c), (d), and (e), and 5.3(c), (d), and (e) show the segmentation results obtained for three example spots via each of the three employed methods.

Table 5.1 presents the proportions of sample points whose labelling was in agreement with the manual annotation, for the three segmentation methods. Also shown are the precision and recall computed for the tumour and normal regions.



(a) Original.

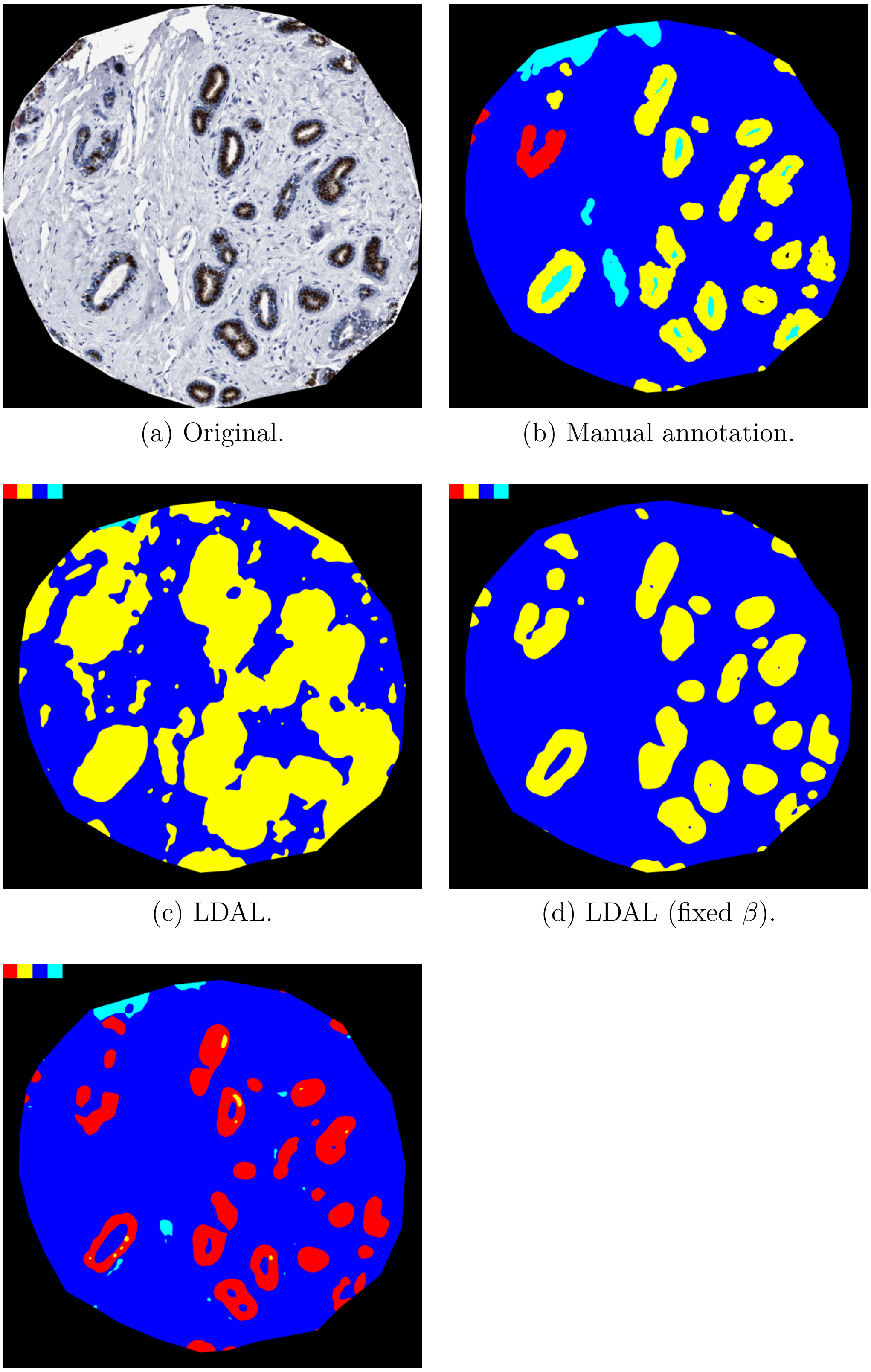
(b) Manual annotation.

(c) LDAL.

(d) LDAL (fixed β).

(e) TF.

Figure 5.1: (a) Example spot, (b) its manually annotated region, and (c,d,e) segmentation results from the three employed methods. (Red: tumour; yellow: normal tissue; blue: stroma; and cyan: fat.)



(a) Original.

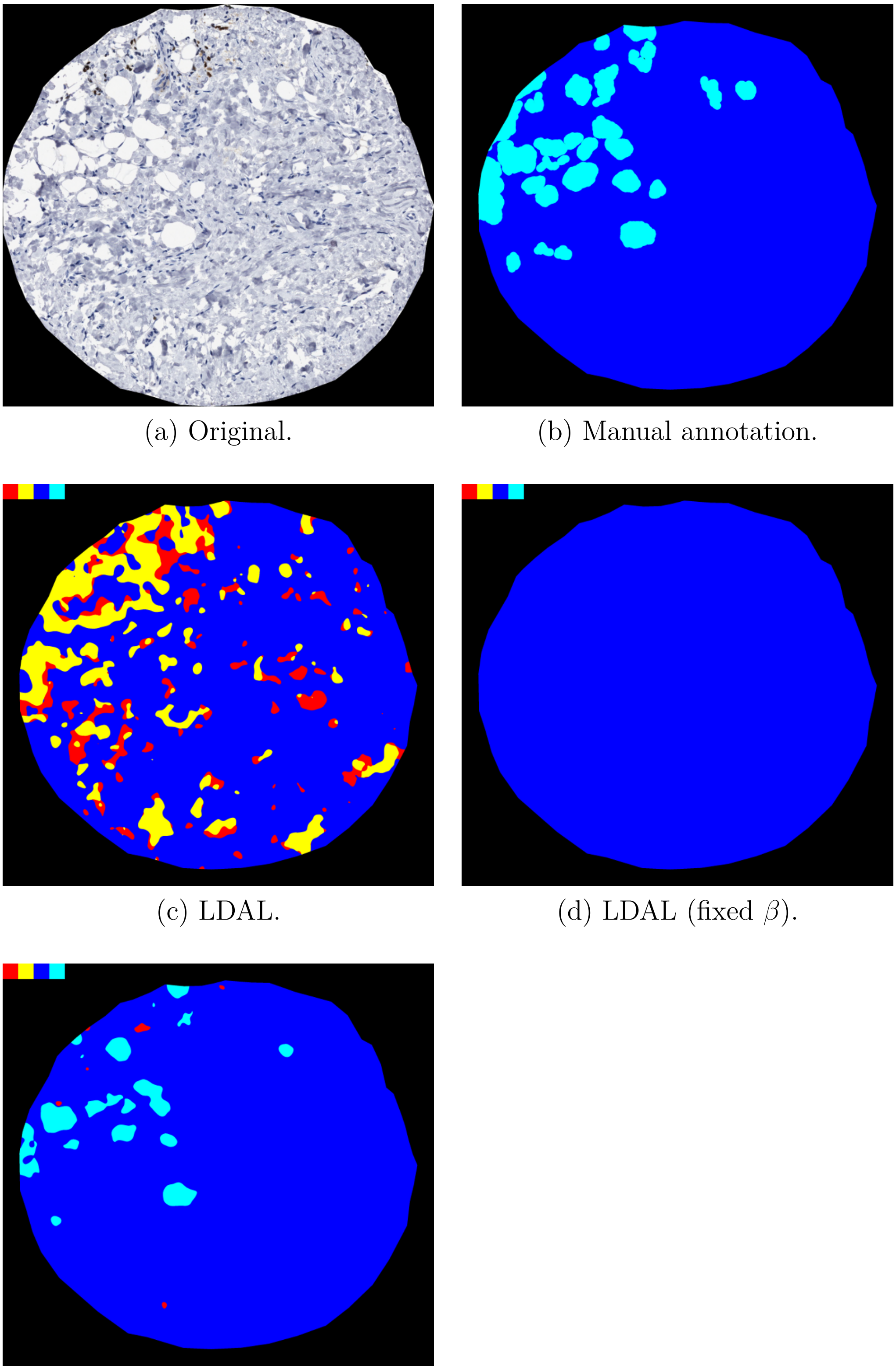
(b) Manual annotation.

(c) LDAL.

(d) LDAL (fixed β).

(e) TF.

Figure 5.2: (a) Example spot, (b) its manually annotated region, and (c,d,e) segmentation results from the three employed methods. (Red: tumour; yellow: normal tissue; blue: stroma; and cyan: fat.)



(a) Original.

(b) Manual annotation.

(c) LDAL.

(d) LDAL (fixed β).

(e) TF.

Figure 5.3: (a) Example spot, (b) its manually annotated region, and (c,d,e) segmentation results from the three employed methods. (Red: tumour; yellow: normal tissue; blue: stroma; and cyan: fat.)

Given that, for global feature extraction, local feature jets were quantised into textons, it may be of interest to note that segmentation based on LDAL can be seen as a further level of quantisation, of textons into region types. Figure 5.4 shows an example TMA spot of the normal type, along with the corresponding map of textons quantising sample jets, the segmentation result based on LDAL, and the manual annotation of the spot, respectively.

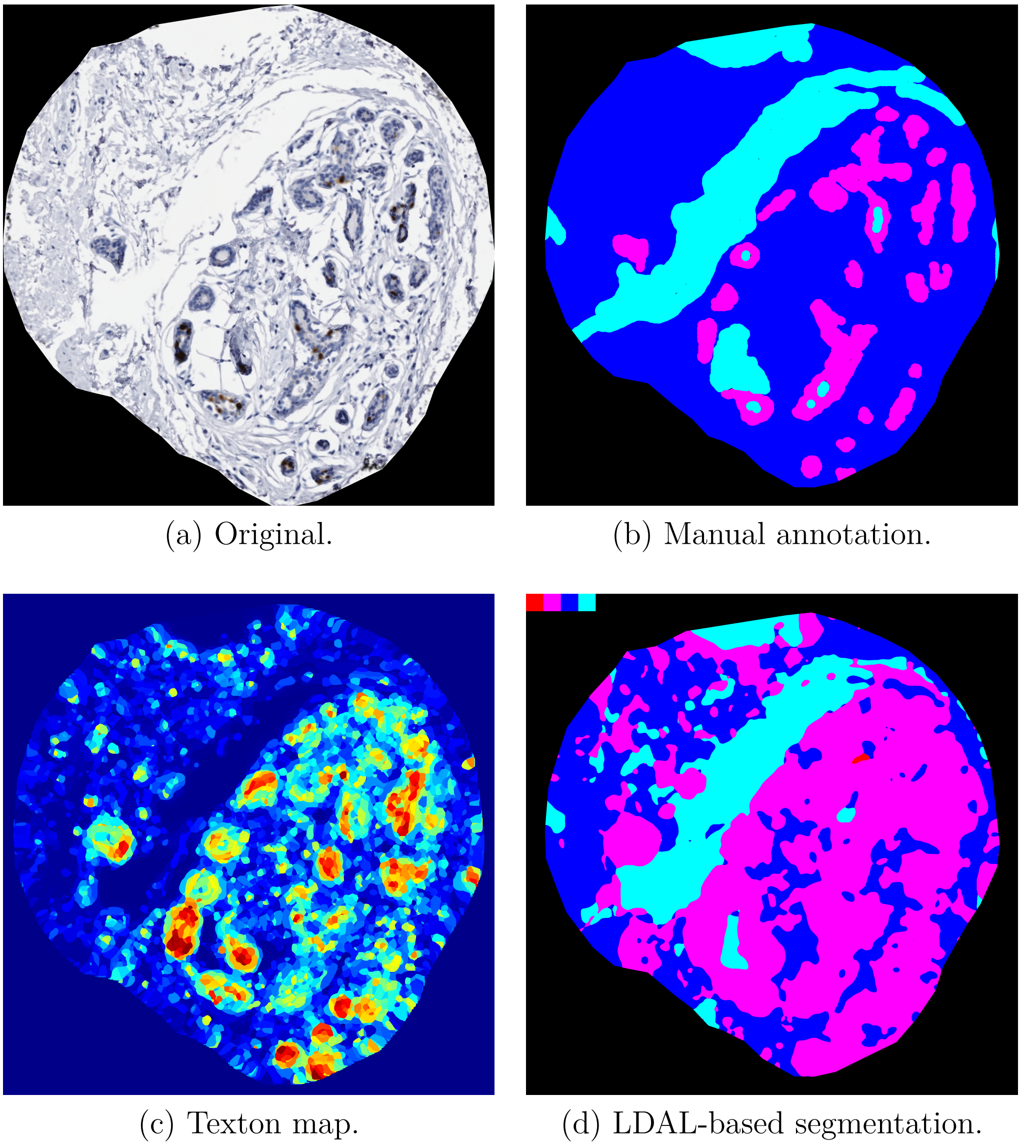
5.5 Discussion and conclusions

As Table 5.1 shows, in relation to the TF segmentation method, the LDAL (fixed β) method resulted in a much better agreement between the manually annotated regions and the segmentation. In addition, the precision and recall for tumour and normal regions achieved much better values. The use of LDAL with inferred β yielded a further improvement in the agreement, but the main difference observed in relation to the LDAL (fixed β) approach was the increase in the recall for tumour and normal areas, although accompanied by lower precision. High recalls for these regions are desirable, as ideally the pathologists would like to detect any existing tumour and normal tissue, even if at the expense of lower precision (which in principle should mainly correspond to dilation of the segmented regions).

Figure 5.1 shows segmentation results for a spot of the tumour type and very low staining strength. As with other spots that shared these characteristics, LDAL was the only method to detect the regions of epithelial tissue, although falsely labelling much of the tumour regions as normal.

Figure 5.2 shows an example normal spot for which the LDAL (fixed β) method yielded the best segmentation results. It can be seen that the TF method detected all of the normal tissue as tumour, whereas LDAL successfully detected these normal regions but exaggeratedly dilated them. An interesting aspect of this example is that the manual annotation contained a mistake, in that the small V-shaped region annotated as tumour on the top-left area of the spot was in fact composed of normal epithelial cells. Nevertheless, both LDAL-based methods segmented this region correctly, as normal tissue.

Figure 5.3 shows an example where the TF method can be said to have yielded the best results. Similar results were obtained for other spots containing dominant regions of stroma: in contrast to the TF method, LDAL (fixed β) tended to interpret the whole of the spot as stroma, whereas LDAL tended to segment



(a) Original.

(b) Manual annotation.

(c) Texton map.

(d) LDAL-based segmentation.

Figure 5.4: (a) A tissue microarray spot of the normal type, (b) its manual annotation, (c) the corresponding texton map, and (d) the result from segmentation based on latent Dirichlet allocation. (Textons are represented by 160 different colours. The segmentation result and the annotation use magenta for normal tissue, blue for stroma, and cyan for fat or glass.)

non-existent portions of normal and tumour tissue scattered amidst the stroma.

The fact that LDAL yielded the best agreement with annotated data and the highest recall for tumour and normal regions is encouraging, but the quality of the results is still far from satisfactory. In particular, even though LDAL was often the only method to correctly detect certain regions of tumour, this was usually accompanied by exaggerated dilation or false detection of normal regions.

5.6 Trial of Aperio Genie tool

The Genie tool from Aperio Technologies, Inc. allows the user to train a classifier from a collection of small annotated regions, called a *montage*. These small regions should be representative of the various tissue types that the user wishes to segment. The trained classifier can then be used to segment whole tissue sections (such as TMA spots) into regions belonging to the tissue types in question.

The trial of this tool involved 16 spots that were manually annotated, to delineate regions of tumour, normal tissue, stroma, and fat. These annotations were created by the author with the Aperio ImageScope tool. Figures 5.5 and 5.6 show, on the left side, examples of annotated spots. With the ImageScope tool, the annotation of tissue areas is done by means of closed contours.

The Genie tool sets a limit to the total annotated area that may be used for training. So, it was necessary to select small portions from the complete annotations to serve as *training* annotations.

The 16 images were used in a leave-one-out experiment. At each pass, a montage was manually created using the training annotations associated with 15 spots, and a new classifier was trained with that montage. The classifier was then used to segment the remaining test spot.

Genie allows the user to set a number of parameters that affect aspects of the training and classification stages, such as the types of features taken into account and the number of iterations executed. In this experiment, however, all parameters were kept at their default values.

Genie does not report measures of performance of the segmentation procedure (excepting a so-called training accuracy, if the user chooses to test the classifier on the same annotated data it was trained with). The visual results output by the tool show the segmented regions as semi-transparent coloured areas overlaid on the original images. Figures 5.5 and 5.6 show, on the right side, the segmentation results obtained for the spots annotated on the left side.

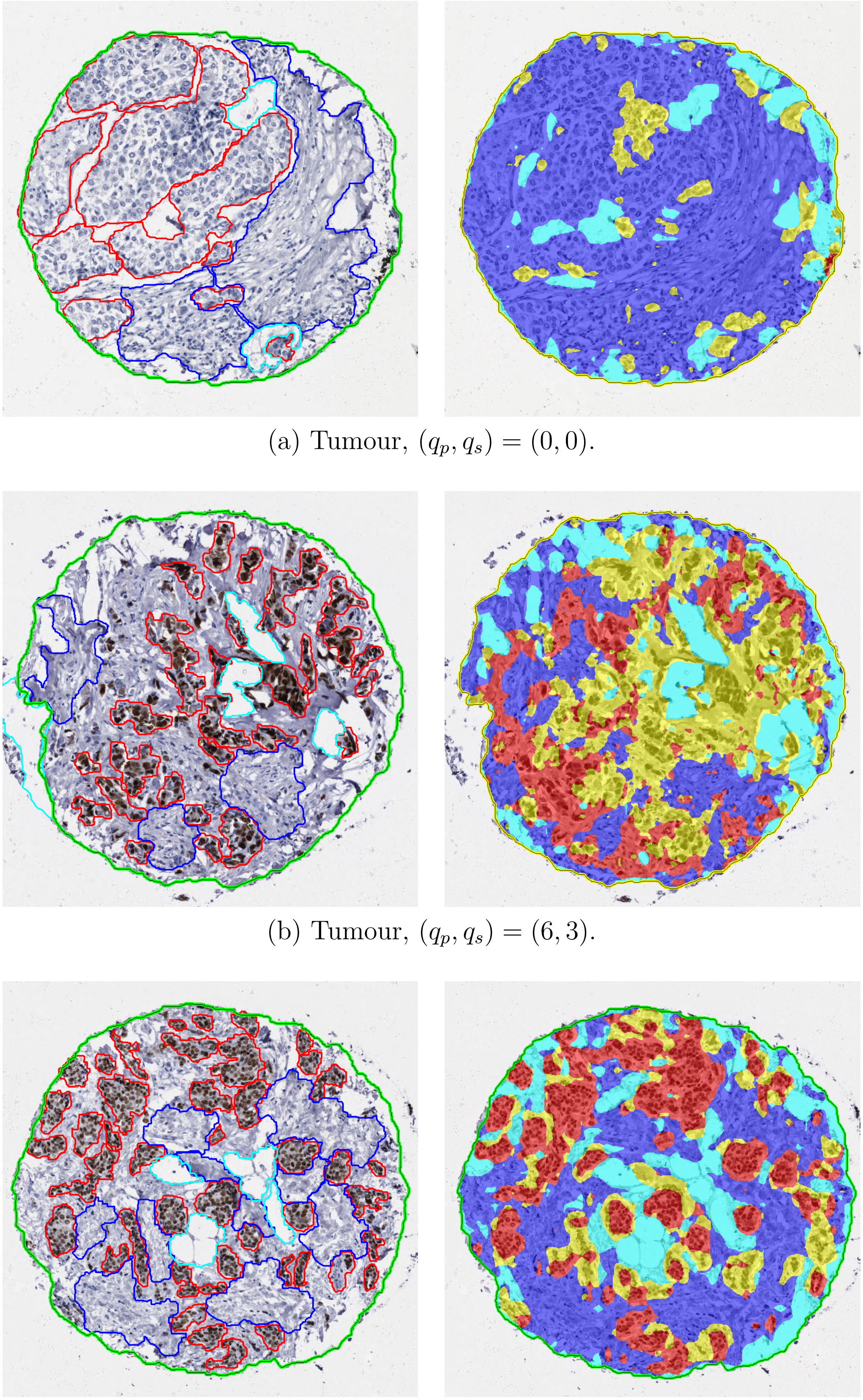
The tool had particular difficulty in segmenting regions of immunonegative (unstained) tumour. This is illustrated with the spot shown in Figure 5.5(a), where a large region of tumour is mostly interpreted as an extension of the existing stroma and partly segmented as normal tissue. In some cases, the segmentation of immunopositive tumour was flawed as well. For example in the spot shown in Figure 5.5(b), the tumour regions were in great part segmented as normal tissue. However, the segmentation of immunopositive tumour was quite successful for some other spots, such as that shown in Figure 5.5(c).

The segmentation performance for immunopositive normal regions was ambiguous. In certain cases, normal tissue was partially segmented as tumour. This is illustrated in Figure 5.6(b), where it can also be seen that large portions of stroma were interpreted as normal tissue or even as tumour. The segmentation of immunopositive normal tissue achieved much better results with some other spots, such as that shown in Figure 5.6(c). Regions of immunonegative normal tissue are those best segmented by Genie. This is exemplified with the spot shown in Figure 5.6(a).

***

This chapter addressed the segmentation of breast TMA spots subjected to progesterone receptor immunohistochemistry into regions of four types, namely tumour, normal tissue, stroma, and fat. The ways in which local and global features were computed from TMA image data were described. Three alternative segmentation approaches were described. A description of the carried out experiments was given and their results reported. These results were discussed and some conclusions presented. The trial of Aperio's Genie tool was also described and its outcome discussed.

The next chapter will focus on the prediction of quickscores of breast TMA spots.

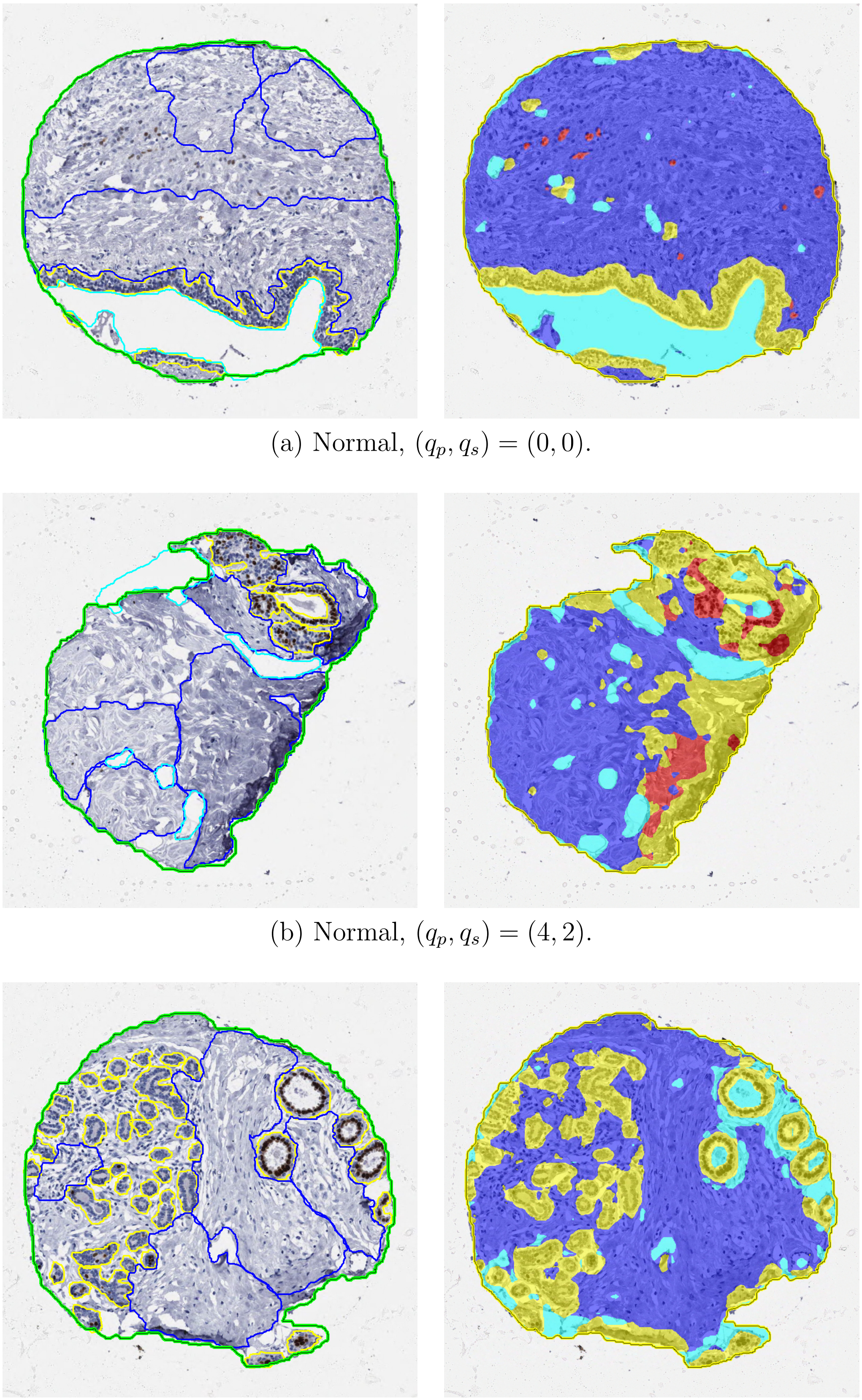


(a) Tumour, $(q_p, q_s) = (0, 0)$.

(b) Tumour, $(q_p, q_s) = (6, 3)$.

(c) Tumour, $(q_p, q_s) = (6, 2)$.

Figure 5.5: (Left) Manually annotated regions on three examples of tumour spots, and (right) corresponding Genie segmentation results. (Red: tumour; yellow: normal tissue; dark blue: stroma; and cyan: fat.)



(a) Normal, $(q_p, q_s) = (0, 0)$.

(b) Normal, $(q_p, q_s) = (4, 2)$.

(c) Normal, $(q_p, q_s) = (2, 3)$.

Figure 5.6: (Left) Manually annotated regions on three examples of normal spots, and (right) corresponding Genie segmentation results. (Red: tumour; yellow: normal tissue; dark blue: stroma; and cyan: fat.)

Chapter 6

Scoring of spots

6.1 Introduction

This chapter addresses the prediction of quickscores of breast tissue microarray (TMA) spots subjected to progesterone receptor immunohistochemistry.

Global features analogous to those used by pathologists (namely an estimate of the proportion of epithelial nuclei that are immunopositive and an estimate of the strength of staining of those nuclei) were computed for each spot, via a method based on local patches. Using those features as input, the prediction performance of ordinal regression based on Gaussian processes was then compared with that of classification with neural networks. Ordinal regression differs from classification in that the existence of an order between different categories is taken into account. So, in the prediction of tumour scores, ordinal regression was expected to achieve better results than classification.

Some variations on the computation of global features were also investigated, by assessing their impact on the prediction performance of ordinal regression. Specifically, the use of local patches was compared with an alternative method involving explicit nuclear segmentation. Moreover, in both cases, computation based on the whole of each spot was compared with computation restricted to a manually drawn region of interest (ROI). The ROIs essentially delineated regions of either tumour or normal tissue within the spots, excluding regions of fat and, more importantly, stroma.

It is worth pointing out that the employed nuclear segmentation technique cannot be considered fully automated, in the sense that it involved heuristics and free parameters that were manually tuned to achieve good performance with the analysed data set. The segmentation method and the manually drawn ROIs served as a means to assess how the system would perform if its ability to identify immunopositive and immunonegative epithelial regions in the tissue were nearly perfect (in other words, the aim was to to determine how far from that ideal performance the truly automated system is, relying on local patches without the aid of ROIs).

The remainder of this chapter is organised as follows. Section 6.2 describes the techniques used to extract both local and global features from TMA spot images. Section 6.3 provides an overview of Gaussian processes and their application to ordinal regression, and explains how this technique was use to predict scores of TMA spots. Section 6.4 provides details on the experiments carried out and their results. Finally, Section 6.5 discusses the results and presents some conclusions.

6.2 Feature extraction

6.2.1 Extraction of local feature jets

A jet of local features was obtained for each pixel in each image, using the method described in Section 4.2.1. These jets contained 15 features, namely the r, g, and b colour values, and four grey-level invariants computed for three different scales, denoted by $d_{k,\sigma}$, $k \in \{1, 2, 3, 4\}$, $\sigma \in \{8, 16, 32\}$.

6.2.2 Extraction of global features

As explained in Section 2.3, the *quickscore* used by pathologists is composed of two integer values, one reflecting the perceived *proportion* of epithelial nuclei that are immunopositive, and the other reflecting the perceived *strength* of staining of those nuclei. In the remainder of this chapter, these two quickscore values are denoted by q_p and q_s, respectively.

For each spot, two real numbers x_p and x_s were computed as global features that formalised the q_p and q_s values used by pathologists. Two distinct methods were employed to obtain these values, as explained in the following sections.

Local patch-based method

Manually annotated subregions of TMA spots were used to estimate class-conditional probability distributions of local features for three classes of pixel, namely background (B), epithelial immunonegative (E_-), and epithelial immunopositive (E_+). Denoting the pixel class as $v \in \{B, E_-, E_+\}$, the estimated distributions can be expressed as $P(r, g, b|v)$ for colour features and $P(d_{k,\sigma}|v)$ for each differential invariant feature $d_{k,\sigma}$. Each of these distributions was estimated as a histogram with a certain number of bins.

Although colour components were considered inter-dependent, the differential invariants were assumed to be independent of one another and independent of colour, given the pixel class. Thus, the class-conditional distributions estimated through training could be used to factor the likelihood of the local features of a pixel given the class, as in $P(r, g, b|v) \prod_{k,\sigma} P(d_{k,\sigma}|v)$.

A prior $P(v)$ over pixel classes was also computed, from the frequencies of pixels belonging to each class, as observed in the annotated training data.

Given a new image (more precisely, the local feature jets for all of its pixels), class posterior probabilities for each pixel in the image could be estimated as in Equation (6.1), where the likelihood term is multiplied by the prior $P(v)$ and $\mathbf{u}$ denotes the pixel's local feature jet $(r, g, b, d_{1.8}, ..., d_{4.32})^{\mathrm{T}}$.

$$P(v|\mathbf{u}) = \frac{P(r, g, b|v) \prod_{k,\sigma} P(d_{k,\sigma}|v) P(v)}{P(\mathbf{u})} \tag{6.1}$$

Each pixel was then labelled as belonging to the class with the highest posterior. The posterior probability values associated with the class E_+, however, were not discarded, since they played a role in the computation of global features, as explained in the following.

The first global feature, x_p, was computed as the number of pixels labelled as E_+, divided by the total number of pixels labelled as epithelial (both E_- and E_+). This is shown in Equation (6.2), where N_{E_-} and N_{E_+} denote the numbers of pixels labelled as E_- and E_+, respectively. This feature formalised the q_p value. The second feature, x_s, was obtained as the mean posterior probability of a pixel belonging to the E_+ epithelial class, computed over all pixels assigned to that class, as shown in Equation (6.3). This feature formalised the q_s value.

$$x_p = \frac{N_{E_+}}{N_{E_-} + N_{E_+}} \tag{6.2}$$

$$x_s = \frac{\sum_{n=1}^{N_{E_+}} P(v = E_+|\mathbf{u}_n)}{N_{E_+}} \tag{6.3}$$

Nuclear segmentation-based method

An alternative method for the computation of global features was based on a multi-level algorithm for monochromatic images that explicitly segments nuclei, developed and implemented by Dr Michele Sciarabba [91]. Given a grey-level image $\mathbf{A}$, 16 nested binary images (called *levels*) $\mathbf{A}_i$ $(i = 1, \ldots, 16)$ were created such that $\mathbf{A}_i(x, y) = 1$ iff $\mathbf{A}(x, y) \leq T_i$, where $T_i = i \times 16$ denotes an intensity threshold and x and y denote pixel coordinates. Connected components at the multiple levels formed a tree structure, each component at a given level having as its parent a component that contained it at the level above. For each component, a global shape index was computed, based on elongation and solidity, to reflect the compatibility of the component's shape with nuclear shape. The leaves of the component tree whose shape was deemed compatible with that of a nucleus are referred to as cores, and corresponded to the darkest compatible regions within each component in the original image $\mathbf{A}$. These cores were grown level by level as far as they kept a compatible shape and did not join with other compatible components.

This algorithm was applied to four different grey-level images derived from the original colour image. The first image was obtained by turning to white all pixels of colour dissimilar to that of fully unstained nuclei (here also referred to as E_-), and then converting the whole image to grey-scale. The second image was obtained similarly, but by turning to white all pixels of colour dissimilar to that of fully stained nuclei (referred to as E_+). Using the above-describe algorithm on these two images, heavily stained nuclei were detected. This approach, however, was not well suited to deal with textured connected components, as was the case with less stained nuclei (both E_- and E_+). These were recognised using the third and fourth grey-level images, created via two pseudo-hue functions (specifically, $2 \times b - r - g$ and its opposite $-2 \times b + r + g$) and rescaling the results so as to have values in the $[0, \ldots, 255]$ range. So, in the third image, stained pixels (regardless of their intensity) had values near to zero, and unstained pixels had values near to 255, while in the fourth image the opposite held. In both cases, pixels that could safely be classified as background (via the previously described pixel labelling technique) were discarded.

In order to compute an x_p global feature formalising the q_p value, each detected nucleus was first classified as either E_+ or E_-. If the nucleus contained more than 20% of stained pixels, it was marked as E_+, otherwise as E_-. Nuclei near the decision boundary of 20% were not marked. The x_p feature was then computed as the proportion of epithelial nuclei (both E_- and E_+) that were E_+. In turn, to obtain a global feature x_s formalising the q_s value, the strength of staining was first estimated for each individual E_+ nucleus, and then for the whole spot from individual values. Each nucleus was associated with the intensity of the 20%-quantile darkest pixel, and the formalised predictor for the whole spot was determined as the 20%-quantile darkest nucleus.

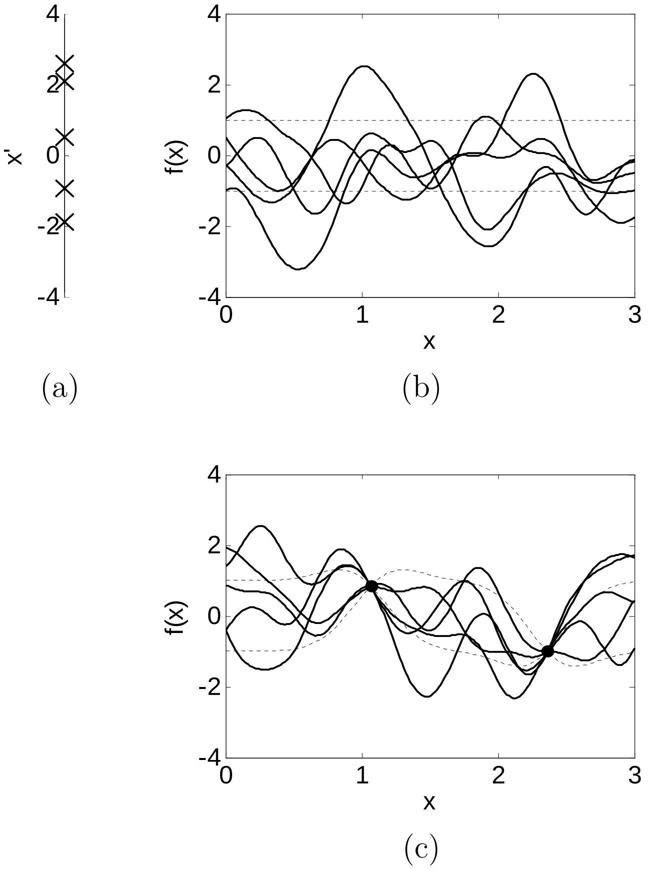
6.3 Gaussian processes for ordinal regression

6.3.1 Overview

This section presents a overview of Gaussian processes applied to ordinal regression, mainly based on the 2005 paper by Chu and Ghahramani [22] and the Introduction to the book *Gaussian Processes for Machine Learning* by Rasmussen and Williams [81]. For a comprehensive discussion, those references should consulted.

The Gaussian probability *distribution* can be generalised into the concept of Gaussian *process*. A uni-variate (or multi-variate) probability distribution governs random variables that are scalars (or vectors), while a stochastic process describes the properties of functions. This is illustrated in Figure 6.1. Figure 6.1(a) shows five samples drawn from a zero-mean Gaussian distribution. In turn, Figure 6.1(b) shows five functions sampled randomly from the *prior* of a Gaussian process that favours smooth functions (by using a radial basis covariance function). At any given point x the possible values of $f(x)$ follow a zero-mean Gaussian distribution, of which the values of the five depicted functions are samples. Standard deviation dotted lines suggest the (constant) width of the probability distribution of $f(x)$ for all input values.

If observations of $f(x)$ at two points are then incorporated into the process, five new functions can be sampled from the *posterior* distribution of the Gaussian process, as shown in Figure 6.1(c). At each point x the possible values of $f(x)$ no longer necessarily follow a distribution centred around zero. Moreover,



(a)

(b)

(c)

Figure 6.1: Samples from (a) a Gaussian distribution, (b) a Gaussian process prior, and (c) a Gaussian process posterior. (Plots obtained with the Gaussian Process Toolbox developed by Neil Lawrence [56].)

all sample functions pass through the points at which $f(x)$ is observed and, as one approaches those points, the possible values of $f(x)$ follow increasingly narrower and more peaked normal distributions (which become delta functions at the actual observed points). This illustrates a simple one-dimensional regression problem, but the considered function could naturally have a multi-dimensional domain.

In Gaussian process ordinal regression, the ordinal target t associated with an input x is assumed to be dependent on the function $f(x)$. This is modelled by dividing the real co-domain of $f(x)$ into a series of contiguous intervals, which map real values of $f(x)$ into ordinal targets while enforcing the ordinal constraints. If these intervals were assumed equal in width, the ordinal target for a test input x could be easily predicted by choosing the interval containing the peak of the posterior probability distribution of $f(x)$ associated with the test input. However, the intervals that map real values into ordinals are not assumed equal in width (in fact, the first and last intervals are defined as extending from and to infinity, respectively). The boundaries between these intervals are part of the model's parameters that need to be learned from training data. Figure 6.2 illustrates an example where a Gaussian process posterior is used to predict the ordinal target for a given test input. The posterior distribution of $f(x)$ is explicitly plotted (sideways) at the test input. This example involves four ordinal targets, which correspond to the intervals $(-\infty, b_1]$, $(b_1, b_2]$, $(b_2, b_3]$, and (b_3, ∞). Intuitively, it can be seen that the predicted target should be that associated with the largest partial integral of the distribution of $f(x)$ associated with the test input, in this case the third ordinal (whose interval happens not to contain the peak of the distribution). As previously in Figure 6.1(c), standard deviation dotted lines suggest the non-constant width of the probability distribution of $f(x)$ (so that, at the test input, the peak of the depicted distribution lies exactly between the two standard deviation lines).

A general problem can now be considered, involving C ordinal categories $t \in \{1, ..., C\}$ and a training set of N observations $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$, where each observation is composed of an input vector $\mathbf{x}_n$ and associated observed target t_n. Since the categories are assumed to be consecutive positive integers (without loss of generality), the upper and lower boundaries of the interval associated with each target t can be denoted by b_t and b_{t-1}, respectively. These boundaries are part of the considered Gaussian process hyper-parameters, which can be denoted by θ.

It can be demonstrated that, given a test input vector $\mathbf{x}$, the predictive probability

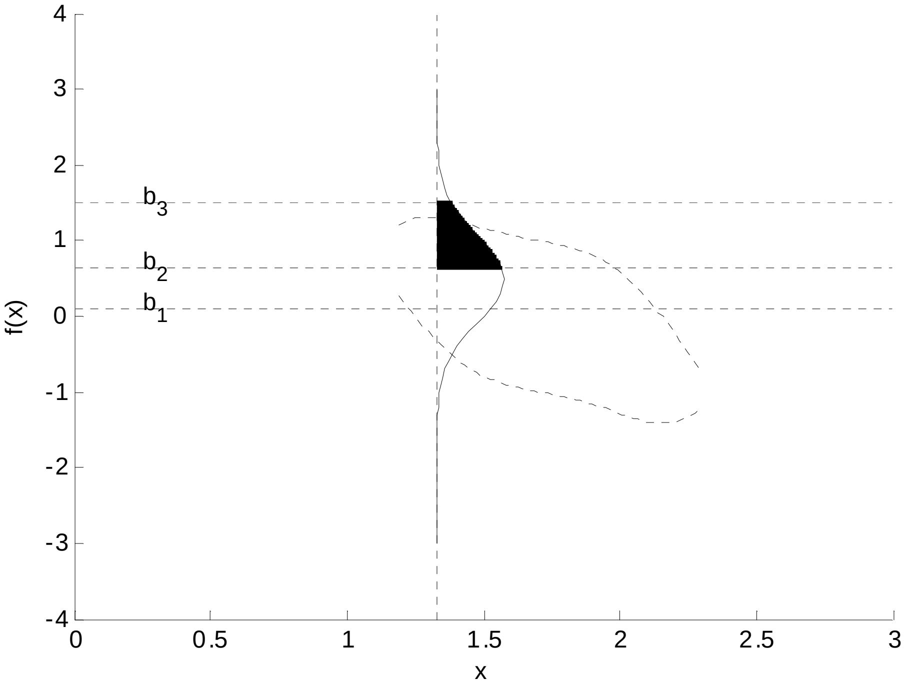


Figure 6.2: Prediction of ordinal target for a test input, based on a Gaussian process.

$P(t|\mathbf{x}, \mathcal{D}, \theta)$ of an ordinal target $t \in \{1, 2, ..., C\}$ is given by an integral over the posterior distribution $P(f(\mathbf{x})|\mathcal{D}, \theta)$, as would be expected from the above discussion of a simple one-dimensional example. This predictive probability is expressed in Equations (6.4) and (6.5), where μ and σ are the mean and standard deviation of the posterior of $f(\mathbf{x})$, respectively, and σ^2_{noise} is a noise variance that shall be explained below. The notation $\Phi(a)$ refers to the integral of a Gaussian with zero mean and unit variance, from $-\infty$ to a.

$$P(t|\mathbf{x}, \mathcal{D}, \theta) = \int_{b_{t-1}}^{b_t} P(t|f(\mathbf{x}), \theta) P(f(\mathbf{x})|\mathcal{D}, \theta) df(\mathbf{x}) \tag{6.4}$$

$$= \Phi(\frac{b_t - \mu}{\sqrt{\sigma^2_{noise} - \sigma^2}}) - \Phi(\frac{b_{t-1} - \mu}{\sqrt{\sigma^2_{noise} - \sigma^2}}) \tag{6.5}$$

The predictive probabilities $P(t|\mathbf{x}, \mathcal{D}, \theta)$ for all targets $t \in \{1, 2, ..., C\}$ provide a probabilistic output in a regression problem, in a similar way to the estimates of class posterior probabilities output by neural networks for all classes $t \in \{T_1, \ldots, T_C\}$ in a classification problem, as discussed in Section 4.3. Given a test input $\mathbf{x}$, the predicted ordinal target can be chosen as that associated with the highest probability. However, as discussed later in Sections 6.5 and 6.4, the probabilistic outputs themselves proved to be useful in the present work.

The likelihood term $P(t|f(\mathbf{x}),\theta)$ in Equation (6.4) is the probability of observing the ordinal target t given a value of the function $f(\mathbf{x})$. In ideal noise-free cases, this term would simply be a "switch" function $P_{ideal}(t|f(\mathbf{x}),\theta)$ equating to 1 for values of $f(\mathbf{x})$ falling within the interval associated with the ordinal target t and to 0 otherwise. In the presence of noise from training inputs or targets, the function $f(\mathbf{x})$ is assumed to be contaminated by Gaussian noise with zero mean and variance σ^2_{noise}. This noise variance is included in the model's parameters θ. The likelihood term then assumes the form shown in Equation (6.6), where $\mathcal{N}(\delta;0,\sigma^2_{noise})$ denotes the Gaussian distribution of noise variable δ.

$$P(t|f(\mathbf{x}),\theta) = \int P_{ideal}(t|f(\mathbf{x}),\theta)\mathcal{N}(\delta;0,\sigma^2_{noise})d\delta \tag{6.6}$$

Applying Bayes' rule, the term $P(f(\mathbf{x})|\mathcal{D},\theta)$ in Equation (6.4), corresponding to the posterior probability distribution, can be expressed as in Equation (6.7).

$$P(f(\mathbf{x})|\mathcal{D},\theta) = \frac{P(\mathcal{D}|f(\mathbf{x}),\theta)P(f(\mathbf{x}))}{P(\mathcal{D}|\theta)} \tag{6.7}$$

Considering that each training ordinal target t_n is paired with a training input vector $\mathbf{x}_n$, the likelihood term $P(\mathcal{D}|f(\mathbf{x}),\theta)$ is the joint probability of observing the training data $\mathcal{D} = \{(\mathbf{x}_1,t_1),\ldots,(\mathbf{x}_N,t_N)\}$ given the values $\{f(\mathbf{x}_1),\ldots,f(\mathbf{x}_N)\}$, and can be evaluated generally as a product of the likelihoods for individual observations, as shown in Equation (6.8).

$$P(\mathcal{D}|f(\mathbf{x}),\theta) = \prod_{n=1}^{N} P(t_n|f(\mathbf{x}_n),\theta) \tag{6.8}$$

Each individual likelihood $P(t_n|f(\mathbf{x}_n),\theta)$ can be expressed in a form similar to that shown in Equation (6.6) (considering a training target t_n instead of a test target t).

In Equation (6.7), $P(f(\mathbf{x}))$ is the prior distribution of $f(\mathbf{x})$. The Gaussian process prior can be fully specified by the covariance matrix for the finite set of zero-mean random variables $\{f(\mathbf{x}_1),\ldots,f(\mathbf{x}_N)\}$, associated with the training input vectors $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$. In turn, the covariance between any two such variables $f(\mathbf{x}_n)$ and $f(\mathbf{x}_m)$ can be defined by Mercer kernel functions [99, 88], such as the linear and Gaussian kernels defined in Equations (6.9) and (6.10), respectively. In these formulas, the elements of a D-dimensional input vector $\mathbf{x}_n$ are denoted by x_d^n ($d \in \{1, \ldots, D\}$). The constants κ_o and κ_a are included in the model's parameters θ.

$$\begin{aligned} Cov[f(\mathbf{x}_n), f(\mathbf{x}_m)] &= \kappa_o \sum_{d=1}^{D} \kappa_a x_d^n x_d^m && (6.9) \\ Cov[f(\mathbf{x}_n), f(\mathbf{x}_m)] &= \kappa_o exp(-\frac{\kappa_a}{2} \sum_{d=1}^{D} (x_d^n - x_d^m)^2) && (6.10) \end{aligned}$$

Considering the case in which a Gaussian kernel is employed, the Gaussian process prior $P(f(\mathbf{x}))$ can be expressed as the multi-variate Gaussian shown in Equation (6.11), where Z_f is a normalisation factor equal to $(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}$, $\mathbf{f}$ denotes the vector of random variables $(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))^{\mathrm{T}}$, and $\mathbf{\Sigma}$ is the covariance matrix whose nm-th element was defined in Equation (6.10).

$$P(f(\mathbf{x})) = \frac{1}{Z_f} \exp(-\frac{1}{2}\mathbf{f}^T \mathbf{\Sigma}^{-1} \mathbf{f}) \qquad (6.11)$$

The model's hyper-parameters θ, as well as the normalisation factor $P(\mathcal{D}|\theta)$ in Equation (6.7) (the evidence for θ), can be learned through a model adaptation technique, such as maximum *a posteriori* (MAP) estimation or expectation propagation (EP). These parameters include the ordinal interval boundaries b_t, $t \in \{1, ..., C-1\}$, the noise variance σ^2_{noise}, and the kernel constants κ_o and κ_a. It is also worth noting that the automatic relevance determination method proposed by MacKay [61] and Neal [69] can be embedded into the employed covariance function.

6.3.2 Scoring of spots

Models based on Gaussian processes for ordinal regression, as well as generalised linear models (GLMs) and multi-layer perceptrons (MLPs), were trained to predict the q_p and q_s values of new spots, using as input the pair of global features extracted for each spot, that is, $\mathbf{x} = (x_p, x_s)^{\mathrm{T}}$. In other words, the employed algorithms were trained to predict the two quickscore *integer* values for each test spot from the pair of *real* numbers that constituted the formalised quickscore extracted for that spot. It is fair to assume that valuable information could be lost in this process; nevertheless, pathologists are trained to base their decisions on quickscore integer values. This was, therefore, the type of output required for the developed system.

All GLMs were trained through the iterated re-weighted least squares algorithm, whereas the learning algorithm used with all MLPs was scaled conjugate gradients optimisation. For both types of network, softmax was chosen as the activation function.

6.4 Experiments and results

6.4.1 Data and code

The data used in the scoring experiments consisted of colour images of 190 breast TMA spots subjected to progesterone receptor immunohistochemistry. These spots belonged to the tumour and normal types and the quickscores assigned to them by a pathologist were known (scoring data was available for two separate manual assessment sessions, but only the first session was considered here). For all spots, a manually drawn ROI was also available, although these were not created by a pathologist but by Dr Michele Sciarabba. Figures 6.9(d) and (f) (on page 107) show examples of manually drawn ROIs.

On each of 20 spots, a circular subregion 500 pixels in diameter was randomly selected and manually annotated by the author. These annotations were partially reviewed by a pathologist. In this way, the contours of approximately 700 epithelial nuclei were marked and labelled as either immunonegative or immunopositive. Figure 6.3 shows an example of an annotated subregion (although only the contouring of the nuclei is shown, not their labelling). The left third of this circular subregion is populated with epithelial cells, both stained and non-stained, whereas its remainder contains connective tissue.

The code for extraction of local and global features, as well as the code used to control the scoring experiments (that is, to prepare the data, call the algorithms, and process their results), was implemented in Matlab. Local feature extraction made use of the Sussex convolution function [103]. The Netlab [67] implementations of the GLM and the MLP were used. The C implementation of Gaussian process ordinal regression developed and made publicly available by Chu Wei [21, 22] was used as well.

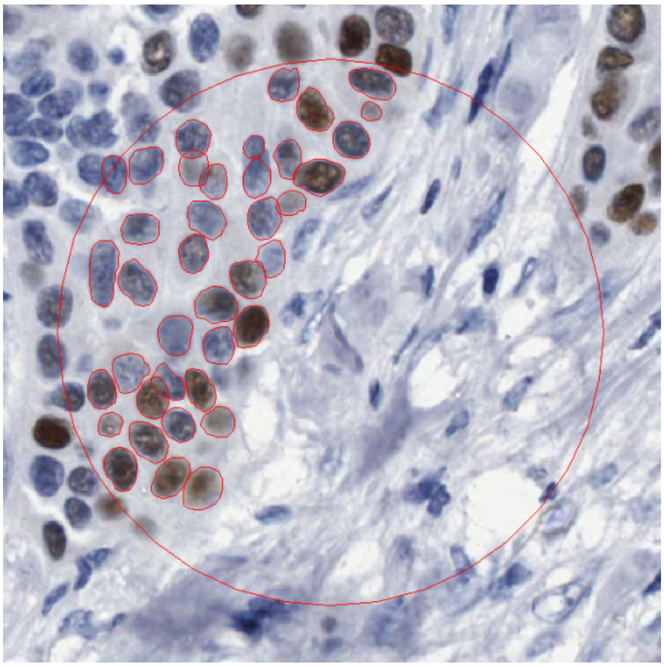


Figure 6.3: Manual contouring (in red) of epithelial nuclei, within a circular subregion of a tissue microarray spot stained for progesterone receptor.

6.4.2 Pixel labelling

In order to separately assess the performance of the pixel labelling technique described in Section 6.2.2, a leave-10-out experiment was executed over the 20 annotated subregions. At each pass of the experiment, the class-conditional probability distributions of local features given the pixel class were estimated from data associated with 10 subregions. More precisely, in order to estimate the $P(r, g, b|v)$ distributions, each of the r, g, and b components of the colour space was divided into 16 bins of equal width, whereas the $P(d_{k,\sigma}|v)$ distributions were estimated after dividing the range of each differential invariant feature into 64 equal bins.

Equation (6.1) was then used for each pixel class, in order to label pixels in the remaining 10 subregions. This was done first using local feature jets containing only the r, g, and b colour values, and then using jets containing colour values and the four differential invariants $d_{1,8}$ to $d_{4,8}$ (that is, considering only the scale $\sigma = 8$). Table 6.1 shows the confusion matrices that resulted from the leave-10-out experiment, and Figure 6.4 shows the pixel labelling results for the subregion shown in Figure 6.3.

For the simpler case in which only r, g, and b colour information was used, Figures 6.5(a), (b), and (c) illustrate the estimated $P(r, g, b|v)$ class-conditional probability distributions, and Figures 6.5(d), (e), and (f) show the resulting class

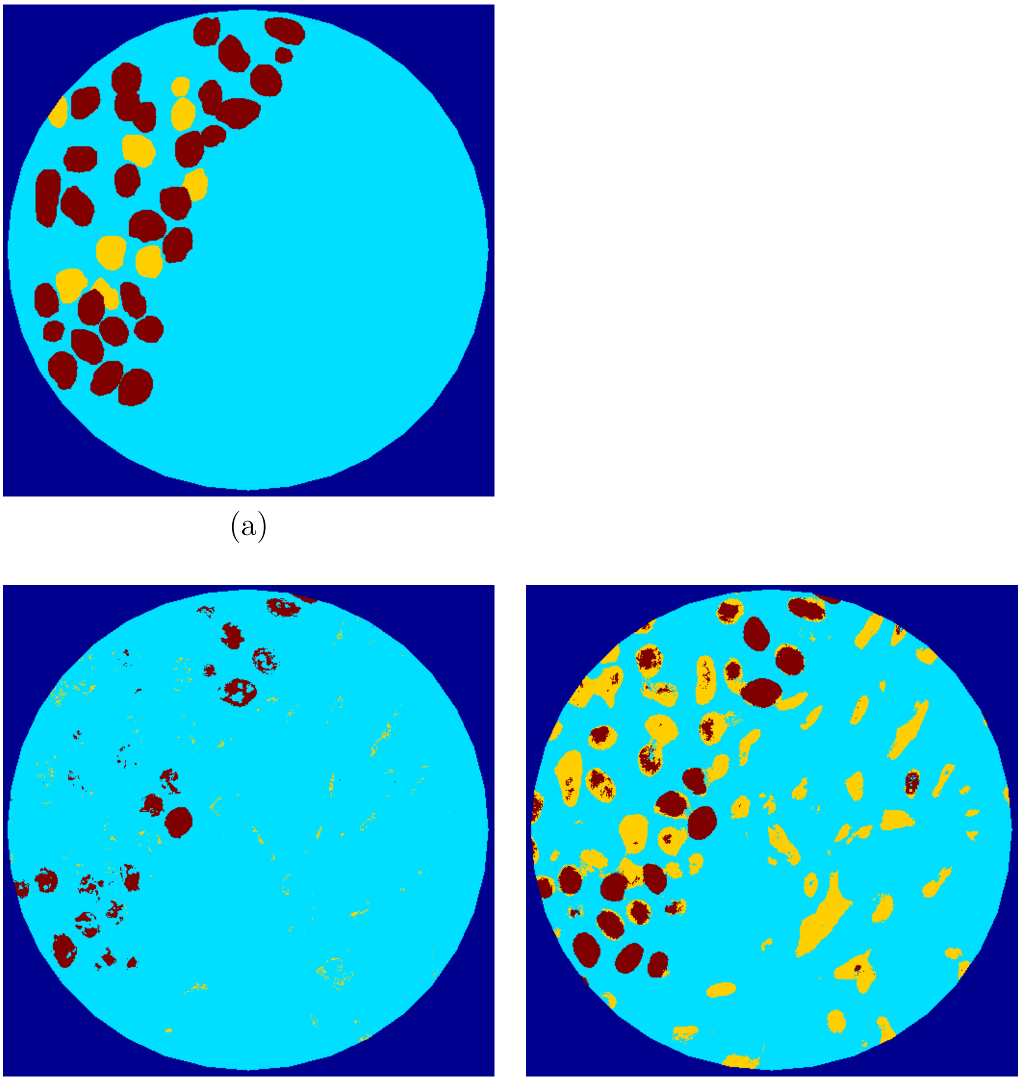
Table 6.1: Confusion matrices for pixel labelling (in thousands of pixels), (a) using only r, g, and b and (b) using also differential invariants $d_{1,8}$ to $d_{4,8}$.

(a)

True	Predicted		
	B	E_-	E_+
B	1666	14	6
E_-	98	17	0
E_+	93	1	62

(b)

True	Predicted		
	B	E_-	E_+
B	1503	155	27
E_-	31	81	3
E_+	35	15	106



(a)

(b)

(c)

Figure 6.4: (a) Manual labelling of a circular subregion of tissue, and pixel labelling results (b) using only r, g, and b and (c) using also differential invariants $d_{1,8}$ to $d_{4,8}$. (Blue: background; yellow: E_- pixels; and brown: E_+ pixels.)

posterior probabilities $P(v|r, g, b)$.

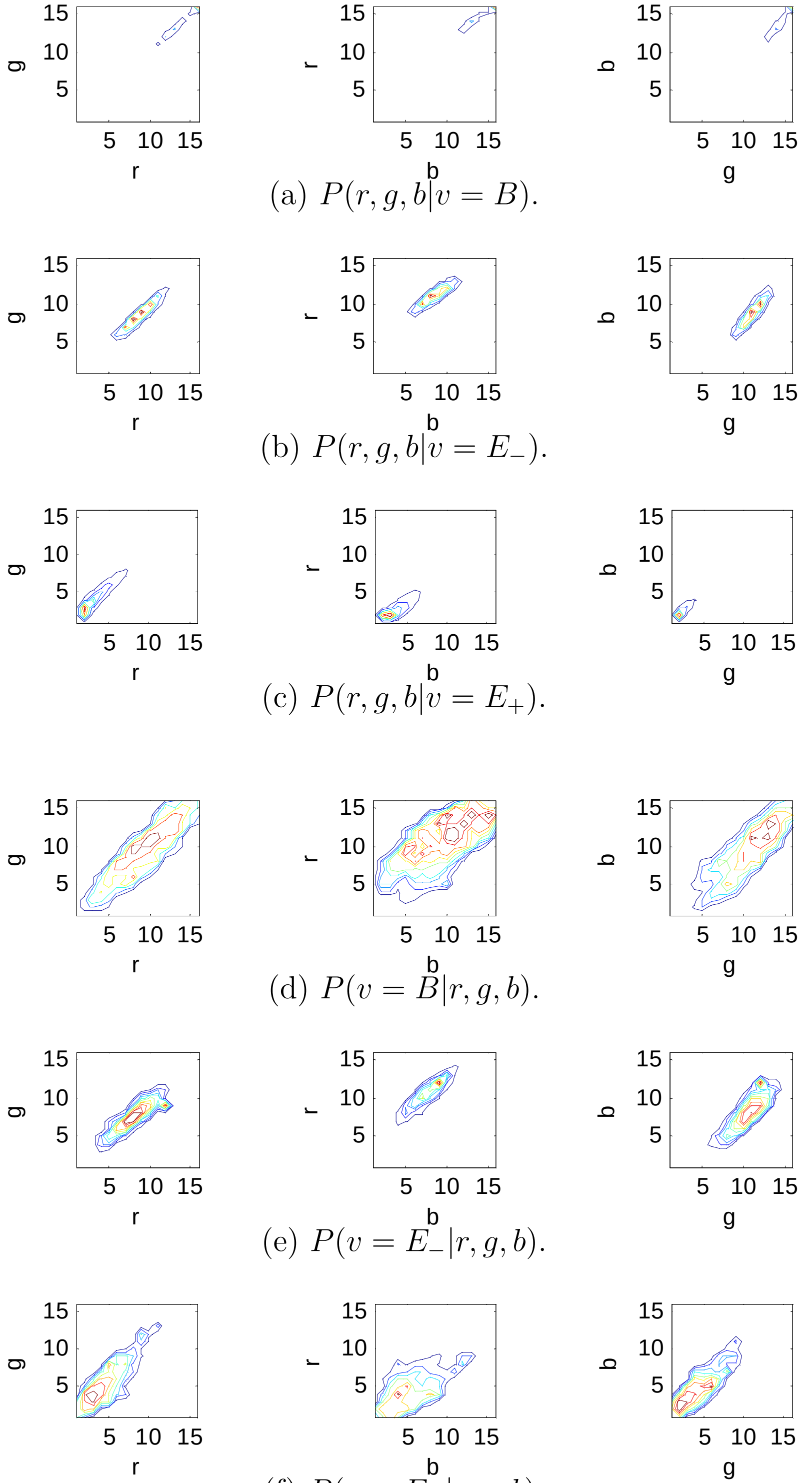
The performance of the nuclear segmentation method was not evaluated quantitatively. Figures 6.6(b) and (c) illustrate the differences in quality of the results obtained via pixel labelling (with full local jets containing 15 features) and via nuclear segmentation, respectively, for the small tissue region shown in Figure 6.6(a). Figures 6.6(e) and (f) illustrate the results obtained for the subregion shown in Figure 6.6(d). (The dark blue region around the border of Figure 6.6(e) constitutes the outside of a mask defining the subregion, so that only pixels or nuclei within the mask were labelled or segmented.)

6.4.3 Comparison of ordinal regression with neural networks

A first scoring experiment was carried out to compare the performance of classification based on neural networks with that of ordinal regression based on Gaussian processes. This experiment used only global features extracted via the local patch-based technique explained in Section 6.2.2. All 20 annotated subregions contributed to the estimation of class-conditional probabilities of local features given the pixel class.

In the prediction of scores through classification, GLMs and MLPs were compared. The number of hidden units M and the regularisation constant A for MLPs were fixed at 3 and 0.1, respectively. In the prediction of scores through ordinal regression, the two parameter learning approaches referred to in Section 6.3 were compared, namely MAP (maximum *a posteriori*) estimation and EP (expectation propagation). In addition, for each learning technique, linear and Gaussian kernels were compared. Neither of these kernels incorporated the automatic relevance determination mechanism.

Leave-one-out experiments were carried out to predict the q_p and q_s values of the 190 available spots. These experiments were then repeated to predict *collapsed* quickscore values, obtained as shown in Equations (6.12) and (6.13).



(a) $P(r, g, b|v = B)$.

(b) $P(r, g, b|v = E_-)$.

(c) $P(r, g, b|v = E_+)$.

(d) $P(v = B|r, g, b)$.

(e) $P(v = E_-|r, g, b)$.

(f) $P(v = E_+|r, g, b)$.

Figure 6.5: (a,b,c) Class-conditional probability distributions of binned colour values given each pixel class, estimated from annotated data. (d,e,f) Class posterior probabilities for each pixel class given the binned colour value. (Colder colours: probabilities closer to 0; and hotter colours: larger probabilities.)

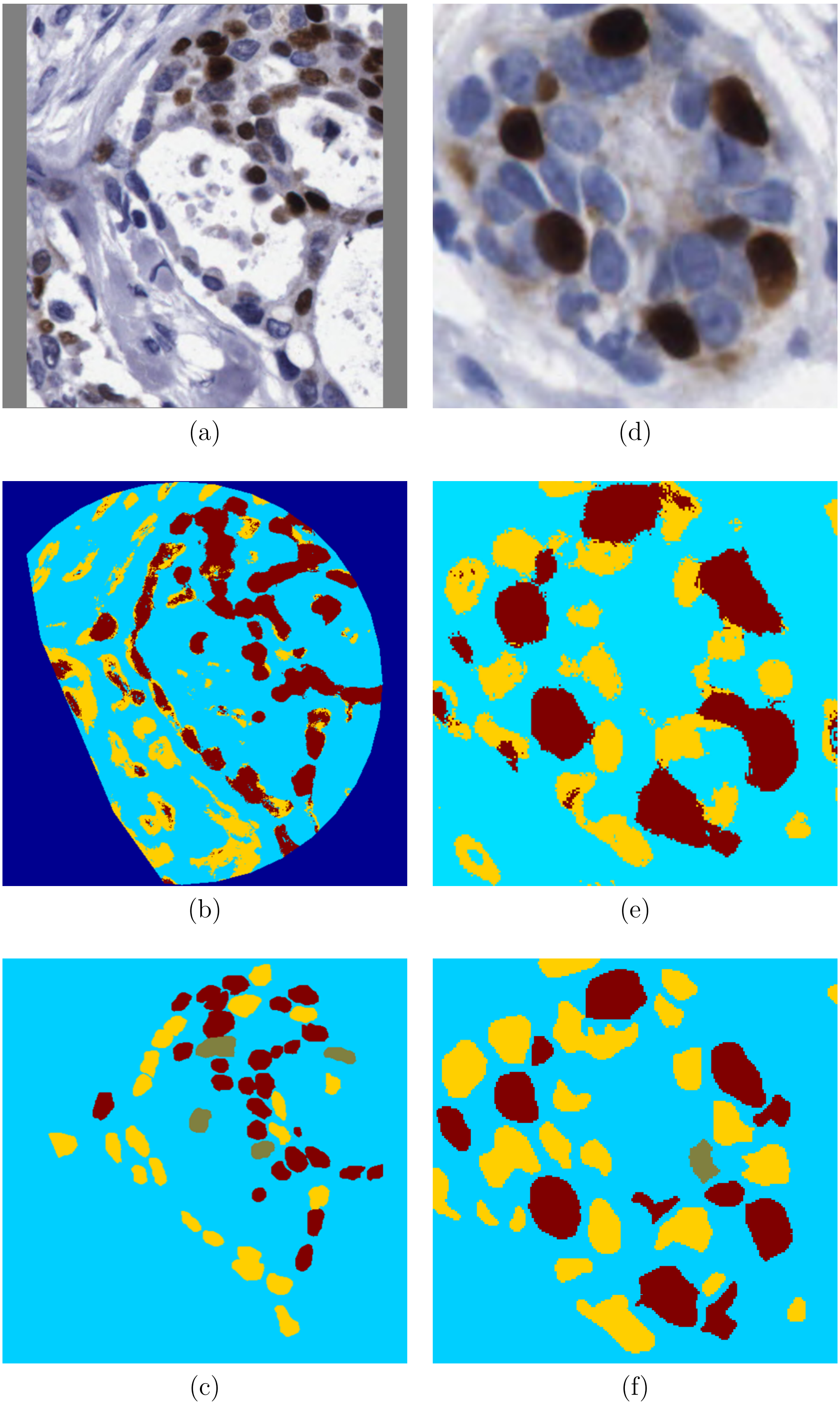


Figure 6.6: (a,d) Small tissue regions, and corresponding results of (b,e) pixel labelling and (c,f) nuclear segmentation. (Light blue: background; yellow: E_{-} pixels or nuclei; dark brown: E_{+} pixels or nuclei; and light brown: nuclei near the decision boundary.)

$$q_{p.collapsed} = \begin{cases} 0 & \text{if } q_p = 0 \\ 1 & \text{if } q_p \in \{1, 2\} \\ 2 & \text{if } q_p \in \{3, 4\} \\ 3 & \text{if } q_p \in \{5, 6\} \end{cases} \tag{6.12}$$

$$q_{s.collapsed} = \begin{cases} 0 & \text{if } q_s = 0 \\ 1 & \text{if } q_s \in \{1, 2\} \\ 2 & \text{if } q_s = 3 \end{cases} \tag{6.13}$$

The obtained results are presented in Table 6.2 in the form of a mean absolute error for each experiment and associated standard deviation. This error corresponds to the average deviation of the predictions from the true targets, as defined in Equation (6.14), where N is the number of predictions, $\hat{t}_n \in \{1, ..., C\}$ denotes the n-th predicted target, and $t_n \in \{1, ..., C\}$ denotes the n-th true target.

$$\overline{e}_{abs} = \frac{\sum_{n=1}^{N} |\hat{t}_n - t_n|}{N} \tag{6.14}$$

In Table 6.2, the lowest errors on each row are printed in boldface. The predictions obtained for non-collapsed targets q_p and q_s were also collapsed *a posteriori* and their mean absolute errors recomputed, so as to render them comparable with the results obtained for collapsed targets. In Table 6.2, these targets collapsed *a posteriori* are denoted by $q_{p.cap}$ and $q_{s.cap}$. Below each result in the table, a normalised result (obtained through division by the number of targets) is presented as well, in green colour.

For each experiment, a confusion matrix was computed. The matrices for some of the experiments are shown in Table 6.3.

As mentioned in Sections 4.3 and 6.3, both the classification and the ordinal regression algorithms output, along with each prediction, a posterior probability distribution over the output targets. The entropy of a posterior distribution can be used as a simple measure of classification or regression confidence, since lower entropies correspond to more peaked posterior distributions and therefore to more confident predictions. For two of the experiments, Figure 6.7 shows the fraction of test spots that can be predicted below a given entropy threshold. Also shown is the mean absolute error computed over each fraction of spots.

Table 6.2: Mean and standard deviation of the absolute error, for the various experiments. Normalised values in green.

Predicted target	Algorithm					
	Classification		Gaussian process ordinal regression			
	GLM	MLP	MAP		EP	
			Linear	Gaussian	Linear	Gaussian
q_p	1.400 ±1.677	0.926 ±1.215	1.126 ±1.397	0.921 ±1.172	0.900 ±1.129	**0.888 ±1.175**
	0.200 ±0.240	0.132 ±0.174	0.161 ±0.200	0.132 ±0.167	0.129 ±0.161	0.127 ±0.168
$q_{p.cap}$	0.774 ±0.935	0.516 ±0.733	0.626 ±0.805	0.537 ±0.702	**0.500 ±0.680**	0.503 ±0.698
	0.194 ±0.234	0.129 ±0.183	0.157 ±0.201	0.134 ±0.176	0.125 ±0.170	0.126 ±0.175
$q_{p.collapsed}$	0.684 ±0.870	0.432 ±0.677	0.579 ±0.757	**0.426 ±0.619**	0.463 ±0.639	**0.426 ±0.611**
	0.171 ±0.218	0.108 ±0.169	0.145 ±0.189	0.107 ±0.155	0.116 ±0.160	0.107 ±0.153
q_s	0.937 ±1.097	**0.763 ±0.988**	0.937 ±1.106	0.784 ±1.003	0.800 ±1.025	0.779 ±0.994
	0.234 ±0.274	0.191 ±0.247	0.234 ±0.277	0.196 ±0.251	0.200 ±0.256	0.195 ±0.249
$q_{s.cap}$	0.674 ±0.727	**0.547 ±0.655**	0.663 ±0.729	0.558 ±0.662	0.568 ±0.677	0.553 ±0.655
	0.225 ±0.242	0.182 ±0.218	0.221 ±0.243	0.186 ±0.221	0.189 ±0.226	0.184 ±0.218
$q_{s.collapsed}$	0.589 ±0.626	0.495 ±0.589	0.526 ±0.606	0.495 ±0.561	**0.489 ±0.589**	**0.489 ±0.561**
	0.196 ±0.209	0.165 ±0.196	0.175 ±0.202	0.165 ±0.187	0.163 ±0.196	0.163 ±0.187

Table 6.3: Confusion matrices for some of the experiments.

(a) q_p, EP, Gaussian.

Test	Predicted						
	0	1	2	3	4	5	6
0	65	0	4	0	1	0	0
1	16	0	2	0	0	0	0
2	13	0	8	0	3	0	1
3	3	0	8	0	4	3	0
4	4	0	4	0	4	4	2
5	3	0	3	0	2	1	7
6	0	0	1	0	0	4	17

(b) q_s, MLP.

Test	Predicted			
	0	1	2	3
0	68	0	0	4
1	15	0	0	16
2	8	0	0	29
3	13	0	2	35

(c) $q_{p.collapsed}$, EP, Gaussian.

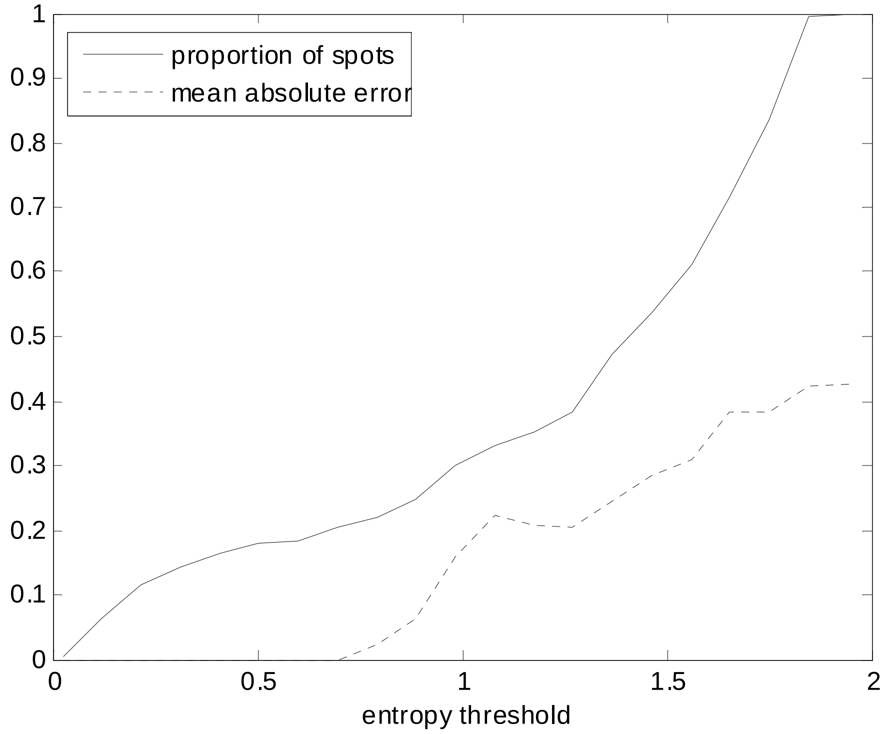
Test	Predicted			
	0	1	2	3
0	59	11	2	0
1	21	15	6	1
2	3	9	18	7
3	1	3	6	28

(d) $q_{s.collapsed}$, EP, Gaussian.

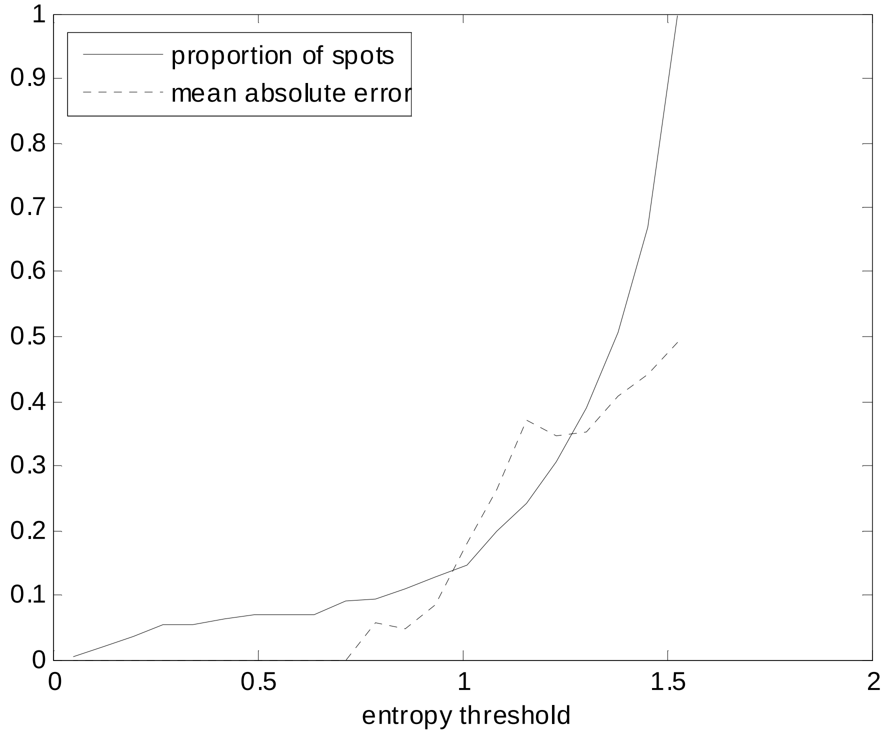
Test	Predicted		
	0	1	2
0	52	20	0
1	16	28	24
2	6	21	23

(e) $q_{p.collapsed}$, GLM.

Test	Predicted			
	0	1	2	3
0	65	0	3	4
1	39	0	2	2
2	19	0	9	9
3	5	0	5	28



(a) $q_{p.collapsed}$.



(b) $q_{s.collapsed}$.

Figure 6.7: Fraction of scored spots and associated mean absolute error, for different entropy thresholds and for both quickscore values. From the experiments based on expectation propagation and Gaussian kernels. (Lower entropy means higher confidence.)

6.4.4 Comparison of global feature extraction methods

A second scoring experiment relied only on ordinal regression, using Gaussian kernels that incorporated automatic relevance determination and learning the model's hyper-parameters via the MAP estimation technique. The main goal of this experiment was to assess how the regression performance would be affected if global features were computed not via the technique based on local patches but via nuclear segmentation, and if that computation were based not on the whole of each spot but only on its manually defined ROI.

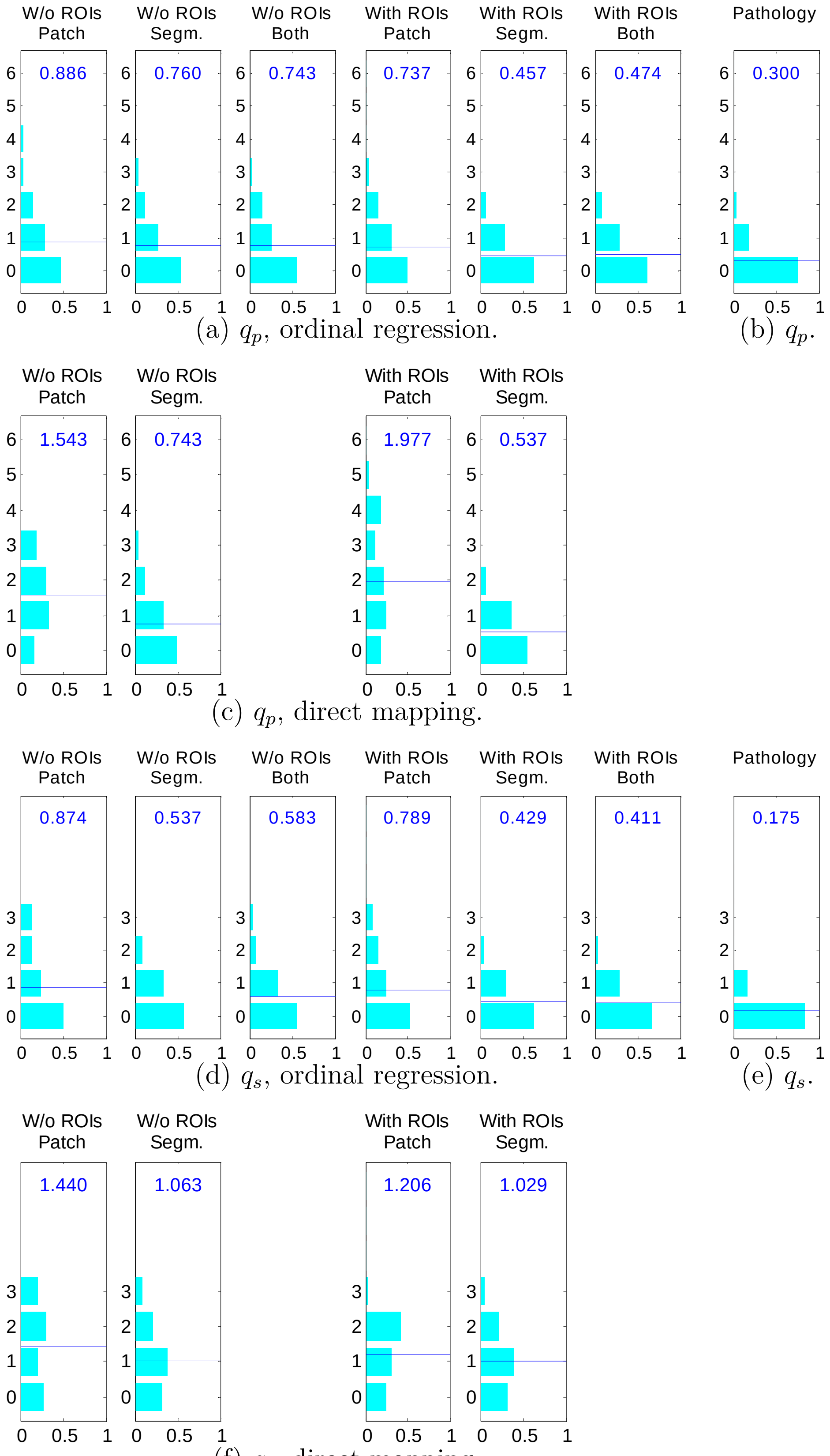
To achieve this, for each quickscore value (q_p and q_s), several leave-one-out experiments were carried out over 175 spots. Specifically, three experiments were run without relying on ROIs, using global features obtained first via local patches, then via nuclear segmentation, and finally via both methods (that is, using not two but four global features to characterise each spot). These three experiments were then repeated taking the ROIs into account.

Figures 6.8(a) and (d) summarise the results obtained for the prediction of q_p and q_s values, respectively. For each individual experiment, the distribution of absolute errors over the 175 predictions is shown by means of a sideways histogram. The mean absolute error is also shown, both numerically and as a horizontal line overlaid on each histogram.

Table 6.4 shows example confusion matrices for two of the q_p prediction experiments: without ROIs and local patch-based; and with ROIs and nuclear segmentation-based.

Figures 6.9(a) and (c) show two interesting examples of spots. Spot 6.9(a) had true q_p and q_s values of 0 and 0, respectively. Without using ROIs, formalised quickscores obtained via local patches led to predictions of 4 and 3, respectively; but, keeping the non-use of ROIs and switching to nuclear segmentation, correct predictions were achieved, both for q_p and for q_s. Figure 6.9(b) shows the (largely erroneous) result of pixel classification for the same spot. In turn, spot 6.9(c) had true q_p and q_s values of 2 and 1, respectively. Without using ROIs, formalised quickscores based on nuclear segmentation led to predictions of 0 and 0, respectively; but, using the spot's ROI while keeping the segmentation-based approach, correct predictions were achieved, again both for q_p and for q_s. Figure 6.9(d) shows the same spot with its ROI highlighted.

For the experiment based on ROIs and segmentation, Figures 6.10(a) and (b) show the fraction of test spots that can be processed below a given entropy



(a) q_p, ordinal regression.

(b) q_p.

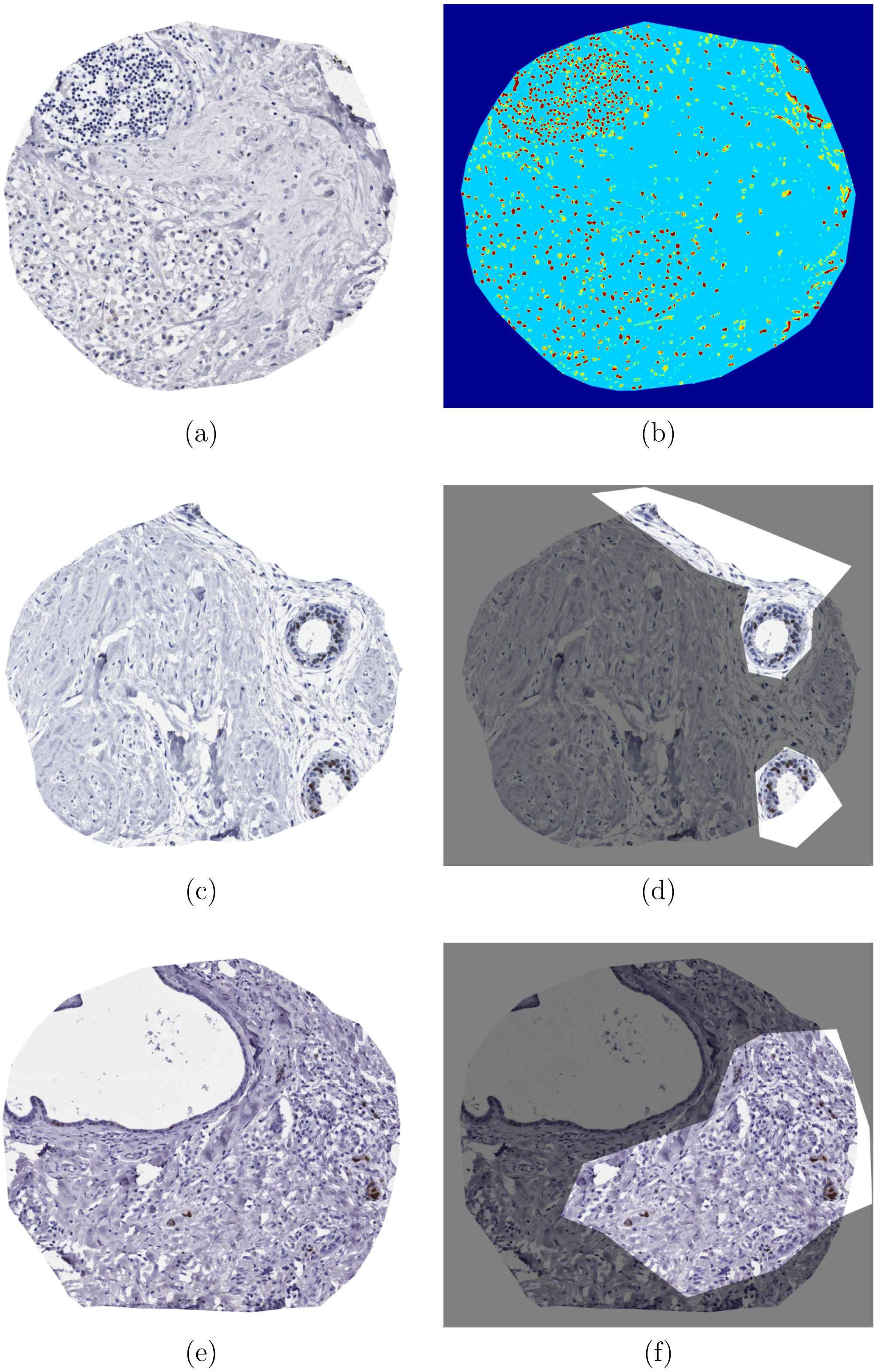
(c) q_p, direct mapping.

(d) q_s, ordinal regression.

(e) q_s.

(f) q_s, direct mapping.

Figure 6.8: Distributions of q_p and q_s absolute prediction errors for different global feature extraction approaches, (a,d) using ordinal regression and (c,f) via direct mapping. (b,e) Distributions of q_p and q_s absolute intra-observer disagreements between two scoring sessions.



(a) (b)

(c) (d)

(e) (f)

Figure 6.9: (a,b) A spot correctly predicted after switching from pixel classification to nuclear segmentation, along with pixel classification results. (Light blue: background; yellow: E_- pixels; and red: E_+pixels.) (c,d) A spot correctly predicted only when its manually drawn region of interest was taken into account, along with a depiction of that region. (e,f) A heterogeneous spot containing both normal and tumour tissue, along with its manually drawn region of interest excluding the normal tissue.

Table 6.4: Confusion matrices for two of the experiments.

(a) q_p, W/o ROIs, Patch-based.

Test	Predicted						
	0	1	2	3	4	5	6
0	54	0	06	0	1	0	0
1	16	0	2	0	0	0	0
2	10	0	10	1	2	0	1
3	4	0	6	0	7	1	0
4	4	0	4	0	4	4	2
5	1	0	3	0	3	0	8
6	0	0	1	0	0	4	16

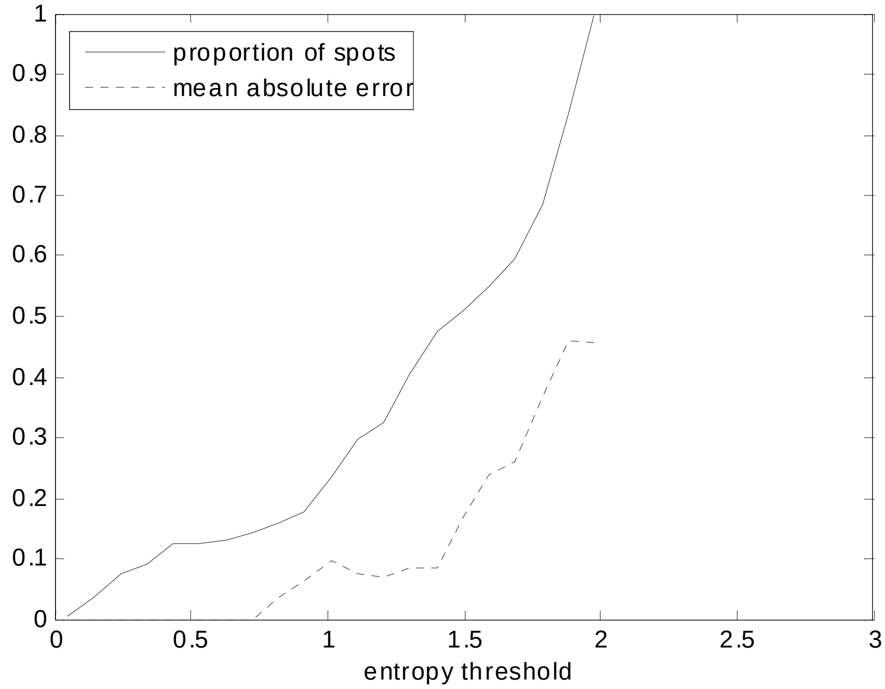
(b) q_p, With ROIs, Segmentation-based.

Test	Predicted						
	0	1	2	3	4	5	6
0	60	0	1	0	0	0	0
1	11	0	7	0	0	0	0
2	4	2	13	5	0	0	0
3	1	0	5	5	7	0	0
4	0	0	4	2	9	2	1
5	0	0	0	2	0	4	9
6	0	0	0	0	1	1	19

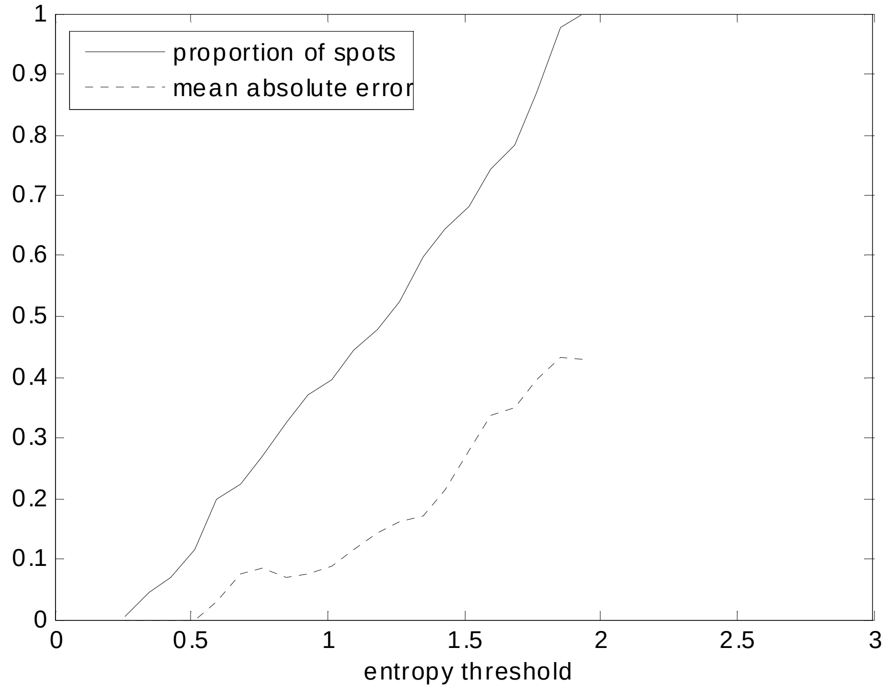
threshold, in the prediction of q_p and q_s values, respectively. Also shown is the mean absolute error computed over each fraction of spots.

As mentioned in Section 2.5, a trial was conducted in which a set of 686 tumour and normal spots were scored by the same pathologist on two separate sessions (this larger set included all 175 spots used in the previously described scoring experiments). This resulted in mean absolute disagreements of 0.300 and 0.175 and in linearly weighted Cohen's kappa coefficients of 0.892 and 0.866, for the assessments of q_p and q_s values, respectively. Figures 6.8(b) and (e) show the distributions of absolute disagreements between the two scoring sessions for q_p and q_s values, respectively, while Tables 6.5(a) and (b) show the corresponding contingency tables.

Finally, some experiments were carried out to assess the extent to which the adopted formalised quickscore values (that is, the global features characterising each spot) could be directly mapped into predicted quickscores, without relying on classification or ordinal regression. To this effect, those experiments that involved only one pair of global features $\mathbf{x} = (x_p, x_s)^{\mathrm{T}}$ per spot were repeated, replacing ordinal regression by direct mappings between x_p values and predicted q_p values and between x_s values and predicted q_s values. These mappings were



(a) q_p.



(b) q_s.

Figure 6.10: Fraction of scored spots and associated mean absolute error, for different entropy thresholds and for both quickscore values. From the experiments based on regions of interest and nuclear segmentation. (Lower entropy means higher confidence.)

Table 6.5: Contingency tables of the intra-observer scoring trial.

(a) q_p.

Session 1	Session 2							
	0	1	2	3	4	5	6	
0	278	2	4	0	1	0	0	285
1	7	7	16	2	1	0	0	33
2	3	2	32	3	3	0	0	43
3	0	0	16	8	7	3	0	34
4	2	0	1	4	27	15	8	57
5	0	0	1	5	14	32	25	77
6	0	0	0	0	1	17	139	157
	290	11	70	22	54	67	172	

(b) q_s.

Session 1	Session 2				
	0	1	2	3	
0	278	6	1	0	285
1	9	89	11	0	109
2	3	27	89	31	150
3	0	1	26	115	142
	290	123	127	146	

loosely based on the definition of quickscore (as given in Section 2.3) and are defined in Equations (6.15) and (6.16).

$$q_p = \begin{cases} 0 & \text{if } x_p \in [0.00,\ 0.01) \\ 1 & \text{if } x_p \in [0.01,\ 0.05) \\ 2 & \text{if } x_p \in [0.05,\ 0.20) \\ 3 & \text{if } x_p \in [0.20,\ 0.40) \\ 4 & \text{if } x_p \in [0.40,\ 0.60) \\ 5 & \text{if } x_p \in [0.60,\ 0.80) \\ 6 & \text{if } x_p \in [0.80,\ 1.00] \end{cases} \tag{6.15}$$

$$q_s = \begin{cases} 0 & \text{if } x_s \in [0.00,\ 0.25) \\ 1 & \text{if } x_s \in [0.25,\ 0.50) \\ 2 & \text{if } x_s \in [0.50,\ 0.75) \\ 3 & \text{if } x_s \in [0.75,\ 1.00] \end{cases} \tag{6.16}$$

The distributions of absolute errors resulting from the direct mapping experiments are shown in Figures 6.8(c) and (f), for the prediction of q_p and q_s values, respectively.

6.5 Discussion and conclusions

6.5.1 Pixel labelling

It is interesting to note that, in the simpler pixel labelling experiment that relied only on r, g, and b colour information, the class-conditional distributions estimated from annotated data reflected the fact that training pixel colours were lighter within the background, darker within unstained nuclei and even darker (closer to the origin of the colour space) within stained nuclei. This is visible in Figures 6.5(a), (b), and (c). Accordingly, the obtained class posteriors implied that the probability of the E_- class assumed higher values for lighter test pixel colours, whereas the probability of the E_+ class assumed higher values for darker test pixel colours. This is visible in Figures 6.5(e) and (f). The probability of the background class, however, assumed considerably high values for test pixels ranging from light to dark colours. This can be seen in Figure 6.5(d), where the region of high probabilities within the colour space extends roughly over a diagonal from black to white. In principle, this means that, for test pixels exhibiting certain intermediate colours, the system had difficulty in making a decision between the immunonegative and background labels.

The use of differential invariants in addition to colour resulted in a substantially higher number of pixels correctly labelled as belonging to epithelial nuclei (both E_- and E_+). The accompanying higher number of false positives was mostly due to under-segmentation of the nuclear regions (in the sense that the labelling technique did not yield segments corresponding to cleanly separated nuclei, but rather regions containing several nuclei merged together). However, for the purpose of computing formalised quickscore features, the benefit of having more correctly classified epithelial pixels outweighs the disadvantage of more epithelial false positives.

6.5.2 Comparison of ordinal regression with neural networks

Models trained to predict collapsed quickscore values consistently yielded better mean absolute errors than models trained to predict the same quickscores in non-collapsed format (collapsed only *a posteriori* for the purpose of comparison). This difference in quality of the results was also reflected in the confusion matrices. All matrices for the prediction of non-collapsed quickscores (q_p or q_s, regardless of the algorithm) showed one or two middle targets with zero predictions, but this effect was not observable in the matrices for the prediction of collapsed quickscores. This is illustrated in Tables 6.3(a) and (c) for non-collapsed and collapsed q_p values, respectively, and in Tables 6.3(b) and (d) for non-collapsed and collapsed q_s values, respectively. These results may indicate inadequacy of the global features used to characterise TMA spots, or a lack of training examples for middle targets. In the case of q_p predictions, there is also the possibility that the number of targets defined by the quickscore is itself excessive. In fact, from the contingency table shown in Table 6.5(a), it can be seen that the 0 and 6 q_p scores are those assigned most frequently by the pathologist, the intermediate scores being comparatively little used.

The GLM-based algorithm performed poorly. Besides yielding the highest mean absolute error in every experiment, the prediction of collapsed q_p values yielded a confusion matrix that showed a middle target with no predictions, something that did not happen with any other algorithm. This matrix is shown in Table 6.3(e).

Based solely on average absolute errors, ordinal regression using EP and Gaussian kernels could be said to be the best algorithm, as it yielded mean errors that were always either the lowest or very close to the lowest. However, as shown in Table 6.2, the large standard deviations associated with prediction errors seem to render a comparison between algorithms inconclusive. This is nevertheless an interesting result, given that the experiments carried out by Chu and Ghahramani [22] suggested that Gaussian process ordinal regression was capable of performing convincingly better than classification (in that instance based on support vector machines) on a variety of data sets. However, the result reported by Chu and Ghahramani for each individual experiment consisted of the mean absolute error over *all* test samples, coupled with a standard deviation over *partial mean* absolute errors. For each data set, a number of random partitions was defined and, for each partition, a partial mean error was computed. A standard deviation computed in this way has the disadvantage of depending on the partitioning of the data (that is, on the number of test samples per partition). In contrast, in the present work, both the mean and the standard deviation of the absolute error were computed over all the samples involved in each experiment.

In the prediction of q_p and q_s (non-collapsed) values, ordinal regression with EP and Gaussian kernels yielded mean absolute errors of 0.888 and 0.779, respectively, which are far from the intra-observer disagreements of 0.300 and 0.175. Nevertheless, it should be noted that the *inter*-observer disagreement would presumably be higher and therefore constitute a more realistic term of comparison.

The MLP-based algorithm performed surprisingly well, when compared with the methods that relied on ordinal regression. This suggests that further research is needed to improve the ordinal regression approach, given the expectation that formulating the tissue scoring problem as ordinal regression should represent an advantage over classification. Predictions made by the MLP also consumed a computational time per TMA spot that was at least one order of magnitude below that taken by ordinal regression (tenths of second versus several seconds).

As the entropy threshold set on predictions was decreased (that is, as the minimum confidence threshold was increased), the mean absolute error tended to decrease, as exemplified in Figures 6.7(a) and (b). This suggests that it would be possible to automatically process, with quite low mean errors, reasonable fractions of spots that are more unequivocal, while identifying the more difficult spots that cannot dispense with human assessment (in a similar way to the MLP-based approach used to classify spots into types, reported in Chapter 4).

6.5.3 Comparison of global feature extraction methods

As shown in Figures 6.8(a) and (d), in relation to the use of local patches without ROIs, the combined use of nuclear segmentation and ROIs led to large reductions in the mean absolute errors associated with the predictions of both quickscore values, more specifically a reduction of 0.430 (48%) in q_p predictions and of 0.450 (51%) in q_s predictions.

The isolated effect of replacing the local patch-based method with nuclear segmentation was considerable. Without ROIs, error reductions of 0.126 and 0.337 were achieved for q_p and q_s predictions, respectively; with ROIs, these reductions were 0.280 and 0.360, respectively. This appears to stem from the fact that the local patch-based method tended to incorrectly label many E_- pixels as E_+, whereas the segmentation-based method was more successful at identifying E_- nuclei. This is illustrated in Figures 6.6(a), (b), and (c), where a string of E_- nuclei (forming a C-shaped region across the centre of the image) had most of their pixels incorrectly labelled as E_+ but were correctly segmented and classified as E_-. These results suggests the need to improve the quality of local patch-based posteriors, if segmentation is to be avoided in the computation of formalised quickscore features. Alternatively, a truly automated method for accurate segmentation of epithelial nuclei should be adopted (avoiding reliance on heuristics and free parameters in need of manual tuning).

The isolated effect of introducing ROIs was considerable, too. Using local patches, error reductions of 0.149 and 0.085 were obtained for q_p and q_s predictions, respectively; with nuclear segmentation, these reductions were 0.303 and 0.108, respectively. It is worth pointing out that, when assessing the quickscores of certain heterogeneous spots that contain *both* regions of normal epithelial tissue and regions of tumour, pathologists ignore the normal tissue and focus their attention only on the tumour portions. This constitutes a somewhat liberal interpretation of the definition of quickscore, which (as seen in Section 2.3) does not actually involve a distinction between normal and tumour epithelial cells. Given that the methods used in the computation of global features, too, did not make such a distinction, this specific aspect of scoring could explain to a certain extent the impact of ROI usage on prediction performance. Figure 6.9(e) showed an example spot containing a partial ring of normal epithelial nuclei on the top-left region and a large portion of tumour tissue on the bottom-right region. Figure 6.9(f) showed the corresponding manually drawn ROI, which excluded the region of normal tissue.

In practice, however, heterogeneous spots containing both tumour and normal tissue are rare, which suggests that the effect of using manually drawn ROIs was essentially that of masking the relatively poor efficacy of both methods employed in the computation of global features. In fact, when using local patches, many pixels belonging to non-epithelial nuclei (such as stromal and inflammatory) were erroneously labelled as epithelial. This is illustrated with the spot shown in Figure 6.9(a), whose top-left region contains a concentration of inflammatory cells. In Figure 6.9(b), it can be seen how the patch-based approach classified many of the pixels in these cells as E_+, which led to over-estimated formalised q_p and q_s values. This mislabelling of pixels was probably due to the fact that the training set of manually annotated nuclei contained a certain proportion of nuclei that were neither fully non-stained nor fully stained. In turn, even though the segmentation-based approach, too, wrongly detected the inflammatory cells in this spot as epithelial nuclei, they were assigned to the E_- class, thus having a less harmful effect on the computed formalised quickscore and allowing a correct prediction.

Within the spot shown in Figure 6.9(c), the segmentation-based method misidentified many of the stromal cells in the region of connective tissue (that is, the two left thirds of the spot) as E_- nuclei. This led to under-estimated formalised q_p and q_s values. When the manually annotated ROI shown in Figure 6.9(d) was used, the stromal cells were excluded from the computations, which allowed a correct prediction of the spot's quickscore.

From Figure 6.10(a), it can be seen that, below an entropy threshold of about 1.75, over 60% of the spots could still be scored as to q_p with a mean absolute error equal to the mean intra-observer disagreement of 0.30. This result (which relied on nuclear segmentation and the use of ROIs) suggests that an improved scoring procedure could ultimately be used to automatically process large proportions of spots with average errors comparable to the variability inherent to a human observer.

Figures 6.8(c) and (f) show that the direct mapping of formalised quickscores into predicted scores yielded a result similar (and actually slightly better) to that obtained via ordinal regression only in the case of q_p predictions based on nuclear segmentation without ROIs. Especially in the case of q_s predictions, it can be seen that direct mapping is inadequate. This indicates the inexistence of a linear relationship between the computed global features and the spots' scores.

***

This chapter addressed the prediction of quickscores of breast TMA spots subjected to progesterone receptor immunohistochemistry. The techniques used to extract local and global features from TMA spot images were specified. An overview of the technique of Gaussian processes for ordinal regression was given. A description of the carried out experiments was given and their results reported. These results were discussed and some conclusions presented.

The next chapter will conclude this thesis, summarising the contributions of the reported work and providing an overview of the main conclusions drawn from it. Possible future directions of work will also be discussed.

Chapter 7

Conclusions and future directions

This chapter provides, in Section 7.1, a summary of the contributions of the work reported in this thesis. Section 7.2 discusses the main conclusions drawn from the carried out experiments, whereas section 7.3 suggests potential future directions of work. Both these sections are divided into subsections associated with the classification of spots into types, their scoring, and their segmentation into regions of different types. At the end of Section 7.3, an additional subsection addresses possible future work of a more generic nature.

7.1 Summary of contributions

The following bullet points list the main contributions of this work.

- An approach was developed to classify breast tissue microarray spots into their four main types, namely tumour, normal, stroma and fat, with the purpose of identifying tumour and normal spots that needed to be subsequently scored, while discarding spots containing only stroma and fatty tissue.

- The developed method was applied to spots subjected to progesterone receptor nuclear immunostaining.

- The implemented method was based on the technique introduced by Varma and Zisserman [98], so that a histogram of texton frequencies was computed for each spot and a classifier was trained to classify spots based on their texton histograms.

- The classification performance of a multi-layer perceptron (MLP) was compared with that of simpler classifiers.
- The performance of the MLP was also compared with that of a classifier based on latent Dirichlet allocation (LDAL) models.
- By associating distinct types of tissue with latent variables of the LDAL model, a method for the segmentation of TMA spots into regions of different types was explored.
- An approach was developed to predict the quickscores of tumour and normal breast tissue microarray spots subjected to progesterone receptor nuclear immunostaining.
- Both quickscore integer values were predicted, to reflect the proportion of epithelial nuclei that were stained as well as the strength of their staining.
- The developed method was based on the hypothesis that the prediction of scores would not need to rely on an accurate segmentation technique. The basis for the computation of global features was the labelling of pixels as to the probability of their belonging to each of three classes, namely background, immunopositive nucleus, and immunonegative nucleus. From this labelling, features formalising the quickscore values used by pathologists were computed.
- In the prediction of quickscores of spots based on their global features, the performance of Gaussian processes for ordinal regression was compared with that of MLP classifiers.
- Different methods of global feature extraction were compared in terms of their impact on the scoring performance.
- The posterior probabilities output by the MLP classifier (for classification) and by the ordinal regression algorithm (for scoring) were used to compute a simple measure of prediction confidence. This allowed to set confidence thresholds that helped to distinguish the “easier” spots that could be processed automatically with high confidence from the more “difficult” spots that should be referred for manual assessment.

7.2 Conclusions

7.2.1 Classification of spots

In the classification of breast tissue microarray spots into four types, the multilayer perceptron (MLP) performed better than the nearest-neighbour classifier and the generalised linear model (GLM). The highest accuracy achieved was 74.6±0.9%.

This result may be compared with an intra-observer agreement of 94.0%, while keeping in mind that the *inter*-observer agreement would in principle be lower and therefore constitute a more realistic criterion for comparison.

By setting classification confidence thresholds, higher accuracies were achieved for fractions of the data. Thus, the proposed system could be used to reduce the workload of pathologists, classifying the more unequivocal spots and identifying the more difficult spots in need of manual assessment.

The performance of a classifier based on latent Dirichlet allocation (LDAL) models was comparable to that of the MLP. This was an interesting result, given that LDAL is a generative model, unlike the MLP. The developed approach, however, relied on separate LDAL models (one for each spot type) whose latent topics were not shared, and therefore a probabilistic output similar to that of the MLP was not available.

The analysis of the class posterior probabilities output by the MLP for certain misclassified spots raised a number of questions. Tumour spots whose tumour epithelial nuclei were very scattered tended not to be classified as tumour spots. This suggests that the employed local features may have been too localised to capture the high-level texture of scattered nuclei.

On the other hand, large regions of non-epithelial cells (such as inflammatory cells) seem to have caused spots containing only stroma to be misclassified as either tumour or normal. This suggests that the used local features did not capture differences between the internal textures of different types of nuclei.

In some cases, the slight scattering of normal epithelial nuclei appears to have been sufficient to cause the misclassification of normal spots as tumour spots. This probably reflects the fact that local texture features are incapable of capturing the morphology of tissue structures, such as ring-like arrangements of normal epithelial nuclei that form the walls of tubules.

When the area occupied by tumour or normal epithelial tissue was very small, spots tended to be misclassified as either stroma or fat. This may indicate a lack of training examples for spots with such characteristics.

7.2.2 Scoring of spots

The use of differential invariant features in addition to colour improved the pixel labelling that formed the basis of global feature extraction in scoring experiments, in the sense that substantially more pixels were correctly classified as belonging to epithelial nuclei. However, the quality of the results is still far from ideal, with large numbers of pixels being misclassified as epithelial, generally contributing to under-segmented nuclear regions.

In the prediction of quickscores, classification with GLMs and MLPs was compared against Gaussian process ordinal regression (the latter having been tested with two different types of kernel and two distinct hyper-parameter learning techniques). However, the large standard deviations associated with the mean absolute errors obtained for the different algorithms rendered their comparison inconclusive. This result was interesting, because it indicates that, when standard deviations are computed from the absolute errors of all individual predictions, the performance of ordinal regression can be less impressive than the work of Chu and Ghahramani [22] suggested. On the other hand, the inconclusive nature of the comparison was disappointing, given the expectation that formulating the tissue scoring problem as ordinal regression (thus incorporating into the model the existence of an order between targets) should represent an advantage over classification.

In particular, the MLP performed surprisingly well (besides taking considerably less time to train than a Gaussian process for ordinal regression). Nevertheless, ordinal regression using a Gaussian kernel and the expectation maximisation learning technique yielded the best mean absolute error, of 0.888, in the prediction of q_p values (which reflect the proportion of epithelial nuclei that are stained). That same technique yielded an error of 0.779, very close to the best, in the prediction of q_s values (which reflect the strength of staining).

These results may be compared with intra-observer disagreements of 0.300 and 0.175, in the assessment of q_p and q_s values, respectively. It should be kept in mind, however, that *inter*-observer disagreements would in principle be larger and therefore constitute a more realistic term of comparison.

The direct mapping of global features (that is, formalised quickscores) into predicted scores yielded considerably worse results than prediction through ordinal regression. This indicates the absence of a linear relationship between the computed global features and the spots' scores.

By taking advantage of the probabilistic output of ordinal regression and setting scoring confidence thresholds, lower mean absolute errors were achieved for fractions of the data. Thus, as with the MLP-based classification of spots, the proposed scoring system could be used to reduce the pathologists' workload, scoring the more unequivocal spots and identifying the more difficult spots in need of manual assessment.

The prediction of collapsed quickscores was consistently better than the prediction of non-collapsed quickscores. Presumably, collapsed quickscores are of no interest to pathologists, but this result nevertheless suggests that the used training data may have been insufficient to adequately represent the whole range of possible quickscore values. In fact, by inspecting the available expert annotations, it was observed that the two quickscore integer ranges (q_p values between 0 and 6, and q_s values between 0 and 3) are used by pathologists in a way that is far from uniform, especially in the case of q_p values.

The replacement of the global feature extraction method based on pixel labelling with a method based on nuclear segmentation (although not fully automated, in the sense that many free parameters were manually tuned to achieve good performance with the available data), together with the use of manually drawn regions of interest, yielded large reductions in the observed mean absolute errors: 0.430 (48%) in q_p predictions and 0.450 (51%) in q_s predictions. By setting an appropriate scoring confidence threshold, the system thus modified was capable, for example, of predicting the q_p scores of more than 60% of the available spots with a mean absolute error equal to the mean intra-observer disagreement of 0.300. These results provide a notion of how much the performance of the developed scoring approach could be improved, should the pixel labelling technique be enhanced to yield results comparable to those of a fully automated and accurate nuclear segmentation method.

In fact, it could be observed that the pixel labelling technique tended to mislabel many of the pixels that belonged to immunonegative nuclei as belonging to immunopositive nuclei. In addition, pixels belonging to non-epithelial nuclei (such as stromal and inflammatory) were often mislabelled as epithelial. It is interesting to note that the misperception of non-epithelial regions as epithelial appeared to be an issue in the classification of spots, too. This suggests that the problem may be largely due to inadequacy of the used local features, given that, even though the spot classification method did not rely on pixel labelling, it nevertheless employed the same local features as the scoring approach.

7.2.3 Segmentation of tissue regions

The segmentation methods based on LDAL models performed significantly better than a simple method based on texton frequencies directly learned from annotated data. The best agreement between segmentations and manual annotations, of 0.695, was achieved by the method based on LDAL with inferred parameters β.

In relation to the method based on fixed β, the inference of β also permitted considerable increments in the recalls for tumour and normal areas. In principle, this corresponded to a better ability of the system not to miss epithelial regions, which are of particularly importance for the pathologist (even if at the expense of lower precision, manifested as under-segmentation).

At any rate, the quality of the segmentation results obtained via LDAL was far from ideal. In spots with low staining strength, tumour regions were often wrongly detected as normal. The segmentation of regions of normal tissue often exhibited exaggerate dilation. In spots containing large regions of stroma, epithelial regions tended to be wrongly interpreted as stroma.

It is interesting to note that the qualitative segmentation results obtained with an existing commercial tool (the Genie module from Aperio, Inc.), although better than the results achieved in the present work, were not particularly impressive. Specifically, the tool had considerable difficulty in segmenting regions of immunonegative (unstained) tumour, and dealt ambiguously with immunopositive normal regions.

7.3 Future directions

7.3.1 Classification of spots

It would be worth adding at least one level to the Gaussian pyramid used in the extraction of local features (thus taking into account a larger scale), in an attempt to capture the high-level texture of regions containing very scattered nuclei. On the other hand, higher-order differential invariant features should be tried, in order to capture more detail on the internal texture of different types of nuclei.

It would also be interesting to experiment with different types of local features, including the “Gabor-like” filter bank proposed by Schmid [89] and distribution-based descriptors such as intensity domain spin images and the Rotation-Invariant Feature Transform (RIFT) proposed by Lazebnik et al. [57]. In image retrieval and classification tasks, “Gabor-like” filters have performed better than differential invariants, and spin images have performed better than “Gabor-like” filters. The combined use of spin images and RIFT has resulted in an additional improvement in performance.

Texton histograms were used as global features for classification, because it was felt that they would be capable of characterising the textural content of the images better than a set of global statistical features computed directly from the results of local filtering. There is evidence, however, that frequency-based histograms tend to yield better classification results for textures that are statistically stationary, whereas spatial statistics of texture elements tend to be more effective when applied to textures that are non-regularly and sparsely distributed [27]. Given that the textural content of tissue microarray spots falls into the latter category, it would be very interesting to compare the performance of global statistical features (such as the popular Haralick features [42]) with the results obtained in this work.

The model of tissue section underlying the classification of spots should incorporate not only texture information provided by local features, but also morphological information capable of reflecting different arrangements of nuclei. This could be achieved, for example, by modelling tissue sections as graphs whose nodes are the centres of pre-detected nuclei and computing summary features that characterise those graphs, as in the work of Rodenacker and Bischoff [85] and Geusebroek et al. [36]. When detected nuclei are reduced to graph nodes, however, much information about the context surrounding each nucleus is lost. Therefore, it would also be interesting to explore techniques such as those proposed by Ren et al. [84] and Heitz and Koller [45], which involve the modelling of spatial context.

The spot classification method based on LDAL should be improved to feature not a separate model per class but one single model sharing latent topics across classes, as in the work reported by Fei-Fei and Perona [34]. Not only this approach would be likely to yield better results, it would also provide a probabilistic output that could be used to estimate the confidence associated with each prediction. In addition, the use of hierarchical Dirichlet processes (HDPs) [95] would allow to automatically determine the optimal number of latent topics.

7.3.2 Scoring of spots

The pixel labelling method used in the extraction of global features should be either greatly improved, or replaced with a fully automated technique for accurate segmentation of nuclei and their classification (as either immunopositive or immunonegative).

It is expected that the use of better local texture features (such as those previously discussed in Section 7.3.1) would improve the pixel labelling performance. Further improvement might be achieved by replacing the simple labelling method based on Bayes' rule with MLP-based classification, possibly incorporating automatic relevance determination into the model's estimation (so that certain local features could play a more relevant role than others).

In addition, it would be worth testing the results of pixel labelling using as colour features not the r, g, and b components, directly, but rather their mapping into a more perceptually uniform colour space, such as an *Lab* colour space. Given that, for the purpose of learning the densities of features for each class, the colour feature space is divided into bins (in this work, 16^3 bins), it is reasonable to assume that, by using a perceptually uniform colour space, pixels of perceptually close colours would be more likely to contribute to bins that are close together. This could improve the quality of the estimated densities.

The fact that ordinal regression incorporates knowledge about the order between targets was not reflected in terms of improved performance, in relation to MLP-based classification. This suggests that further research may be needed, to take full advantage of the ordinal regression model. A possibility would be to investigate modifications that allowed to accurately model the way in which pathologists mislabel the ground-truth, based on observer variability data. Even though the used implementation of Gaussian process ordinal regression already models the existence of noisy data, this is done through a single σ_{noise}^2 noise variance hyper-parameter. A set of variability parameters on which the user could set a prior might be more appropriate.

The size of the data set of tumour and normal spots used in scoring experiments should be increased, to ensure that the whole ranges of q_p and q_s quickscore values are adequately represented in the training data.

7.3.3 Segmentation of tissue regions

It would be interesting to develop Bayesian hierarchical models more sophisticated and better suited to the problem of segmenting TMA spots than the LDAL models used in the present work. In LDAL models featuring a single layer of latent topics, the optimal number of topics tends to be considerably larger than the range of tissue types that may occur within a TMA spot. The addition of a second layer of latent topics (modelled as distributions over topics of the first layer) might therefore help to successfully represent different tissue types. The inclusion of additional layers of topics could also be used to model the relationship between epithelial regions and the immunopositivity of nuclei, in that both tumour and normal regions may contain both immunonegative and immunopositive nuclei. In fact, a system capable of segmenting regions of tumour, normal tissue, fat, and stroma, further segmenting tumour and normal regions into stained and unstained subregions, could in principle be used as a basis for both classification of TMA spots into types and their scoring.

7.3.4 Overall system

In the present work, the type (or score) of a given spot was predicted simply by choosing the target associated with the highest probability, as output by the classification (or regression) algorithm. This relied on the implicit assumption that all errors had the same cost, which can hardly be realistic (for example, the cost of mistaking a tumour spot for normal is certainly higher than that of mistaking a normal spot for tumour). It would be useful to estimate more sensible values for the various errors that may occur when classifying and scoring TMA spots, based on expert knowledge and a cost-benefit analysis.

The carried out classification and scoring experiments did not take into account the provenance of the involved TMA spots (that is, which patients they originated from). Thus, at each pass of a leave-one-out experiment, the training stage may have often involved a small number of spots that originated from the same patient as the spot being tested. It is expected that this did not affect the obtained results significantly, as the used spots were randomly selected from TMA slides that originated from a relatively large number of patients (specifically, 112 patients). In addition, there is large variability even in the appearance of spots associated with the same patient. Nevertheless, it would be advisable to design new experiments taking into account the provenance of spots, for example to test all the spots associated with a given patient, after a training stage based on those spots associated with all the remaining patients.

The available data on observer variability was very limited, being based on the assessment of a collection of TMA spots on glass (that is, by observing physical TMA slides under the microscope) by one pathologist on two separate occasions. Thorougher trials should be conducted, involving more than one pathologist analysing the same spots more than once, both on glass and on screen (that is, from digitised images of the slides), so that better estimates may be obtained for the intra- and inter-observer variabilities associated with the classification and scoring of spots.

It would be equally important to determine whether there is significant variability between images of TMA spots acquired with the same digital scanner and with different scanners, for example due to changes in lighting conditions.

The developed methods should be tested on TMA spots subjected to other forms of nuclear immunostaining, besides progesterone receptor (PR). Alternative stains should include oestrogen receptor (ER, considered the most useful stain from the point of view of diagnosis and survival analysis) and tumour protein 53 (p53). It is worth noting that, according to pathologists, the PR stain (used in this work) is “dirtier” and less “crisp” than ER and p53. This presumably means that, from the point of view of automated analysis, PR is a less friendly nuclear stain than ER and p53.

Once the developed system achieves suitable performance on TMA spots subjected to nuclear immunostaining, it should be extended so as to be capable of dealing with the immunostaining of other sub-cellular compartments, such as membrane staining and cytoplasmic staining (for which appropriate scoring systems are available). These stains are expected to be more difficult to analyse than nuclear stains.

It would be important to investigate the extent to which automation of TMA assessment may contribute to survival analysis and discovery experiments. It is possible that information other than the types and scores of spots could be extracted from image data and applied with benefit to survival analysis. In addition, traditional survival analysis methods (such as Cox’s proportional hazard model estimation) are known to suffer from over-fitting and instability. Therefore, it would be worth experimenting with less conventional approaches, such as methods based on neural networks and also Bayesian methods.

Bibliography

[1] Adjuvant Breast Cancer Trials Collaborative Group. Polychemotherapy for early breast cancer: Results from the international Adjuvant Breast Cancer chemotherapy randomized trial. *Journal of the National Cancer Institute*, 99(7):506–515, 2007. 30

[2] D. Allred, J. Harvey, M. Berardo, and G. Clark. Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Mod. Pathol.*, 11(2):155–168, 1998. 26

[3] Aperio Technologies, Inc. Genie Histology Pattern Recognition. `http://www.aperio.com/imageanalysis/Genie-Histology.asp`. Checked on April 9, 2010. 72

[4] M. Arif and N. Rajpoot. Detection of nuclei by unsupervised manifold learning. In *Medical Image Understanding and Analysis*, pages 96–100, 2007. 34, 35

[5] A. Barnes, S. Pinder, J. Bell, E. Paish, P. Wencyk, J. Robertson, C. Elston, and I. Ellis. Expression of p27kip1 in breast cancer and its prognostic significance. *The Journal of Pathology*, 201(3):451–459, 2003. 26

[6] G. Begelman, E. Gur, E. Rivlin, M. Rudzsky, and Z. Zalevsky. Cell nuclei segmentation using fuzzy logic engine. In *IEEE International Conference on Image Processing*, volume 5, pages 2937–2940, 2004. 36

[7] J. Bejar, E. Sabo, S. Eldar, M. Lev, I. Misselevich, and J. Boss. The prognostic significance of the semiquantitatively determined estrogen receptor content of breast carcinomas: a clinicopathological study. *Pathology-Research and Practice*, 198(7):455–460, 2002. 26

[8] A. Berger, D. Davis, C. Tellez, V. Prieto, J. Gershenwald, M. Johnson, D. Rimm, and M. Bar-Eli. Automated quantitative analysis of activator protein-2 {alpha} subcellular expression in melanoma tissue microarrays

correlates with survival prediction. *Cancer Research*, 65(23):11185, 2005. 46

[9] C. Bilgin, P. Bullough, G. Plopper, and B. Yener. ECM-aware cell-graph mining for bone tissue modeling and classification. *Data Mining and Knowledge Discovery*, pages 1–23, 2008. 43

[10] C. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., 2006. ISBN 0387310738. 4, 51, 54, 56

[11] D. Blei. Topic Modeling. http://www.cs.princeton.edu/~blei/topicmodeling.html. Checked on March 23, 2010. 63, 74

[12] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. 4, 51, 58, 60, 61, 63, 74

[13] H. Bloom and W. Richardson. Histological grading and prognosis in breast cancer. *Br. J. Cancer*, 11(3):359–77, 1957. 27

[14] Breastcancer.org. What Is Breast Cancer? http://www.breastcancer.org/symptoms/understand_bc/what_is_bc.jsp. Checked on September 30, 2009. 23

[15] A. Brook, R. El-Yaniv, E. Isler, R. Kimmel, R. Meir, and D. Peleg. Breast cancer diagnosis from biopsy images using generic features and SVMs. Technical report, Technion - Israel Institute of Technology, 2006. 42

[16] P. Burt and E. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31:532–540, 1983. 52, 53

[17] J. Caicedo, A. Cruz, and F. González. Histopathology image classification using bag of features and kernel functions. *Artificial Intelligence in Medicine*, pages 126–135, 2009. 19

[18] R. Camp, L. Charette, and D. Rimm. Validation of tissue microarray technology in breast carcinoma. *Laboratory Investigation*, 80(12): 1943–1949, 2000. 29

[19] Robert Camp, Gina Chung, and David Rimm. Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nature Medicine*, 8(11):1323–1327, Nov 2002. 45

[20] Cancer Research UK. UK Breast cancer incidence statistics. `http://info.cancerresearchuk.org/cancerstats/types/breast/incidence/`. Checked on September 30, 2009. 23, 24

[21] W. Chu. Gaussian Processes for Ordinal Regression v1.0. `http://www.gatsby.ucl.ac.uk/~chuwei/README.gpor`. Checked on April 1, 2010. 95

[22] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005. 89, 95, 112, 119

[23] J. Coligan, A. Kruisbeek, D. Margulies, E. Shevach, and W. Strober, editors. *Current Protocols in Immunology.* John Wiley & Sons, Inc., 2003. 4, 25

[24] C. Conway, L. Dobson, A. O'Grady, E. Kay, S. Costello, and D. O'Shea. Virtual microscopy as an enabler of automated / quantitative assessment of protein expression in TMAs. *Histochemistry and Cell Biology*, 130(3): 447–463, 2008. 45

[25] A. Coons, H. Creech, and R. Jones. Immunological properties of an antibody containing a fluorescent group. *Proc. Soc. Exp. Biol. Med.*, 47: 200–202, 1941. 24

[26] M. Cregger, A. Berger, and D. Rimm. Immunohistochemistry and quantitative analysis of protein expression. *Archives of Pathology & Laboratory Medicine*, 130(7):1026–1030, 2006. 45

[27] G. Dahme, E. Ribeiro, and M. Bush. Spatial statistics of textons. In *International Conference on Computer Vision Theory and Applications*, pages 13–19, 2006. 122

[28] J. Dalle, W. Leow, D. Racoceanu, A. Tutac, and T. Putti. Automatic breast cancer grading of histopathological images. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3052–3055, 2008. 36, 37, 38, 39, 40, 41, 46

[29] C. Demir and B. Yener. Automated cancer diagnosis based on histopathological images: a systematic survey. Technical report, Rensselaer Polytechnic Institute, Department Of Computer Science, 2005. 33

[30] S. Detre, G. Saccani Jotti, and M. Dowsett. A "quickscore" method for immunohistochemical semiquantitation: validation for oestrogen receptor in breast carcinomas. *Journal of Clinical Pathology*, 48(9):876–878, 1995. 18, 19, 26, 31, 47

[31] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski. Automated grading of prostate cancer using architectural and textural image features. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1284–1287, 2007. 41

[32] N. Elie, B Plancoulaine, J. Signolle, and P. Herlin. A simple way of quantifying immunostained cell nuclei on the whole histologic section. *Cytometry Part A*, 56A, Issue 1:37–45, 2003. 39, 40

[33] A. Esgiar, R. Naguib, B. Sharif, M. Bennett, and A. Murray. Fractal analysis in the detection of colonic cancer images. *IEEE Transactions on Information Technology in Biomedicine*, 6(1):54–58, 2002. 43

[34] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 524–531, 2005. 122

[35] R. Fernandez-Gonzalez and C. de Solorzano. A tool for the quantitative spatial analysis of mammary gland epithelium. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1549–1552, 2004. 38

[36] J. Geusebroek, A. Smeulders, F. Cornelissen, and H. Geerts. Segmentation of tissue architecture by distance graph matching. *Cytometry Part A*, 35(1):11–22, 1999. 122

[37] J. Giltnane, J. Murren, D. Rimm, and B. King. AQUA and FISH analysis of HER-2/neu expression and amplification in a small cell lung carcinoma tissue microarray. *Histopathology*, 49(2):161–169, 2006. 46

[38] D. Glotsos, P. Spyridonos, P. Petalas, D. Cavouras, P. Ravazoula, P. Dadioti, I. Lekka, and G. Nikiforidis. Computer-based malignancy grading of astrocytomas employing a support vector machine classifier, the WHO grading system and the regular hematoxylin-eosin diagnostic staining procedure. *Analytical and Quantitative Cytology and Histology*, 26(2):77–83, 2004. 36, 42

[39] C. Gunduz, B. Yener, and S. Gultekin. The cell graphs of cancer. *Bioinformatics*, 20(Suppl 1):i145, 2004. 43

[40] M. Gurcan, T. Pan, H. Shimada, and J. Saltz. Image analysis for neuroblastoma classification: segmentation of cell nuclei. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, page 4844, 2006. 35

[41] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener. Histopathological image analysis: a review. *IEEE Reviews In Biomedical Engineering*, 2:147–171, 2009. 34

[42] R. Haralick and L. Shapiro. *Computer and Robot Vision*. Addison-Wesley Publishing Company, 1992. 51, 122

[43] M. Harigopal, J. Heymann, S. Ghosh, V. Anagnostou, R. Camp, and D. Rimm. Estrogen receptor co-activator (AIB1) protein expression by automated quantitative analysis (AQUA) in a breast cancer tissue microarray and association with patient outcome. *Breast Cancer Research and Treatment*, 115(1):77–85, 2009. 46

[44] M. Hayat. *Microscopy, immunohistochemistry, and antigen retrieval methods: for light and electron microscopy*. Plenum Pub. Corp., 2002. 24, 25, 26

[45] G. Heitz and D. Koller. Learning spatial context: using stuff to find things. In *European Conference on Computer Vision*, page 30, 2008. 122

[46] K. Jafari-Khouzani and H. Soltanian-Zadeh. Multiwavelet grading of pathological images of prostate. *IEEE Transactions on Biomedical Engineering*, 50(6):697–704, 2003. 41

[47] T. Jones, A. Carpenter, and P. Golland. Voronoi-based segmentation of cells on image manifolds. In *International Workshop on Computer Vision for Biomedical Image Applications*, volume 3765, pages 535–543, 2005. 35

[48] B. Karaçali and A. Tözeren. Automated detection of regions of interest for tissue microarray experiments: an image texture analysis. *BMC Medical Imaging*, 7(1):2, 2007. 38, 42

[49] R. Katz, S. Patel, N. Sneige, H. Fritsche, G. Hortobagyi, F. Ames, T. Brooks, and N. Ordonez. Comparison of immunocytochemical and biochemical assays for estrogen receptor in fine needle aspirates and

histologic sections from breast carcinomas. *Breast Cancer Research and Treatment*, 15(3):191–203, 1990. 26

[50] S. Keenan, J. Diamond, W. McCluggage, H. Bharucha, D. Thompson, P. Bartels, and P. Hamilton. An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN). *J. Pathol.*, 192:351–362, 2000. 42

[51] J. Koenderink and A. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55(6):367–375, 1987. 49

[52] M. Komosinski and K. Krawiec. Evolutionary weighting of image features for diagnosing of CNS tumors. *Artificial Intelligence in Medicine*, 19(1): 25–38, 2000. 42

[53] J. Kong, H. Shimada, K. Boyer, J. Saltz, and M. Gurcan. Image analysis for automated assessment of grade of neuroblastic differentiation. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 61–64, 2007. 39

[54] J. Kononen, L. Bubendorf, A. Kallionimeni, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. Mihatsch, G. Sauter, and O. Kallionimeni. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, 4(7):844–847, 1998. 27

[55] S. Kostopoulos, D. Cavouras, A. Daskalakis, P. Bougioukos, P. Georgiadis, G. Kagadis, I. Kalatzis, P. Ravazoula, and G. Nikiforidis. Colour-texture based image analysis method for assessing the hormone receptors status in breast tissue sections. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4985–4988, 2007. 40

[56] N. Lawrence. Gaussian Process Software. `http://www.cs.man.ac.uk/~neill/gp/`. Checked on March 26, 2010. 5, 90

[57] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005. 122

[58] K. Lee and W. Street. Model-based detection, segmentation, and classification for image analysis using online shape learning. *Machine Vision and Applications*, 13(4):222–233, 2001. 37

[59] K. Lee and W. Street. An adaptive resource-allocating network for automated detection, segmentation, and classification of breast cancer nuclei topic area: image processing and recognition. *IEEE Transactions on Neural Networks*, 14, Issue 3:680–.687, 2003. 37

[60] C. Loukas and A. Linney. A survey on histological image analysis-based assessment of three major biological factors influencing radiotherapy: proliferation, hypoxia and vasculature. *Computer Methods and Programs in Biomedicine*, 74(3):183–199, 2004. 33

[61] D. MacKay. *Bayesian methods for backpropagation networks*, pages 211–254. Springer-Verlag, 1994. 58, 94

[62] K. Masood and N. Rajpoot. Texture based classification of hyperspectral colon biopsy samples using CBLP. In *International Symposium on Biomedical Imaging*, 2009. 43

[63] A. Materka and M. Strzelecki. Texture analysis methods - a review. Technical report, Technical University of Lodz, Institute of Electronics, 1998. 33

[64] K. McCarty, Jr, L. Miller, E. Cox, J. Konrath, and K. McCarty, Sr. Estrogen receptor analyses. correlation of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies. *Arch. Pathol. Lab. Med.*, 109:716–721, 1985. 26

[65] D. McKee and W. Land, Jr. An adaptive image segmentation process for the classification of lung biopsy images. In *SPIE Medical Imaging*, volume 6144, pages 1628–1637, 2006. 34, 35, 36

[66] T. Mouroutis, S. Roberts, and A. Bharath. Robust cell nuclei segmentation using statistical modelling. *Bioimaging*, 6(2):79–91, 1998. 35

[67] I. Nabney. *NETLAB: algorithms for pattern recognition.* Springer-Verlag, 2002. ISBN 1-85233-440-1. 54, 63, 95

[68] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 284–287, 2008. 38

[69] R. Neal. *Bayesian learning for neural networks.* Number 118. Springer-Verlag, 1996. 58, 94

[70] OpenWetWare. Griffin: Antigen Retrieval Technique. `http://openwetware.org/index.php?title=Griffin:Antigen_Retrieval_Technique&oldid=376340`. Checked on March 2, 2010. 25

[71] S. Perkins, K. Edlund, D. Esch-Mosher, D. Eads, N. Harvey, and S. Brumby. Genie Pro: robust image classification using shape, texture and spectral information. In *SPIE Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery*, volume 5806, pages 139–148, 2005. 72

[72] S. Petushi, C. Katsinis, C. Coward, A Tözeren, and F. Garcia. Automated identification of microstructures on histology slides. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 424–427, 2004. 36, 46

[73] S. Petushi, F. Garcia, M. Haber, C. Katsinis, and A. Tozeren. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC Medical Imaging*, 6:14, 2006. 37, 38, 41, 46

[74] A. Psyrri, M. Kassar, Z. Yu, A. Bamias, P. Weinberger, S. Markakis, D. Kowalski, R. Camp, D. Rimm, and M. Dimopoulos. Effect of epidermal growth factor receptor expression level on survival in patients with epithelial ovarian cancer. *Clinical Cancer Research*, 11(24):8637, 2005. 46

[75] A. Psyrri, Z. Yu, P. Weinberger, C. Sasaki, B. Haffty, R. Camp, D. Rimm, and B. Burtness. Quantitative determination of nuclear and cytoplasmic epidermal growth factor receptor expression in oropharyngeal squamous cell cancer by using automated quantitative analysis. *Clinical Cancer Research*, 11(16):5856, 2005. 46

[76] H. Qureshi. *Meningioma classification using an adaptive discriminant wavelet packet transform.* PhD thesis, University of Warwick, 2009. 44

[77] H. Qureshi and N. Rajpoot. Comparative analysis of spatial and transform domain methods for meningioma subtype classification. In *Medical Image Understanding and Analysis*, pages 209–213, 2010. 44

[78] N. Rajpoot, M. Arif, and A. Bhalerao. Unsupervised learning of shape manifolds. In *British Machine Vision Conference*, 2007. 35

[79] E. Rakha, M. El-Sayed, A. Lee, C. Elston, M. Grainge, Z. Hodi, R. Blamey, and I. Ellis. Prognostic significance of Nottingham histologic

grade in invasive breast carcinoma. *J. Clin. Oncol.*, 26(19):3153–3158, 2008. 27

[80] J. Ramos-Vara. Technical aspects of immunohistochemistry. *Veterinary Pathology Online*, 42(4):405–426, 2005. 24, 25

[81] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. 89

[82] A. Reiner, G. Reiner, J. Spona, B. Teleky, R. Kolb, and J. Holzner. Estrogen receptor immunocytochemistry for preoperative determination of estrogen receptor status on fine-needle aspirates of breast cancer. *Am. J. Clin. Pathol.*, 88(4):399–404, 1987. 26

[83] W. Remmele and H. Stegner. Immunohistochemischer Nachweis von Ostrogenrezeptoren (ER-ICA) in Mammakarzinomgewebe: Vorschlag zur einheitlichen Formulierung des Untersuchungsbefundes. *Dtsch. Arztebl.*, 83:3362–3364, 1986. 26

[84] X. Ren, C. Fowlkes, and J. Malik. Scale-invariant contour completion using conditional random fields. In *International Conference on Computer Vision*, volume 2, pages 1214–1221, 2005. 122

[85] K. Rodenacker and P. Bischoff. Quantification of tissue sections. *Pattern Recognition Letters*, 11(4):275–284, 1990. ISSN 0167-8655. 122

[86] M. Rubin, M. Zerkowski, R. Camp, R. Kuefer, M. Hofer, A. Chinnaiyan, and D. Rimm. Quantitative determination of expression of the prostate cancer protein {alpha}-methylacyl-CoA racemase using automated quantitative analysis (AQUA): a novel paradigm for automated and continuous biomarker measurements. *American Journal of Pathology*, 164 (3):831, 2004. 46

[87] T. Sanders, T. Stokes, R. Moffitt, Q. Chaudry, R. Parry, and M. Wang. Development of an automatic quantification method for cancer tissue microarray study. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, page 3665, 2009. 44

[88] B. Schölkopf and A. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond.* The MIT Press, 2002. 93

[89] C. Schmid. Constructing models for content-based image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 39–45, 2001. 122

[90] C. Schmid and R. Mohr. Matching by local invariants. Technical Report RR-2644, INRIA, 1995. 49, 53

[91] M. Sciarabba, G. Serrao, D. Bauer, F. Arnaboldi, and N. Borghese. Automatic detection of neurons in large cortical slices. *Journal of Neuroscience Methods*, 182(1):123–140, 2009. 88

[92] J. Sont, W. de Boer, W. van Schadewijk, K. Grünberg, J. van Krieken, P. Hiemstra, and P. Sterk. Fully automated assessment of inflammatory cell counts and cytokine expression in bronchial tissue. *American Journal of Respiratory and Critical Care Medicine*, 167:1496–1503, 2003. 40

[93] P. Spyridonos, D. Cavouras, P. Ravazoula, and G. Nikiforidis. Neural network based segmentation and classification system for the automatic grading of histological sections of urinary bladder carcinoma. *Analytical and Quantitative Cytology and Histology*, 24(6):317–324, 2002. 36, 42

[94] A. Tabesh, M. Teverovskiy, H. Pang, V. Kumar, D. Verbel, A. Kotsianti, and O. Saidi. Multi-feature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Transactions on Medical Imaging*, 2007. 41

[95] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 123

[96] D. Turbin, S. Leung, M. Cheang, H. Kennecke, K. Montgomery, S. McKinney, D. Treaba, N. Boyd, L. Goldstein, S. Badve, et al. Automated quantitative analysis of estrogen receptor expression in breast carcinoma does not differ from expert pathologist scoring: a tissue microarray study of 3,484 cases. *Breast Cancer Research and Treatment*, 110(3):417–426, 2008. 45

[97] G. van de Wouwer, B. Weyn, P. Scheunders, W. Jacob, E. van Marck, and D. van Dyck. Wavelets as chromatin texture descriptors for the automated identification of neoplastic nuclei. *Journal of Microscopy*, 197 (1):25–35, 2000. 41

[98] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1-2):61–81, 2005. 19, 51, 54, 116

[99] G. Wahba. *Spline models for observational data.* Society for Industrial Mathematics, 1990. 93

[100] D. Weaver, D. Krag, E. Manna, T. Ashikaga, S. Harlow, and K. Bauer. Comparison of pathologist-detected and automated computer-assisted image analysis detected sentinel lymph node micrometastases in breast cancer. *Modern Pathology*, 16(11):1159–1163, 2003. 45

[101] J. Weaver and J. Au. Comparative scoring by visual and image analysis of cells in human solid tumors labeled for proliferation markers. *Cytometry Part A*, 27, Issue 2:189–199, 1997. 40

[102] B. Weyn, G. van de Wouwer, A. van Daele, P. Scheunders, D. van Dyck, E. van Marck, and W. Jacob. Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry Part A*, 33(1):32–40, 1998. 41

[103] D. Young. Convolutions in Matlab. `http://www.cogs.susx.ac.uk/courses/compvis/matlab_demos/convolution_demo.html`. Checked on March 24, 2008. 63, 95

[104] Y. Zhu, S. Williams, and R. Zwiggelaar. Computer technology in detection and staging of prostate carcinoma: a review. *Medical Image Analysis*, 10(2):178–199, 2006. 33