Gland Instance Segmentation in Colon Histology Images

By

Li Yang Wang

汪立阳

A thesis submitted in partial fulfilment for the requirements for the degree of

Doctor of Philosophy

at

the University of Central Lancashire

October, 2019

Student Declaration

Concurrent registration for two or more academic awards

Either *I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution

or ~~*I declare that while registered for the research degree, I was with the University's specific permission, a *registered candidate/*enrolled student for the following award:~~

_____

Material submitted for another award

Either *****I declare that no material contained in the thesis has been used in any other submission for an academic award and is solely my own work

or ~~*I declare that the following material contained in the thesis formed part of a submission for the award of:~~

_____

(state award and awarding body and list the material below):

Collaboration

Where a candidate's research programme is part of a collaborative project, the thesis must indicate in addition clearly the candidate's individual contribution and the extent of the collaboration. Please state below:

Signature of Candidate Li Yang Wang

Type of Award Doctor of Philosophy

School School of Engineering

Abstract

This thesis looks at approaches to gland instance segmentation in histology images. The aim is to find suitable local image representations to describe the gland structures in images with benign tissue and those with malignant tissue and subsequently use them for design of accurate, scalable and flexible gland instance segmentation methods.

The gland instance segmentation is a clinically important and technically challenging problem as the morphological structure and visual appearance of gland tissue is highly variable and complex. Glands are one of the most common organs in the human body. The glandular features are present in many cancer types and histopathologists use these features to predict tumour grade. Accurate tumour grading is critical for prescribing suitable cancer treatment resulting in improved outcome and survival rate. Different cancer grades are reflected by differences in glands morphology and structure. It is therefore important to accurately segment glands in histology images in order to get a valid prediction of tumour grade.

Several segmentation methods, including segmentation with and without pre-classification, have been proposed and investigated as part of the research reported in this thesis. A number of feature spaces, including hand-crafted and deep features, have been investigated and experimentally validated to find a suitable set of image attributes for representation of benign and malignant gland tissue for the segmentation task. Furthermore, an exhaustive experimental examination of different combinations of features and classification methods have been carried out using both qualitative and quantitative assessments, including detection, shape and area fidelity metrics.

It has been shown that the proposed hybrid method combining image level classification, to identify images with benign and malignant tissue, and pixel level classification, to perform gland segmentation, achieved the best results. It has been further shown that modelling benign glands using a three-class model, i.e. inside, outside and gland boundary, and malignant tissue using a two-class model is the best combination for achieving accurate and robust gland instance segmentation results. The deep learning features have been shown to overall outperform hand-crafted features, however proposed ring-histogram features still performed adequately, particularly for segmentation of benign glands. The adopted transfer-learning model with proposed image augmentation has proven very successful with 100% image classification accuracy on the available test dataset. It has been shown that the modified object-level Boundary Jaccard metric is more suitable for measuring shape similarity than the previously used object-level Hausdorff distance, as it is not sensitive to outliers and could be easily integrated with region-based metrics such as the object-level Dice index, as contrary to the Hausdorff distance it is bounded between 0 and 1. Dissimilar to most of the other reported research, this study provides comprehensive comparative results for gland segmentation, with a large collection of diverse types of image features, including hand-crafted and deep features.

The novel contributions include hybrid segmentation model superimposing image and pixel level classification, data augmentation for re-training deep learning models for the proposed image level classification, and the object-level Boundary Jaccard metric adopted for evaluation of instance segmentation methods.

Contents

Abstract 3

List of Figures 8

List of Tables 15

List of Acronyms 18

Thesaurus 19

Acknowledgements 21

Chapter 1 Introduction 22

1.1 Background 22

1.2 Motivation and aims 24

1.3 Image segmentation 25

1.4 Contributions 29

1.5 Thesis Outline 30

Chapter 2 Introduction to gland segmentation 32

2.1 Preparation of Histology Images 32

2.2 Gland segmentation 34

2.3 Gland dataset 38

2.4 Summary 39

Chapter 3 Mathematical foundation of pixel-level classifier 41

3.1 Machine learning 42

3.2 Decision trees 44

3.2.1 Splitting criteria 46

3.2.2 Thresholding 49
3.2.3 Weak learners 52
3.2.4 Termination conditions 53
3.3 Random forest 53
3.3.1 Bagging 54
3.3.2 Random subspace 55
3.3.3 Applications 55
3.4 Comparison of different forest models 57
3.4.1 Evaluation measures for the experiments 58
3.4.2 Experimental setup and results 58
3.5 K-means algorithm 61
3.6 Summary 62
Chapter 4 Feature extraction 64
4.1 Motivation 64
4.2 Grey-level co-occurrence matrix 66
4.3 Histogram 67
4.4 Local Binary Pattern 70
4.5 Histogram of oriented gradients 78
4.6 Deep learning features 81
4.6.1 LeNet-5 architecture 81
4.6.2 GoogleNet architecture 83
4.7 Features discriminative analysis 85
4.8 Summary 86

Chapter 5 Segmentation method and pre-/post-processing 87

5.1 Image classification and segmentation 87

5.2 Segmentation without pre-classification 87

5.3 Limitations of segmentation without pre-classification 89

5.4 Segmentation with pre-classification approach 91

5.4.1 Image-level classification 92

5.4.2 Pixel-level classification 97

5.4.2.1 Two-categories pixel-level classification problem 98

5.4.2.2 Three-categories pixel-level classification problem 98

5.4.2.3Segmentation with pre-classification at different levels 101

5.5 Pre-processing 104

5.6 Post-processing 106

5.6.1 Morphological post-processing 107

5.6.2 Level set algorithm 108

5.7 Summary 111

Chapter 6 Results 113

6.1 Evaluation metrics 113

6.1.1 Region-based evaluation metrics 113

6.1.2 Contour-based evaluation metrics 115

6.1.3 Analysis of evaluation metrics 118

6.2 Ranking strategy 126

6.3 Results for segmentation without pre-classification 127

6.3.1 Segmentation results for histogram features 127

6.3.2 Segmentation results with deep learning features 131

6.3.3 Summary of segmentation without pre-classification 134

6.4 Results for segmentation with pre-classification method 137

6.4.1 Results for image-level classification 137

6.4.2 Summary of pixel-level classification 138

6.5 Comparison of the two segmentation methods 141

6.6 Comparison of three segmentation methods 142

6.7 Summary 144

Chapter 7 Summary, contributions and future work 146

7.1 Summary 146

7.2 Contributions 147

7.3 Future research 149

Appendices 151

A Random forest results on UCI dataset 152

B Visualised extracted features 155

C Evaluation of features discriminative properties 161

D Segmentation results without pre-classification 172

E Segmentation results with pre-classification 175

F Ranking of different methods in gland segmentation 182

G Publications 183

H Screenshot of image-level classification results 184

References: 186

List of Figures

Figure 1.1 Sample images from the gland dataset (Nasir, 2015). 23
Figure 1.2 Tissue components in colon glands visualised using Hematoxylin-Eosin staining method. Gland images are from MICCAI 2015 gland database (Nasir, 2015) 24
Figure 1.3 Example of two images for which region homogeneity principle could be successfully deployed using: (a) pixel intensity (image from MATLAB), (b) texture features (image from http://w3.ualg.pt/~dubuf/pubdat/texture/texture.html; Du buf et al. (1990)) 25
Figure 1.4 Example explaining difference between semantic segmentation and instance segmentation, (from PASCAL VOC 2012; Everingham et al., 2011) 26
Figure 2.1 Preparation steps for histology images 33
Figure 2.2 The example images from MICCAI Gland Segmentation 2015 dataset (Nasir, 2015) 33
Figure 2.3 The example of images and corresponding ground truth from gland dataset (Nasir, 2015). The images in the top row are the images with benign tissue. The images in the bottom row are the images with malignant tissue. The corresponding ground truth is shown on the right of the original images 39
Figure 3.1 Decision tree. a) The simple structure of the decision tree model. b) The progress of different types of image through the decision tree model. 46
Figure 3.2 Different splitting approaches. a) The sample points drawn from Gaussian distributions. b) The probability of each class in the training data. c) The horizontal splitting approach for training samples. d) The vertical splitting method for training samples 49
Figure 3.3 Examples of different forest models. a) The forest model with low randomness and high correlations. b) The forest architecture with low correlations and high randomness (Criminisi et al., 2013): 50
Figure 3.4 Examples of different weak learners. a) Data samples drawn from two Gaussian distributions. b) The splitting results are for using the decision tree with axis-aligned weak learner. c) The results for using the decision tree with oblique weak

learner. 53
Figure 3.5 Validation parameters of different forest models on **Liver** database 61
Figure 4.1 The process of generating the GLCM features (Eleyan and Demirel, 2011) .. 67
Figure 4.2 Different images with the same intensity distributions (Xiaoling, 2009) 68
Figure 4.3 Examples of the extended histogram feature proposed by Xiaoling (2009) . 69
Figure 4.4 Extracting the ring histogram feature from one of the sample images 69
Figure 4.5 Creating the LBP features (Lahdenoja et al., 2013) 71
Figure 4.6 Example of eight neighbours uniform LBP features (Pietikäinen et al., 2011) 72
Figure 4.7 The patterns of rotation-invariant LBP feature (Ojala et al., 2000)..... 74
Figure 4.8 Creating original LBP feature for one of the gland images 75
Figure 4.9 Creating uniform LBP feature for one of the gland images 76
Figure 4.10 The process of extracting rotation-invariant LBP feature from histology image 77
Figure 4.11 The process of extracting rotation-invariant uniform LBP feature from the gland image 77
Figure 4.12 The process of generating the original HOG features 79
Figure 4.13 The process of extracting the local circular Fourier HOG from histology image 80
Figure 4.14 LeNet5 architecture adapted in gland segmentation..... 82
Figure 4.15 The architecture used in segmentation with a pre-classification method for three-classes pixel-level classification method 83
Figure 4.16 GoogleNet architecture to extract the local patterns in gland segmentation 84
Figure 5.1 The process of the training phase of segmentation without pre-classification method (Manivannan et al., 2017). (a) sample image from the gland dataset. (b) represents the label for the same sample image. (c) represents the feature vector generated from the gland image, and (d) the label vector generated from the label image. (e) represents the trained random forest model 88
Figure 5.2 Testing process of segmentation without pre-classification (Manivannan et al.,

2017). (a) sample testing gland image from the gland database. (b) feature vector for the selected area. (c) the trained classifier (the same classifier as shown in 5.1.e). (d) prediction for the selected patches from the testing image..... 89
Figure 5.3 Sample images from the gland dataset 90
Figure 5.4 The process of segmentation with the pre-classification approach 91
Figure 5.5 Structure of the HMAX model (Theriault et al., 2013) 93
Figure 5.6 Original images and images after local image deformation 96
Figure 5.7 Original images and images after colour jitter 97
Figure 5.8 Sample images and corresponding ground truth from gland dataset 99
Figure 5.9 Sample images and the inside gland labels of corresponding images 100
Figure 5.10 Sample image and the boundary label for corresponding image 100
Figure 5.11 Sample images and inside labels for malignant category 101
Figure 5.12 Sample images and boundary labels for malignant category 101
Figure 5.13 The details of three segmentation methods used for gland instance segmentation 103
Figure 5.14 Sample histology images with white areas that need/don't need to be removed 104
Figure 5.15 The results of removing unwanted white areas in histology images 105
Figure 5.16 Sample images after using histogram correction 105
Figure 5.17 Sample of the probability maps in segmentation without pre-classification 106
Figure 5.18 Morphological post-processing of processing probability maps 107
Figure 5.19 Conventions used in the level set algorithm (Zhang et al., 2008) 109
Figure 5.20 Steps in the level set post-processing method..... 111
Figure 6.1 F1 score evaluation example 118
Figure 6.2 Object-level Dice index evaluation example 119
Figure 6.3 Object-level Hausdorff distance evaluation measure 120
Figure 6.4 Example of infra-segmentation and over-segmentation of two different gland objects 121
Figure 6.5 Significance of the object-level Hausdorff distance measure 122

Figure 6.6 Significance of the F1 score 123
Figure 6.7 Significance of the object-level Dice index 124
Figure 6.8 Object-level Boundary Jaccard index VS object-level Hausdorff distance ... 125
Figure 6.9 Example of the ground truth and three different segmentation results 126
Figure 6.10 Comparison of the results of ring histograms with different sizes of input patches in segmentation without pre-classification. Different colours indicate different gland objects. 129
Figure 6.11 Comparison of results of different post-processing method 130
Figure 6.12 Segmentation results of deep learning using LeNet-5 with different size of input patches 131
Figure 6.13 A sample of segmentation results with LeNet-5 features as function of different number of training input patches. 132
Figure 6.14 A sample of segmentation results with GoogleNet deep features using different sizes of input patch 133
Figure 6.15 A sample of segmentation results with the GoogleNet deep features using different numbers of training patches 134
Figure A.1 Different evaluation measures of different forest models on **tic-tac-toe endgame** 152
Figure A.2 Different evaluation measures of different forest models on **Ionosphere** data 153
Figure A.3 Different evaluation measures of different forest models on **Sonar** data .. 153
Figure A.4 Different evaluation measures of different forest models on **Tic-tac-toe** data 154
Figure B.1 Ring histogram after using PCA algorithm in 2D feature space. 155
Figure B.2 The original HOG feature after using PCA algorithm in 2D feature space .. 155
Figure B.3 The circular Fourier HOG feature after using PCA in 2D feature space 156
Figure B.4 Original LBP feature after PCA algorithm in 2D feature space 156
Figure B.5 Original rotation-invariant LBP feature after PCA algorithm in 2D feature space 157
Figure B.6 Uniform LBP after using PCA algorithm in 2D feature space 157

Figure B.7 Rotation-invariant uniform LBP after using PCA algorithm in 2D feature space 157
Figure B.8 Deep feature from LeNet-5 after using PCA algorithm in 2D feature space 158
Figure B.9 Deep features from GoogleNet after using PCA algorithm in 2D feature space 158
Figure B.10 Deep features from LeNet-5 with three classes after using PCA in 2D feature space 159
Figure B.11 Deep features from LeNet-5 with 2 classes after using PCA in 2D feature space 159
Figure B.12 Ring histogram with 2 classes after using PCA in 2D feature space in segmentation with pre-classification 160
Figure B.13 Ring histogram with 3 classes after using PCA in 2D feature space in segmentation with pre-classification 160
Figure C.1 The accuracy of different features generated from different co-occurrence matrices from the gland images 161
Figure C.2 Accuracy of different models (different sized patches) using K-means algorithm 162
Figure C.3 Accuracy of different models (different number of rings in each patch) using K-means 163
Figure C.4 Accuracy of ring histogram features in segmentation with pre-classification method 163
Figure C.5 The estimation results for different LBP features in segmentation without pre-classification 164
Figure C.6 Estimation results for different size of rotation-invariant uniform LBP feature 164
Figure C.7 Accuracy of rotation-invariant uniform LBP features in segmentation with pre-classification 165
Figure C.8 Accuracy of two versions of HOG features in segmentation without pre-classification. 165
Figure C.9 Accuracy for different Fourier HOG features in segmentation with pre-

classification 166
Figure C.10 The accuracy of circular Fourier HOG with different sizes of circles in each patch..... 166
Figure C.11 Accuracy of different circular Fourier HOG in segmentation with pre-classification 167
Figure C.12 Accuracy of deep feature from LeNet-5 architecture with different numbers of input patches in segmentation without pre-classification 167
Figure C.13 Accuracy for different deep learning feature vectors from LeNet-5 architecture with different sizes of input patches in segmentation with pre-classification 168
Figure C.14 Accuracy of different deep learning features with LeNet-5 architecture in segmentation with pre-classification. 168
Figure C.15 Accuracy for different deep learning features from GoogleNet with different sizes of input patches in segmentation with pre-classification..... 169
Figure C.16 Accuracy of deep features from GoogleNet with different input patches after using K-means clustering in segmentation without pre-classification 170
Figure C.17 Accuracy of deep features from GoogleNet with different input patches after using K-means clustering in segmentation without pre-classification 171
Figure D.1 Example of a probability map using features from co-occurrence matrix.. 172
Figure D.2 Probability maps for the two failed LBP features 172
Figure D.3 Example of segmentation performance with the two uniform LBP features 173
Figure D.4 Results of different input patches size using rotation invariant uniform LBP 173
Figure D.5 Probability maps for HOG feature..... 174
Figure D.6 Results of different input patches size using circular Fourier HOG 174
Figure E.1 Examples of results for the benign category with two classes of different features in segmentation with pre-classification 175
Figure E.2 Example results of benign category with three classes of different features in segmentation with pre-classification method..... 176

Figure E.3 Example results of malignant category with two classes of different features in segmentation with pre-classification method. 177
Figure E.4 Example of results of malignant category with two classes of different features in segmentation with pre-classification 178
Figure E.5 Examples of training malignant images and the image after using local image deformation 179
Figure E.6 Example of training benign images and the corresponding images after using local image deformation 180

List of Tables

Table 3.1 The values of different splitting criteria based on the different splitting methods 49
Table 3.2 The details of the UCI datasets used to test different forest models 57
Table 4.1 Texture information generated from the co-occurrence matrix (Haralick and Shanmugan, 1973) 67
Table 6.1 F1 score for the segmentation results in Figure 6.1 119
Table 6.2 Object-level Dice index for the results in Figure 6.2 119
Table 6.3 Shape similarity for the segmentation results shown in Figure 6.3 120
Table 6.4 Evaluation measures for the segmentation results shown in Figure 6.5 122
Table 6.5 Evaluation measures for the segmentation results shown in Figure 6.6 123
Table 6.6 Evaluation measures for the segmentation result shown in Figure 6.6 124
Table 6.7 Evaluation measures for the segmentation result shown in Figure 6.8 125
Table 6.8 Evaluation measures for the segmentation result shown in Figure 6.9 126
Table 6.9 Comparison of results when using different numbers of trees in random forest 128
Table 6.10 Comparison results when using different patch size 129
Table 6.11 Comparison results when using different number of rings..... 129
Table 6.12 Comparison results when using different post-processing methods 130
Table 6.13 The results of deep learning features using LeNet-5 with different size of input patches 131
Table 6.14 Results of deep feature from LeNet-5 with different number of input patches 132
Table 6.15 Evaluation measures for GoogleNet deep features with different sizes of input patch..... 133
Table 6.16 Evaluation measures for the GoogleNet features with different numbers of training patches 133
Table 6.17 The overall performance of different features in segmentation without pre-classification 135

Table 6.18 Performance of different features for the benign category only, using segmentation without pre-classification 136
Table 6.19 Performance of different features for the malignant category only, using segmentation without pre-classification 136
Table 6.20 Image-level classification results of deep learning models (AlexNet, GoogleNet, ResNet-50) on testing data 138
Table 6.21 Overall ranking of segmentation results for different features on images with benign tissue..... 140
Table 6.22 Overall ranking of segmentation results for different features on images with malignant tissue..... 140
Table 6.23 Performance of the best gland segmentation methods with and without pre-classification 141
Table 6.24 Performance of the best gland segmentation methods with and without pre-classification based on the object-level Boundary Jaccard index..... 142
Table 6.25 The overall segmentation results for the three methods shown in Figure 5.13 143
Table 6.26 Malignant category segmentation results for three methods shown in Figure 5.13 143
Table 6.27 Benign category segmentation results for three methods shown in Figure 5.13 144
Table D.1 Ranking of segmentation performance of different uniform LBP features .. 173
Table D.2 Ranking of rotation-invariant uniform LBP feature using different sizes of input patch..... 173
Table D.3 Ranking of Fourier HOG features using different sizes of circle..... 174
Table E.1 Ranking of different feature of two classes of benign category in segmentation with pre-classification 175
Table E.2 Ranking of different features of three categories of benign images in segmentation with pre-classification..... 176
Table E.3 Ranking of segmentation results of different feature of two classes of malignant images 177

Table E.4 Ranking of segmentation results of different features of three classes in the malignant category 178
Table E.5 Ranking of two classes of malignant category of the ring histogram feature with/without local image deformation 179
Table E.6 Ranking of segmentation results of benign category three classes using histogram with/without local image deformation 180
Table E.7 Segmentation results for two types of histogram with three classes of benign category images 181
Table E.8 Segmentation results for two types of histogram with two classes of benign category images 181
Table E.9 Segmentation results for two types of histogram for two classes of malignant category images 181
Table E.10 Segmentation results for two types of histogram for two classes of malignant category images 181
Table E.11 Segmentation results of histogram feature with/without histogram correction in segmentation without pre-classification 181
Table F.1 The ranking of different methods in Gland Segmentation Challenge (Warwick.ac.uk, 2018b) 182

List of Acronyms

CART	Classification And Regression Tree
CNN	Convolutional Neural Network
FCN	Fully Convolutional Network
GLCM	Grey Level Co-occurrence Matrix
H&E method	Hematoxylin-Eosin method
HOG	Histogram of Oriented Gradients
ID3	Iterative Dichotomiser 3
LBP	Local Binary Patterns
LTP	Local Ternary Patterns
MLP	Multilayer Perceptron
PCA	Principle Component Analysis
PDE	Partial Differential Equation
SVM	Support Vector Machine
RBF	Ratial Basic Function
RF	Random Forest
UCI databases	University of California Irvine Databases

Thesaurus

Binary classification problem: A binary classification problem is to separate the unknown samples into two given categories (these two categories are given from training data).

Centroid: This term is a concept in K-means algorithm, and it refers to the centre point of the cluster.

Classification: This refers to image classification in this work. Image classification is the process of grouping all the pixels in an image into one of a number of given classes or categories. This term is related to image segmentation, and the differences between these two terms are discussed in the term 'segmentation'.

Deep features: This refers to deep learning features. These features are generated by deep learning architecture, and could be extracted before fully connected layer, i.e. AlexNet, GoogleNet and ResNet features.

Ensemble learning: This is a process in which a set of classifier are trained independently, such that classifier in the model will make the predictions. The final output of the ensemble learning is to combine the output of each classifier using some strategy (such as majority voting).

Feature space: The term is one of the well-known terms in machine learning. Feature space refers to the abstract space where the training features are contained.

Feature: This term often refers to a feature vector in machine learning. A feature vector is the vector which contains a set of digits has been used to describe the properties of the image, and it has been used as the input data to train the classifier in classification and regression.

Ground truth: This term is widely used in machine learning. The ground truth used in this work is annotated by histology experts, and the ground truth of a histology image is the benchmark which represents the gland and background parts in the images.

Natural image: This is the term used in (Sumengen and Manjunath, 2005), and it refers to the images contain many texture features.

Pixel-level classification: This refers to gland instance segmentation, and it is the second part of the proposed segmentation method. Gland segmentation classifies gland and non-gland parts in histology images and subsequently separates different gland objects. The reason why these two terms are similar is that gland instance segmentation deals with pixel-wise classification for test histology images. The term is defined compared with the image-level classification (discussed at the same page).

Pixel-level classifier: The classifier used to deal with gland segmentation, and the classifier used in this work is the forest model with axis-aligned weak learner, mid-point thresholding and Gini impurity splitting criterion.

Rotation-invariant property: If the extracted feature has rotation-invariant properties, these features are unaffected by the rotation of the images.

Segmentation: This term refers to image segmentation in this work, image segmentation is to divide the images into several regions (sets of pixels in the same image) based on the semantic meanings. Image segmentation could be treated as pixel-wise classification. Gland segmentation is a pixel-wise classification for each histology image. The difference between these two terms (segmentation and classification) is that segmentation deals with each pixel but classification identifies the whole image as a given classes.

Image-level classification: This refers to the first part of the proposed segmentation method (segmentation with pre-classification method) in this work. The purpose of the image-level classification is to separate histology images into benign or malignant case.

Hand-crafted feature: These features are generated by applying different methods using the information in the images, such as histogram, LBP and HOG.

Residual connection: This term is used in the ResNet architecture. It means the output of a layer is a convolution of its output plus input.

Acknowledgements

I would like to express my gratitude to my Director of Studies, Prof. Bogdan Matuszewski, for providing me the useful support and every bit guidance and expertise over last four years. He not only guided me in the area of image segmentation, but also encouraged me to spend more time enjoying the research. I felt lucky that I have a supervisor like him, and he teaches me a hard work attitude in research.

I would also like to thank my second supervisor, Dr. Yu Zhou, for his comments and helpful feedback on the chapters of the thesis. Thanks must go to my colleague, Yun Bo Guo, because the topic of my research developed when working together on an image segmentation challenge. I am also thankful to other friends and the colleagues who helped me whenever I needed.

I would also like to thank my internal examiner, Dr. Jules simo, for his huge support and feedback during my corrections.

I would also like to thank my friends in China for sharing the happiness and patiently listen to my complains during my research. Especially my best friend, Xiong Xu, for giving me the feedback and possible suggestions for improving the structure of the thesis.

I would like to thank my parents in China, especially my mom who had paid a lot of attentions on my life rather than her own life. Her patient attitude taught me not to giving up especially when I am writing the thesis. The thesis would not have been written without her support.

I would also thank to University of Warwick, because they have been made the gland data publicly available. All experiments, which include in this thesis, would not happen without having this data (gland data).

The past four years have been the most memorable time in my life. It was the time that I could learn the things in the most interested area and felt the beauty of the state-of-the-art methods in image segmentation. Finally, I would like to thank myself not giving up even in hard times during the research.

Chapter 1

Introduction

Introduction to histology imaging and gland tissue components, in benign and malignant glands, are briefly discussed at the beginning of this chapter. Subsequently, the motivation and the aims of the research described in this thesis are discussed. The two main types of image segmentation are introduced and their differences described. The original contributions of this research are summarised, and the overall structure of the thesis is described.

1.1 Background

The work described in this thesis deals with segmentation of glands in histology images. This chapter briefly introduces essential information about histology imaging and glands' morphology to facilitate a better understanding of the research objectives as well as methods described in this thesis.

Histology is a study which uses microscopy data to investigate the tissues of animals and plants. For example, identification of different types of cancer, or other diseases, often requires histology image analysis for detection of tissue abnormalities including tissue morphology irregularities. The key steps for acquiring histology images are detailed in Chapter 2. The reason why histology images are important is that these images can help biologists to analyse and understand morphology and function of different tissues in animals and plants.

In recent years, automated gland segmentation has become one of the important topics in biomedical image analysis research. The images with benign tissue are shown in the top row in Figure 1.1 and those with malignant tissue in the bottom row. The artificially highlighted (brighter) parts in these images represent glands. The morphological structure and visual appearance of gland objects can be different even for the same type of tissue (benign or malignant). Gland segmentation is a challenging problem, as the methods have to cope with these variabilities. For example, because

glands in malignant tissue can have very significantly different shape and size, it is rather difficult to design meaningful segmentation shape prior. Furthermore, the image patterns representing glands and surrounding tissue could be very similar leading to under or over segmentation of glands. The proposed methods have to recognise these different variation patterns to establish robust decision about presents of gland tissue. Costantini et al. (2003) and Van Putten et al. (2011) discussed how even histology experts have different subjective views on tissue classification. The objectivity of computational methods can support an improved analysis of histological data (McCann et al., 2015) leading to more robust decisions.

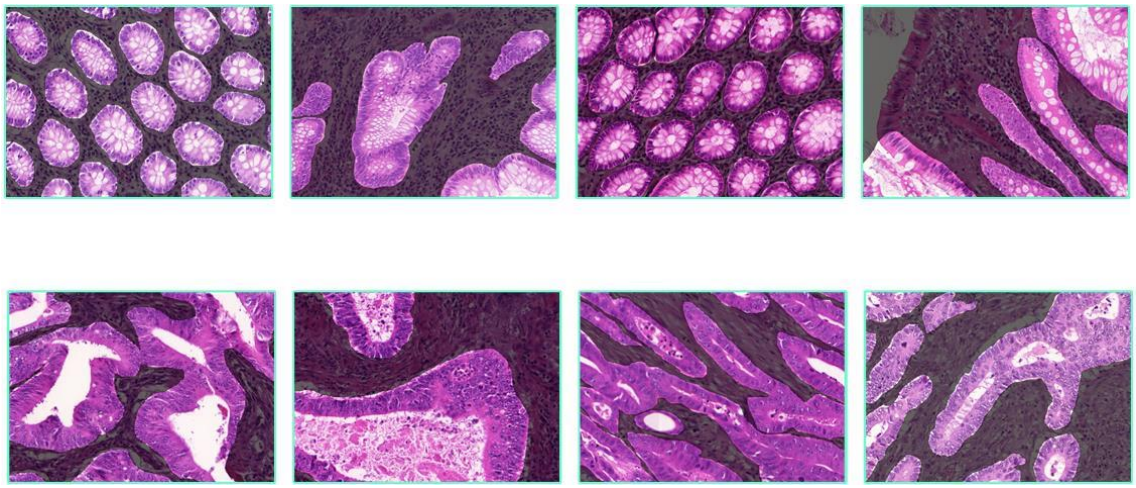


Figure 1.1 Sample images from the gland dataset (Nasir, 2015).

Colorectal cancer is one of the commonly diagnosed cancers in both males and females (Torre et al., 2015). Accurate tumour grading is critical for individual cancer treatment procedure and resulting survival rate. Fleming et al. (2012) pointed out that the different grading of the cancer was reflected by different structure of gland objects. It is therefore important to accurately segment glands in histology images in order to get a valid classification of different tumour grading.

The colon gland (refer to Figure 1.2) could be divided into four tissue components: lumen, cytoplasm, epithelial cells, and stroma. In what follows, the stroma in either benign (normal, Figure 1.2.a) or malignant gland (Figure 1.2.b) is treated as a background structure in the computational approach to gland segmentation. The epithelial cells form the boundary of the gland, which encloses the internal gland structures, i.e. lumen and cytoplasm. The reason for stroma is being treated as background, is that typically, it is

not used for cancer grading and therefore is not of interest for histology experts (Kather et al., 2019).

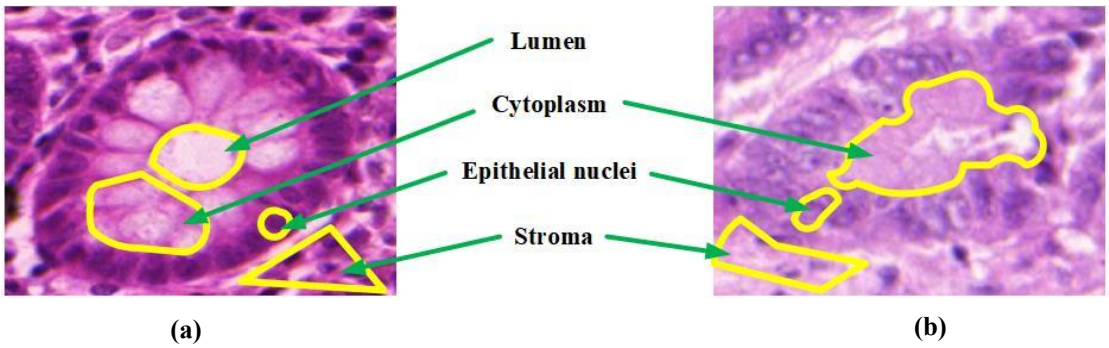


Figure 1.2 Tissue components in colon glands visualised using Hematoxylin-Eosin staining method. Gland images are from MICCAI 2015 gland database (Nasir, 2015)

Typical appearances of tissue components of benign and malignant glands are shown in Figure 1.2. Even from a brief visual inspection, the structure of the malignant gland is different from that of the benign gland. In any of these cases, the gland segmentation method must deal with the variability of glands size, shape and tissue appearance.

1.2 Motivation and aims

Gland segmentation aims to automatically delineate gland and non-gland structures in histology images. This work is essential as analysis of morphology and function of gland objects is useful for detection and quantification of many types of diseases. For example, the morphological representation of the gland structure is employed to describe the degree of malignancy of various adenocarcinomas, including breast and colon.

It has been shown that malignancy of cancer can be assessed by the shape and appearance of glands and therefore gland segmentation is an important step in enabling automatic classification of different tumour types. The main motivation behind this work is twofold. First, it is to investigate the reliable automatic segmentation tools that are, as argued above, essential for accurate cancer grading. Second, segmentation in general is one of the key enabling techniques in image processing and computer vision, therefore the progress made on image segmentation is instrumental in developing a large class of image computational methods. This research aims to find the suitable local image

representations to describe the gland parts in images with benign tissue and those with malignant tissue and investigate effective segmentation algorithms.

In this work, random forest has been used as a primary classifier for the pixel-wise classification of histology images. Although other techniques have been proposed and used for the pixel level classification in histology images, including recently very popular convolutional neural networks, the random forest methodology has been chosen as it provides a good compromise between accuracy, scalability and flexibility of the design.

1.3 Image segmentation

Early image segmentation techniques used homogeneity of regions as the primary criterion for segmentation. The region homogeneity could be defined in terms of: intensity, colour, texture, shape or other relevant features. Figure 1.3 shows an example of two images for which region homogeneity based segmentation could be successfully used with pixel intensity and texture based features, respectively. With a suitable feature defined and calculated the segmentation process usually involves a simple thresholding performed in the feature space.

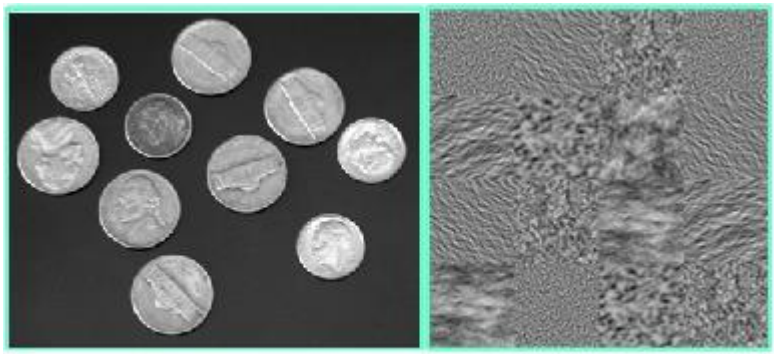


Figure 1.3 Example of two images for which region homogeneity principle could be successfully deployed using: (a) pixel intensity (image from MATLAB), (b) texture features (image from http://w3.ualg.pt/~dubuf/pubdat/texture/texture.html; Du buf et al. (1990))

Semantic segmentation is a much more complex problem than the region homogeneity-based image segmentation. This is because of the significant variation in the appearance (i.e. image patterns leading to differences in image perception) of objects belonging to the same category of segmented entities. For this type of segmentation, a simple region homogeneity principle may not be sufficient as objects in

the same class could have significantly different characteristics. For example, for the image in Figure 1.3, although the bottles belong to the same “bottle” class, they have different size, colour pattern, and shape. The aim of the semantic segmentation in this case is to delineate all the bottles as a single entity despite significant differences in their appearance. In case of the semantic segmentation different object instances of the same class are annotated by the same label (i.e. are assigned to a single, but possibly disconnected, image region).

Instance segmentation differs from the semantic segmentation in the sense that different instances of the same object class are given a different id (i.e. identity number number) in order to classify different objects, i.e. different instances of the same object class are assigned unique image regions. In the case of the “bottle” class in Figure 1.4, all the three bottles are dilated separately, yet still are recognised as representing the same object class.

Original images | Instance segmentation | Semantic segmentation

Figure 1.4 Example explaining difference between semantic segmentation and instance segmentation, (from PASCAL VOC 2012; Everingham et al., 2011)

In case of gland instance segmentation in colon histology images the objective is to classify the gland and non-gland structures and then label different instances of the gland object with unique labels. The following section provides a brief review of image segmentation methods.

There are many methods designed to solve image segmentation problems, and they have been analysed and reported in the literature (Pal et al., 1993; Yuheng and Hao, 2017). One of the simplest approaches used for image segmentation is threshold based methods (Cheriet et al., 1998; Hammouche et al., 2008). The main objective of these thresholding methods is to find the optimal threshold value in the corresponding image feature space. Histogram thresholding method based on the similarity between the grey

levels is described in (Tobias and Seara, 2002). Arifin and Asano (2006) used the clustering similarity measure to determine the value of the threshold, which was the main criteria in their method. Tan and Isa (2011) introduced a colour image segmentation method which combined histogram thresholding and fuzzy *c*-means algorithm. In this hybrid method, the histogram thresholding method was used to determine different regions in the colour image. The fuzzy *c*-means algorithm improves the compactness of different regions generated by histogram thresholding. One of the disadvantages of histogram thresholding techniques is that this method only considers the pixel intensity but does not take into account the spatial relationships present in the images. In this case, histogram thresholding methods may not be useful in segmentation of blurred images.

Another, frequently used image segmentation methodology is based on edge detection. Examples of edge detectors include Sobel (Vincent and Folorunso, 2009) and Canny (Canny, 1986). Sumengen and Manjunath (2005) introduced a multiscale edge detection method applied to natural images. Segmentation results provided by this method improved significantly when compared with one single scale edge detection approach. The key drawback, of these edge detection methods, is that they could not perform well when there are too many edges. In that case, the segmentation methods cannot identify a closed boundary. However, these edge detection based image segmentation methods are useful when detecting the continuous edges in images. Brejl and Sonka (2000) introduced a hybrid method which combined the object shape model and border appearance model. The object shape model (e.g. built using Hough Transform (Ballard, 1981)) is used to describe a likely shape of the object and is often represented as a mean shape and probability map of possible shape variations. These models are usually constructed from set of training images and used to locate an approximate position of the object. The appearance models are constructed in a similar manner but the mean and appearance variability describe pixel intensities rather than images edges.

Region-based techniques are another commonly used methods in image segmentation. Adams and Bischof (1994) introduced a method of region growing for

image segmentation. This seed (where seed referes to a small number of pixels selected in an image) region growing method controlled the initial seeds without tuning the homogeneity parameters. Each seed is responsible for representing each connected object associated with that seed, leading to dividing image into several regions. The final regions in the image were built by merging the pixels to their nearest seed region. Zhang et al. (2008) introduced a hybrid level set algorithm to segment the objects in medical images. The proposed method addressed a so called edge leakage problem, frequent for active shape models. The method is able to locate a correct boundary target object even in images with in a complex background. Morar et al. (2012) introduced an active contour without edges method to segment different gland objects. That method could handle the problem with two different gland objects that were close to each other, and correctly segmented different objects in low contrast images.

Watershed transformation is another useful technique for image segmentation. Wang (1997) introduced a multiscale method for calculating image gradients, handling blurred and step edges in images. Watershed transformation could provide over-segmented results. To overcome that problem, Belaid and Mourou (2011) introduced a hybrid method which combined watershed transformation and topological gradient approaches. Yang et al. (2007) introduced a hybrid method which combined the watershed and normalised cut to solve the over-segmentation problem. Ng et al. (2006) proposed a hybrid method which combined k-means and the improved watershed method for medical image segmentation applications. That method could handle both over-segmentation and noise. Graph partition based methods are also widely used in image segmentation, with the most popular, including: normalized cuts (Shi and Malik, 2000), random walker method (Grady, 2006), minimum cut (Wu and Leahy, 1993), isoperimetric partitioning (Grady and Schwartz, 2006), and minimum spanning tree-based segmentation (Zahn, 1970). In the last two decades, a number of graph partition based methods were introduced and achieved good segmentation results. Shafarenko et al. (1997) introduced a graph partition method to segment the random texture colour images. That method has combined the watershed and merging algorithms, and achieved good results on noisy colour images.

Segmentation methods described so far could be successful when applied to relatively simple problems. They are not suitable for semantic segmentation with complex object. In recent years, machine-learning based methodologies have become popular and have been successfully applied in semantic segmentation. Different from other image segmentation approaches, they are based on feature extraction and machine learning techniques. Wang et al. (2011) used the pixel-level colour feature and the texture feature to train the SVM classifier. The method achieved good segmentation results when compared with then the state-of-the-art methods. Schroff et al. (2008) used the random forest model with textons, colour, filterbank and HOG features. There are many other machine-learning based image segmentation methods (Andrew et al., 2003; Ren and Malik, 2003; Powell et al., 2008). The key differences between different machine-learning based segmentation methods are derived using different image features and different classification methods operating on these featuresThe machine-learning based methods are primary methods used in this research for gland instance segmentation. The key processing components for these approaches involve image feature extraction and subsequent pixel-level feature classification. Different feature extraction methods used to obtain information about local patterns in histology images are detailed in Chapter 4. Various classification methods which could be used with these features for segmentation of histology images are detailed in Chapter 3.

1.4 Contributions

This thesis presents a number of novel approaches for gland instance segmentation in histology images. The main contributions of this research are summarised as follow:

- Two main segmentation processing pipelines, with and without pre-classification, have been proposed in this research. The processing with image pre-classification has been further divided into pre-classification at the feature extraction level and the pixel-classification level. These processing architectures have been extensively tested with different gland categories, number of target classes, and different image feature sets. It has been demonstrated that all these parameters are important when selecting the optimal segmentation method for a given

problem.

- As part of the proposed processing pipeline, an image-level classification algorithm has been employed in order to identify image into benign or malignant cases. It has been shown that this method is able to differentiate between benign and malignant gland images with very high accuracy (100% on the available test data), when used with proposed data augmentation methods.
- A modification of a previously proposed, contour-based, Boundary Jaccard metric has been devised, adopting it for evaluation of the instance segmentation methods. The discussion and experimental evidence demonstrate that this measure is better than the previously used object-level Hausdorff distance metric as it is not sensitive to outliers and it can be easily integrated with region-based metrics such as the object-level Dice index.

1.5 Thesis Outline

The remainder of the thesis is organised as follows. An introduction to gland segmentation problem is given in Chapter 2. It includes a brief description of the key steps needed for production of histology images and a comprehensive review of the existing gland segmentation methods. The gland database used throughout this thesis is also introduced. Chapter 3 presents basic classification methods used in this research for gland segmentation. A comparative analyses of different classification models are provided in order to identify the key models' design parameters and their characteristics. Chapter 4 discusses the different feature extraction methods employed to characterise information present in the gland histology images. Discriminative properties of the described features are also investigated. Chapter 5 describes the complete processing pipelines proposed for the gland segmentation problem. The two main processing structures, with and without pre-classifications are described and the concepts of image- and pixel- level classifications are introduced. This chapter also includes descriptions of histology image pre-processing and the segmented image post-processing algorithms. Chapter 6 describes number of adopted segmentation evaluation measures. Characteristics of these measures are discusses and their complementary properties are

explained. The segmentation evaluation scheme is introduced. This chapter also includes a comprehensive assessment of different segmentation configurations, with tests evaluating different image features, design parameters and segmentation architectures. Chapter 7 summarises the research, highlights the original contributions and draw attention to possible directions for future work.

Chapter 2

Introduction to gland segmentation

The previous chapter introduced the problem of gland segmentation in histology images and presented a brief classification of methods used in image segmentation. This chapter will introduce the process of acquiring the histology images, and previous methods applied to gland segmentation are reviewed subsequently. Limitations of those methods and possible improvement are also discussed. Finally, details of the gland dataset used in this work are also provided.

2.1 Preparation of Histology Images

Chapter 1 Section 1.1 explained the basis of histology and morphology of benign and malignant glands. In order to get the histology images one needs to prepare the histology slide – the steps for the preparation of these slides are briefly discussed in this section. Nicola (2017) described the details of the preparation steps. Figure 2.1 shows five preparation steps and the details of each step are:

- Step 1: Samples of the biological tissue are fixed by using chemical fixation
- Step 2: Water and formalin are removed from the tissues, and an organic solvent is used to remove the alcohol.
- Step 3: Using the paraffin wax for embedding tissue parts, the tissue surrounded by the paraffin wax is treated as a ‘block’. This block supports very thin sectioning.
- Step 4: The embedded tissue are sectioned. This is an important step as it provides thin slides of tissue samples that illustrate the microstructure of the corresponding tissue regions.
- Step 5: The tissue is stained to enhance the contrast and highlight the features of interest. The visual outlook of gland objects in histology images depends strongly on the selected staining method.

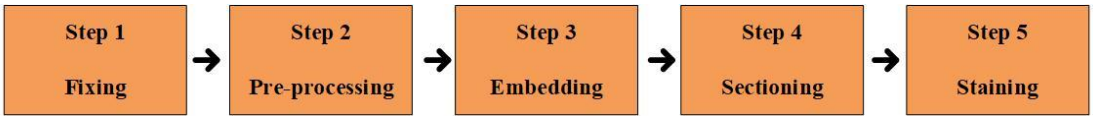


Figure 2.1 Preparation steps for histology images

For the histology sample preparation all these steps are essential for getting a finished histology slide. For researchers working on image segmentation, the staining step is possibly the most important as the appearance of the image strongly depends on this step. The colour in histology images is profoundly affected by the staining methods used to label the specific structures of the tissue. There are many types of staining methods used in this tissue labelling. The H&E (Hematoxylin and Eosin) staining method is one of the staining methods widely used in bio-imaging. The MICCAI 2015 gland dataset (Nasir, 2015) use H&E staining method. Hematoxylin is responsible for nuclei appearing to be blue in the image because the nuclei acids attract Hematoxylin, and Eosin causes cytoplasm to be stained pink. Figure 2.2 shows example images from MICCAI 2015 gland dataset (Nasir, 2015).

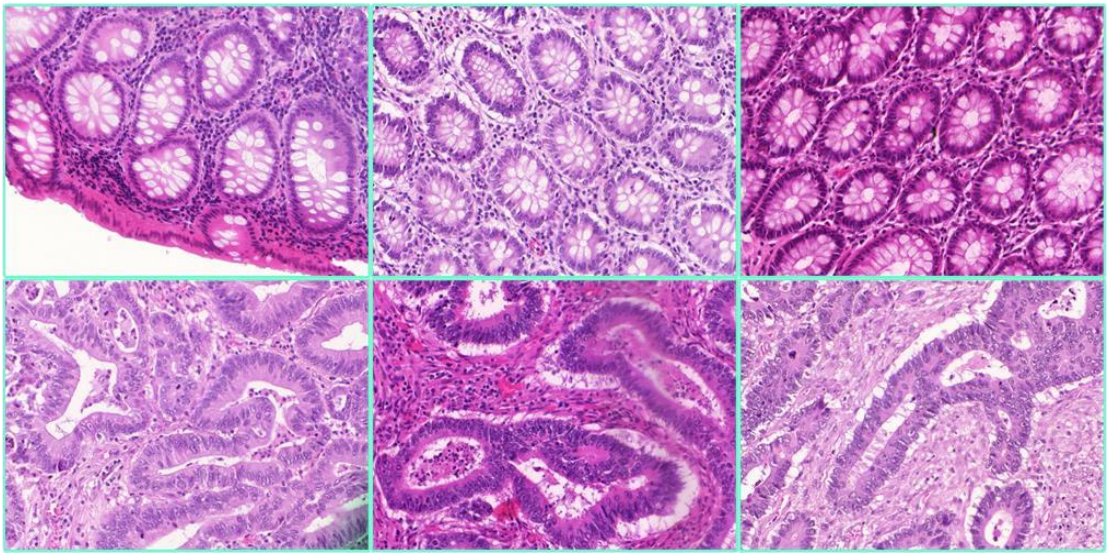


Figure 2.2 The example images from MICCAI Gland Segmentation 2015 dataset (Nasir, 2015)

From visual inspection, the colour pattern in these histology images is different, with the colour being strongly affected by variation in the staining processes applied to different slides. The details of the MICCAI 2015 Gland dataset are provided in Chapter 2 Section 2.3.

2.2 Gland segmentation

This section provides a brief review of the methods used in gland segmentation. The methods have used the morphological techniques and region-growing methods to classify the gland and the background in the histology images; and the machine-learning based methods have been applied to gland segmentation, either as deep learning approaches or traditional machine learning techniques, in order to identify the gland and non-gland parts in histology images. Finally, hybrid methods combine deep learning and traditional machine learning techniques to distinguish the gland and non-gland parts.

Wu et al. (2005a) introduced a method using a 2D-Gaussian low pass filter to find the epithelial cell with the intensity smaller than that of the threshold. The gland area was determined by dilating the pixel classified as the epithelial cell with a circle with radius. In the same year, Wu et al. (2005b) introduced a region growing approach to classify the nuclei region from other parts by setting a threshold in the images. The large empty parts (lumen parts shown in Figure 1.2) are employed to initialise the seed points for region growing. The chain of epithelial nuclei is used as the stop criterion for the region growing. The drawback of these two methods is that the threshold is generated individually for each image. The region growing approach achieved good performance in the images with healthy tissue and those with abnormal intestinal gland, however validation was only done by visual inspection.

Gunduz-Demir et al. (2010) introduced a method that used graph connectivity to identify the initial seed for the region growing. This is different from the methods described above that employed the pixel-level information and used a set of surrounding pixels to represent each tissue part in the images. To growing beyond the gland parts in the images, edges between the nuclear objects are employed to cease region grow. The final step is to remove the region that has no glandular characteristics. However, this method is also limited to images with only benign glands.

Some studies used image features to represent the tissue components and subsequently used classifier trained on these features to make the predictions (Diamond et al., 2004; Farjam et al., 2007; Altunbay et al., 2010; Nguyen et al., 2010; Rathore et al., 2013; Fu et al., 2014; Akbar et al., 2015; Cohen et al., 2015; Ap et al., 2017). Diamond et

al. (2004) used morphological structure and texture features from histology images to classify test histology images into stroma, benign and carcinoma categories, and the image within a window 100-by-100 pixels was selected from whole-mount radical prostatectomy sections at 40× magnification. The details of features generated from co-occurrence matrix are discussed in Chapter 4 Section 4.2.

Farjam et al. (2007) used k-means algorithm for clustering of local features from the histology images, and these features were used to separate the stroma and lumen from regions of nuclei. Altunbay et al. (2010) used the distribution of multiple gland components for the representations of different regions (such as nuclei or stroma) in histology images. The feature representing the tissue components in the images use a new set of structural features, and these structural features are: (i) average degree for: stromal-stromal edges, luminal-stromal edges, and luminal-nuclear edges; (ii) average clustering coefficient for: luminal components and stromal component; (iii) diameters for stromal-stromal edges, and luminal-stromal edges. In the same year, Nguyen et al. (2010) used both statistical and textural features to describe the local patterns in the tissue images, that study addressed three class problem with benign, grade 3 cancer and grade 4 cancer categories. The author used the SVM classification algorithm and the multilayer perceptron (MLP) algorithm to make a prediction based on the extracted features.

Akbar et al. (2015) introduced a method to separate the images into different parts (gland and background), and extract the local features from these parts. The features were used to train the machine learning algorithm in WEKA (Jagtap, 2013) (WEKA is a software implementing many machine learning methods). The extracted features included: the feature generated from grey-level co-occurrence matrix, first-order statistical features and second-order statistical features. First-order statistical features including mean, variance and entropy, and second-order statistical features include correlations, contrast and homogeneity. The comparison results based on the KNN algorithm, SVM algorithm and Bayesian Logistic Regression. A brief review of these classification methods is provided in Chapter 3 Section 3.1.

Rathore et al. (2013) introduced yet another method to segment the tissue in

histology images. The proposed method contains five steps, including pre-processing, feature extraction, feature selection, clustering of histology images, and post-processing. In the pre-processing step, enhancement of the contrast of the images was used to decrease the effect of the image variability due to differences in the staining process. Feature extraction and feature selection were employed to find the best features subsequently used by the classifier. Local Binary Pattern (LBP), Local ternary patterns (Tan and Triggs, 2010), and Haralick texture features (Haralick and Shanmugam, 1973) were used to describe the local information (the details of different versions of LBP features are discussed in Chapter 4, Section 4.4). The post-processing step is to remove noise in the images. The adopted evaluation measure was concerned only with the classification accuracy and the shape similarity was not estimated.

Fu et al. (2014) proposed a segmentation approach based on polar coordinates. This method starts to convert the histology images to polar space. The morphological gland boundary is transformed into a vertical periodic graph, which could be identified using a conditional random field. The weakness of this method is that it achieved excellent performance only for images with benign tissue. The following year, Cohen et al. (2015) proposed a method which combined the pixel-level classification and the active contour to solve the gland segmentation. In pixel-level classification, the first-order statistical features are extracted from different colour spaces, such as RGB, HSL and Lab. This method relied on a gland surrounded by an epithelial layer (refers to epithelial nuclei in benign and malignant tissue in Figure 1.2) that occurr dark in the image.

Ap et al. (2017) presented a method to detect the boundary epithelial cell of the gland, and then constructed the entire gland boundary. The described method employed the histogram feature and Haralick texture features (Haralick and Shanmugam, 1973) to train the random forest models, and to predict the gland objects in test images as thick or thin glands. The performance of that method achieved good results when compared with the top-10 ranked results reported by (Warwick.ac.uk, 2016a). The results of that method are worse than the results obtained by the random forest proposed in this thesis. The details of histogram features are discussed in Chapter 4, Section 4.3.

The above methods have limited capability of handling different shapes or image

staining and are sensitive to intensity and texture variations. Paul et al. (2016) introduced an approach which builds a new informative morphological scale space for gland segmentation. However, the evaluation measures for segmentation results shown in the paper (Paul et al., 2016) were only done on the images already used for training.

In recent years, many researchers have developed segmentation methods based on deep learning techniques. These approaches have been applied to gland segmentation and demonstrate excellent performance. In the MICCAI gland segmentation 2015 competition, the method from our group employed the modified LeNet5 architecture to learn the three categories –"inside gland", "gland boundary" and "outside of gland" from the training images. All the training images were normalised to have similar histogram distribution. The final segmentation results applied the post-processing using level set method applied to the LeNet-5 predicted probability maps (Sirinukunwattana et al., 2017).

Kainz et al. (2015) proposed a new method which employed two convolutional networks as the pixel classifiers. The input for these two convolutional neural networks was pre-processed by deconvolving the red channel in the original images. The final output is generated by a global segmentation based on weighted total variation. The segmentation results of the proposed method performed well in the images with benign tissue and those with malignant tissue.

The method proposed by Chen et al. (2016) achieved the best performance in MICCAI Gland Segmentation 2015 competition. The proposed method has combined the multi-level features representation with the fully convolutional network (FCN).

The same group developed and introduced a novel deep learning architecture (Graham et al., 2018), named 'minimal information loss dilate network', which combined the minimal information loss units, dilated residual units and traditional residual units. The novel architecture for this network is the minimal information loss units which include the training images downsampling into the residual units after max-pooling layer.

Li et al. (2016) introduced a method which combines deep learning features and hand-crafted features to train the SVM algorithm and uses this trained model to predict the tissue parts in test images. In the paper, the comparative results for different sizes of

the patches for the hand-crafted features and deep learning features, and the comparative results for different fusions of the hand-crafted and deep learning features are also represented.

Manivannan et al. (2018) introduced a hybrid method which also combines the hand-crafted features and deep learning features. The hand-crafted features used in the proposed method are root-SIFT, raw-pixel values, and multiresolution local binary patterns (Ojala et al., 2002) and deep learning features which were learnt by a modified FCN architecture.

The methods discussed above have been applied to gland segmentation. Some of them combined different types of features (i.e. LBP, histogram and Haralick texture feature) and machine learning techniques (i.e. random forest and SVM), and other methods have used deep learning architecture. The current approaches nowadays use deep learning and hand-crafted features.

2.3 Gland dataset

Chapter 2 Section 2.2 provided a brief review of gland segmentation methods. Most of the research time has been spent in classifying the benign tissue, and images with malignant tissue have also been investigated but not spend as much time as benign tissue (Gurcan et al., 2009). This research aims to find the best image representation for images with benign tissue and those with malignant tissue. In this research, the MICCAI 2015 gland database (Sirinukunwattana et al., 2015; Sirinukunwattana et al., 2017; Warwick.ac.uk, 2016b) is used because it contains two types of images: the images with benign tissue and those with malignant tissue. This dataset reflects the variation of the gland structure, shape and appearance and is more comprehensive when compared to datasets containing the images with benign tissue only. The MICCAI 2015 gland database is publicly available from (Warwick.ac.uk, 2016b).

Table 2.1 shows the details of the gland segmentation data used in this thesis. The dataset contains in total 165 histology images, divided into **Training Part**, **Testing Part A**, and **Testing Part B**. Different parts of the subset contain different number of the images, and the size of these images are also different. These are all detailed in Table 2.1. There

are 85 training gland images, divided into 37 images with benign tissue and 48 images with malignant tissue.

Table 2.1 The details of gland segmentation database (Sirinukunwattana et al., 2017)

Histological gland category	Number of images (Width × Height in pixel)		
	Training Part	Testing Part A	Testing Part B
Benign cases	$37\begin{cases}1\ (574\times433)\\1\ (589\times453)\\35\ (775\times522)\end{cases}$	$33\begin{cases}1\ (574\times433)\\4\ (589\times453)\\28\ (775\times522)\end{cases}$	4 (775 × 522)
Malignant cases	$48\begin{cases}1\ (567\times430)\\3\ (589\times453)\\44\ (775\times522)\end{cases}$	$27\begin{cases}1\ (578\times433)\\2\ (581\times442)\\24\ (775\times522)\end{cases}$	16 (775 × 522)

Figure 2.3 illustrates sample images from the gland dataset and the corresponding ground truth, which uses different colours to represent different gland objects.

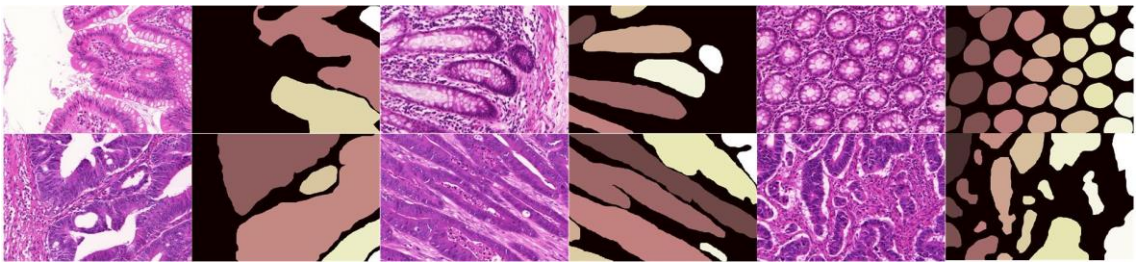


Figure 2.3 The example of images and corresponding ground truth from gland dataset (Nasir, 2015). The images in the top row are the images with benign tissue. The images in the bottom row are the images with malignant tissue. The corresponding ground truth is shown on the right of the original images

The same database was also used in various other studies (Yang et al., 2017; Ravishankar et al., 2017). Different evaluation measures were employed to estimate the results in that study. More measures have been used to evaluate the results reported in this thesis, the details of which are provided in Chapter 6.

2.4 Summary

This chapter provides a brief description of processes involved in generation of the histology images. This has been included so the variation present in the histology images could be attributed to different processes involved in the preparation of tissue specimens. Most of the studies focused on the identification of benign specimens rather than malignant tissue (Gurcan et al., 2009). Finally, the details and structure of

MICCAI 2015 Gland Database are discussed. This database has been selected for development and validation of the methods described in the rest of this thesis.

The following chapter will discuss the mathematical basis of the existing pixel-level classifier applied to gland segmentation.

Chapter 3

Mathematical foundation of pixel-level classifier

The classification methods can be subdivided into approaches using supervised and unsupervised learning techniques. Both of those are used in this work but for different purposes. Many machine learning based methods (i.e SVM and random forest) have been applied to image segmentation, whereas not machine learning based methods (i.e. region-growing method and watershed method) are also used in the same problem (image segmentation). This research focuses on supervised machine learning based methods. In this case, a brief review of machine learning based methods is described at the beginning of this chapter. Random forest techniques are employed as the primary classifier applied in gland segmentation. It is for several reasons. They are inherently designed to solve multiclass classification problems. Furthermore, the forest models are not sensitive to noise and outliers but, and more importantly, they can achieve good results in both image classification and segmentation with efficient implementations. This work not only uses random forest as the pixel-level classifier, but also employs deep learning architectures for feature extraction and image-based classification. There are also many kinds of unsupervised learning techniques often applied to these problems. In this work, the K-means algorithm is used as a validation tool to estimate the discriminative properties of extracted features.

This chapter is organised as follows. Before discussing decision trees and random forest, a review of different machine learning techniques is provided. Subsequently, decision trees and random forest are described in more detail. The applications of different forest models are also discussed. To find the best random forest model for gland segmentation, the comparative results of these forest models are tested on different subsets of the UCI database (Dua and Karra Taniskidou, 2017). Finally, an unsupervised method, K-means clustering, is used to evaluate the discriminative properties of different features.

3.1 Machine learning

This section reviews different machine learning techniques. Increasingly more research is being done into machine learning. It has become a popular field over the last two decades. Machine learning means that learning behaviours are fulfilled by simulation analysis that can be conducted by the machine to change performance on its own. Both classification and regression have used machine learning based methods. Classification tasks are to classify the unknown data to a target class, which is known from training data. Regression tasks are to predict discrete or continuous value. Clustering tasks are to group the data into several groups based on similarity of the data.

Pattern recognition is related to machine learning. Machine learning techniques have been used to deal with pattern recognition tasks. Image segmentation is pixel-wise classification. The reason for pattern recognition being related to image classification and segmentation is that machine learning methods are extracting the patterns from images. Patterns are the most important criterion in image segmentation and classification tasks. Many machine learning techniques applied to pattern recognition are supervised learning (i.e. SVM, random forest and decision trees), and there are also some unsupervised learning methods (i.e. K-means and Fuzzy C-means clustering). For supervised learning, each problem requires a set of input data samples $\mathbf{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$ and corresponding labels $\mathbf{l} = \{\boldsymbol{l_1}, \boldsymbol{l_2}, \cdots, \boldsymbol{l_n}\}$. Unsupervised learning needs only a set of input samples but not labels. This research focuses on the supervised learning algorithm, although unsupervised techniques are also used. There are many approaches to machine learning algorithms, including Support Vector Machine (SVM), Adaboost, decision tree models, random forest techniques and deep learning algorithms.

The SVM algorithm was introduced by Cortes and Vapnik (1995) to solve binary classification problems. Although it can solve multiclass classification problems, the kernel must be determined, which is the hardest open-ended question for SVM algorithms.

Freund et al. (2003) introduced an improved version of Adaboost algorithm, whose output is formed by the superposition of weighted weak classifiers. The most prominent shortcoming of the Adaboost algorithm is that it could be sensitive to both noise and

outliers (Maclin and Opitz, 1997).

Decision tree models classify uncertain data belonging to one category. There are many popular decision tree approaches, including ID3 (Iterative Dichotomiser 3), C4.5 algorithm, and CART (Classification and Regression Tree). Quinlan (1986) introduced the ID3 algorithm, based on entropy and information gain. The C4.5, a modified version of ID3, algorithm was also introduced by Quinlan (1993). However, unlike ID3, the C4.5 algorithm could solve both continuous and discrete problems, and also solve the overfitting problem by pruning the entirely constructed trees. The significant drawback of this method is that there are some empty branches in the decision tree model. CART was introduced by Leo Breiman (2017), and the tree model is constructed based on the Gini index (or Gini impurity). One of the advantages of the CART model is that it consists of binary splitting child nodes rather than multiple child nodes. These three tree models all employ the axis-aligned weak learner, while Murthy et al. (1994) introduced a new type of decision tree consisting of the oblique weak learner.

Random forest techniques are machine learning approaches commonly applied in computer vision applications. Unlike the Adaboost algorithm (Freund and Schapire, 1997), which is another machine learning approach used in computer vision, the random forest can handle both noise and outliers in the training data (Ross and Kelleher,2013). The initial purpose of the random forest was to solve the multi-class classification problem, but over the last decade, forest techniques have been developed and used in many applications, including image classification, image segmentation and object tracking. Menze et al. (2011) introduced a decision tree model with oblique weak learners. The comparative classification results indicate that the oblique decision tree are better than Linear Machine Decision Tree (Brodley and Utgoff, 1992) and Simulated Annealing of Decision Trees (Heath et al., 1993).

Deep learning techniques have recently become one of the most broadly used methods for both image segmentation and image classification. LeNet5 was the first deep learning architecture, used for handwriting recognition application achieving excellent performance (LeCun et al., 1998). Although this neural network is shallow compared with more recent deep learning approaches, most of the deep learning

techniques applied to images are based on this architecture. In 2012, Krizhevsky et al. introduced a deeper network architecture for image classification (so called AlexNet) which outperformed other methdods on the ImageNet Challenge. In 2014, Szegedy et al. introduced a deep 22-layer architecture, known as GoogleNet, helping to reduce computing complexity using the inception models. A year later, a deep neural decision forest combining a deep neural network and random forest technique was introduced and achieved the best image classification performance on the ImageNet database (Kontschieder et al., 2015).

Many approaches applied in image segmentation are based on a sliding window approach. The sliding window uses a fixed moving window to capture local patterns in the training images. Long et al. (2015) introduced an end-to-end deep learning architecture, the fully convolutional network (FCN), and achieved excellent performance in semantic segmentation. One of the significant benefits of this method is that the size of the output is the same as the training images. In the same year, Ronneberger et al. (2015) introduced a new deep learning architecture so called U-Net. This model achieved the best performance in the ISBI challenge on cell tracking that year.

3.2 Decision trees

Before demonstrating the details of forest models (Chapter 3 Section 3.3), a brief description of decision trees is provided. The reason for introducing decision trees before random forest is that random forest consists of hundred or even thousands of decision trees. It is better to understand the basis of decision trees before learning the foundation of random forest. Random forest is the main classifier used in the proposed method (segmentation with pre-classification method, it is discussed in Chapter 5, Section 5.4) in this work. As a stablished algorithm in machine learning, the decision tree model consists of nodes and directed edges. A simple example of the tree model is shown in Figure 3.1. A decision tree is composed of three types of nodes: the root node, internal nodes (or splitting nodes), and terminal nodes (or leaf nodes). The root node stores the input of the decision tree, the internal nodes store the functions that splits the input data, and the terminal nodes are the output of the decision tree.

Figure 3.1.a depicts a simple structure of the decision tree model. The role of grass and desert images is to explain the path of different types of images (grass or desert classes) passing through the decision tree. Figure 3.1.b shows an example of classification process of two types of input images passing through the trees. The input images go through each level of the tree until they reach the terminal nodes. The blue circles indicate the nodes selected by the decision tree, and the direction of the blue arrows indicates the path of the images passing through the tree. The yellow and green squares in Figure 3.1.b represent the desert and the grass image classes. The path of desert images passing through the tree is shown in the right part of Figure 3.1 and the left part demonstrates the path of grass images.

Different splitting criteria, different thresholding approaches, and different weak learners are employed to build the different decision tree models, as explained in the following sections. The role of different splitting criteria is different way to split the samples in each node (except the terminal nodes or leaf nodes) in decision tree. The role of thresholding is different methods to generate the threshold in order to separate the samples. The role of weak learners is to describe different splitting methods (i.e. horizontal, vertical and oblique).

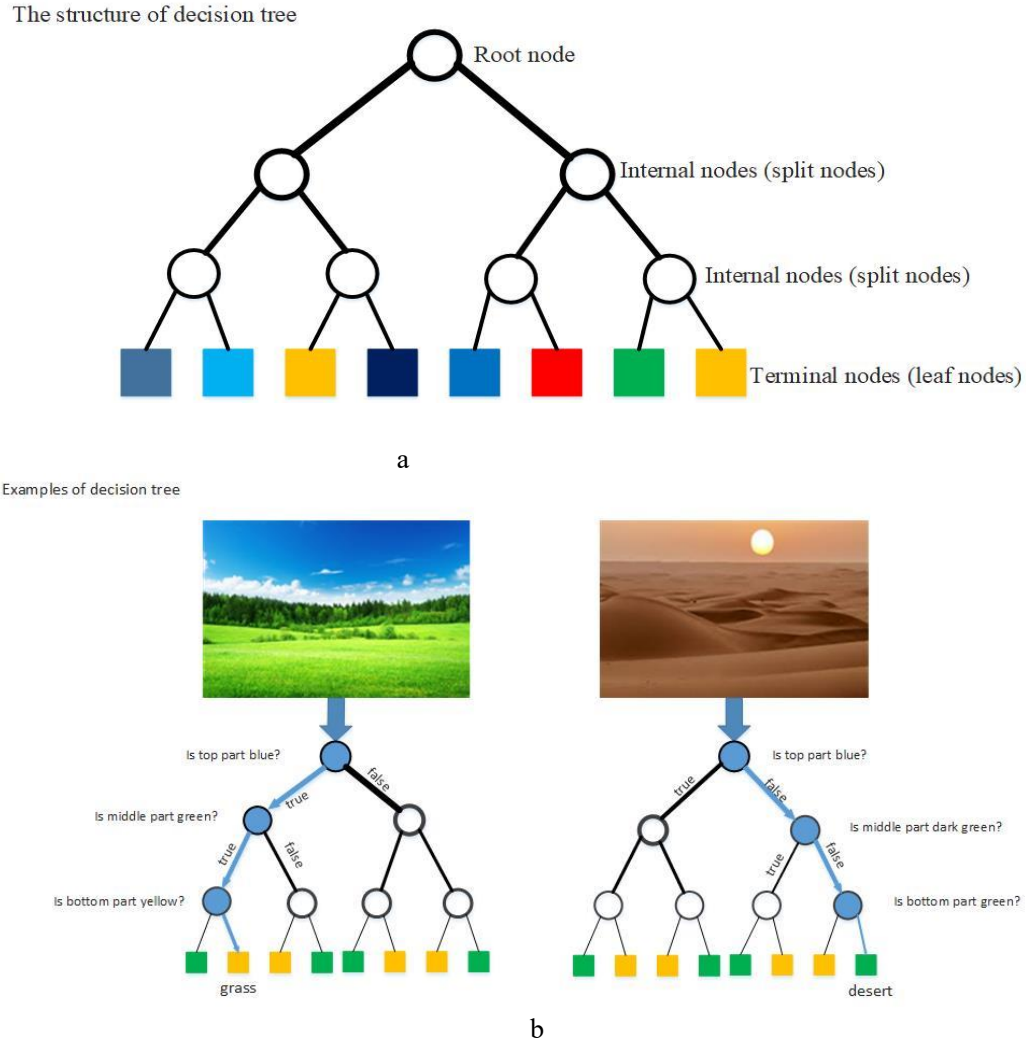


Figure 3.1 Decision tree. a) The simple structure of the decision tree model. b) The progress of different types of image through the decision tree model.

3.2.1 Splitting criteria

Three splitting criteria are widely used when building decision tree models: information gain, gain ratio and Gini index (or Gini impurity).

Entropy and Information gain

In the process of classification, two fundamental concepts of information theory are useful in the construction of the decision tree: entropy and information gain. The former is deployed to demonstrate the impurity of variables in training samples, and its mathematical expression is defined by (Quinlan, 1986):

$$H(\boldsymbol{x}) = -\sum_{c \in \boldsymbol{C}} P_{(c)} \cdot \log P_{(c)} \tag{3.1}$$

where $H(\boldsymbol{x})$ represents the entropy of variable $\boldsymbol{x}$, c is a class index, $\boldsymbol{C}$ is the number of classes represented in the database, and $P_{(c)}$ indicates the probability of class c.

In decision trees, information gain represents the entropy differences before and after data is split in the tree node. In other words, an increase of the information gain indicates a higher node data "purity " in the next level of the decision tree. This concept is a well-known criterion when building the ID3 model, and can be described by:

$$I = H(\mathbf{S}) - \sum_{\boldsymbol{S}^s \in \{\boldsymbol{S}^L, \boldsymbol{S}^R\}} \frac{|\boldsymbol{S}^s|}{|\boldsymbol{S}|} H(\boldsymbol{S}^s) \tag{3.2}$$

where $\mathrm{H}(\mathbf{S})$ indicates the entropy of the data set $\mathbf{S}$, $\boldsymbol{S}^s$ describes the subsets of the entire database $\mathbf{S}$; in the tree model it indicates the left/right child nodes ($\boldsymbol{S}^L$ *or* $\boldsymbol{S}^R$) of the parent nodes, and $|\cdot|$ demonstrates the cardinality of the corresponding data subset.

Gain ratio

The gain ratio is a modified version of information gain and is a well-known concept in building the C4.5 (one of decision tree models) tree model (Quinlan, 1993). The mathematical definition of this concept is defined by (Quinlan, 1993):

$$Gain\ ratio = \frac{Info\ gain}{split\ info} \tag{3.3}$$

where $Info\ gain$ is the information gain I (see equation 3.2), and $split\ info$ is the splitting information of the left and right child node. The mathematical expression is defined by (Quinlan, 1993):

$$split\ info = \sum_{\boldsymbol{S}^s \in \{\boldsymbol{S}^L, \boldsymbol{S}^R\}} \frac{|\boldsymbol{S}^s|}{|\boldsymbol{S}|} H(\boldsymbol{S}) \tag{3.4}$$

where $H(\mathbf{S})$, $\boldsymbol{S}^L, \boldsymbol{S}^R$, $|\cdot|$ have already been defined in equation 3.2.

Gini impurity (or Gini index)

Gini index is a measure of income inequality introduced by Gini (1912). This term has been extended used in decision tree by Breiman which has been widely applied in models built by decision trees. It can be utilised to illustrate a degree of uncertainty in the distribution of the feature space as demonstrated in CART by Breiman (2017). The mathematical expression of Gini impurity is defined by (Breiman,2017) :

$$G(\mathbf{T}) = 1 - \sum_{i=1}^{n} p_i^2 \tag{3.5}$$

where $G(\mathbf{T})$ indicates the Gini impurity of the data set $\mathbf{T}$. n indicates the number of classes in the database. p_i is the probability of the corresponding class i.

Figure 3.2 on page 49 in this section demonstrates an example to illustrate differences between the three splitting criteria concepts described above.

To explain the proposed methods, suppose that there are two sets of samples drawn from Gaussian distributions, as has shown in Figure 3.2.a. This task aims to separate the samples into either blue or red class. Figure 3.2.b demonstrates the probability of these two categories occurring in this dataset. There are many splitting methods, but in this example, it uses two splitting ways to separate the samples shown in Figure 3.2.a. Figures 3.2.c and 3.2.d illustrate the different splitting methods for separating the same data, with horizontal and vertical splitting respectively.

From the probability shown in Figures 3.2.c and 3.2.d, it is easy to observe that the horizontal splitting method is better than the vertical method. Table 3.1 shows the values of the described splitting criteria reflect these two splitting methods. Based on this example, the best performance of classification is provided by choosing the minimum value of the Gini index or the maximum value of the information gain and gain ratio.

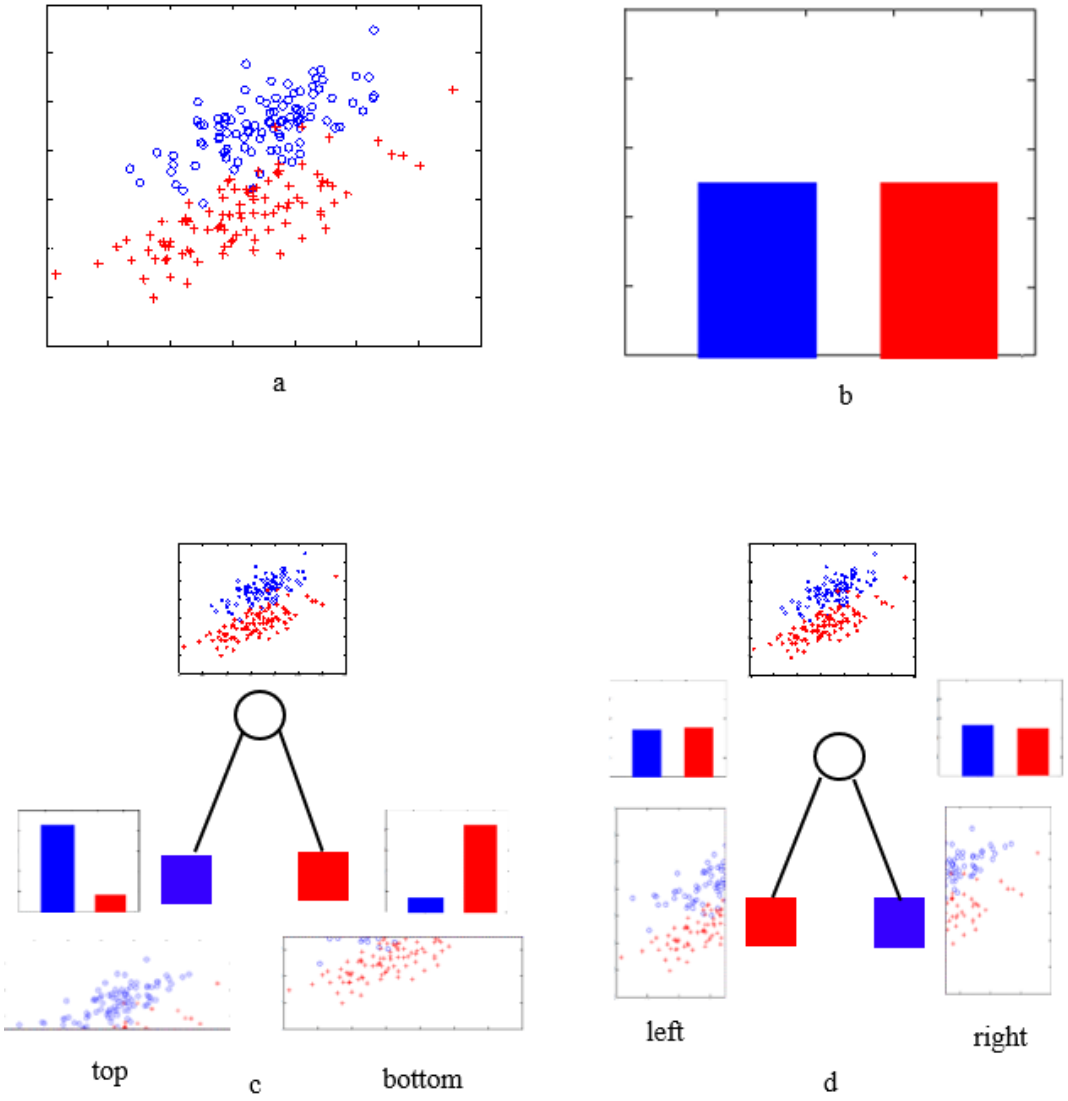


Figure 3.2 Different splitting approaches. a) The sample points drawn from Gaussian distributions. b) The probability of each class in the training data. c) The horizontal splitting approach for training samples. d) The vertical splitting method for training samples.

Table 3.1 The values of different splitting criteria based on the different splitting methods

Different methods	Gini index	Information gain	Gain ratio
Horizontal	0.55	0.68	0.45
Vertical	0.72	0.49	0.32

3.2.2 Thresholding

In random forest techniques, the correlations between trees play an essential role. If these correlations are too high, the behaviours of the forest models are similar to those of a single decision tree. Figure 3.3.a is an example of tree models with high correlations and low randomness, and the model with low correlations and high randomness is

shown in Figure 3.3.b. In practical applications, trees with low correlations and high randomness (Figure 3.3.b) are preferred.

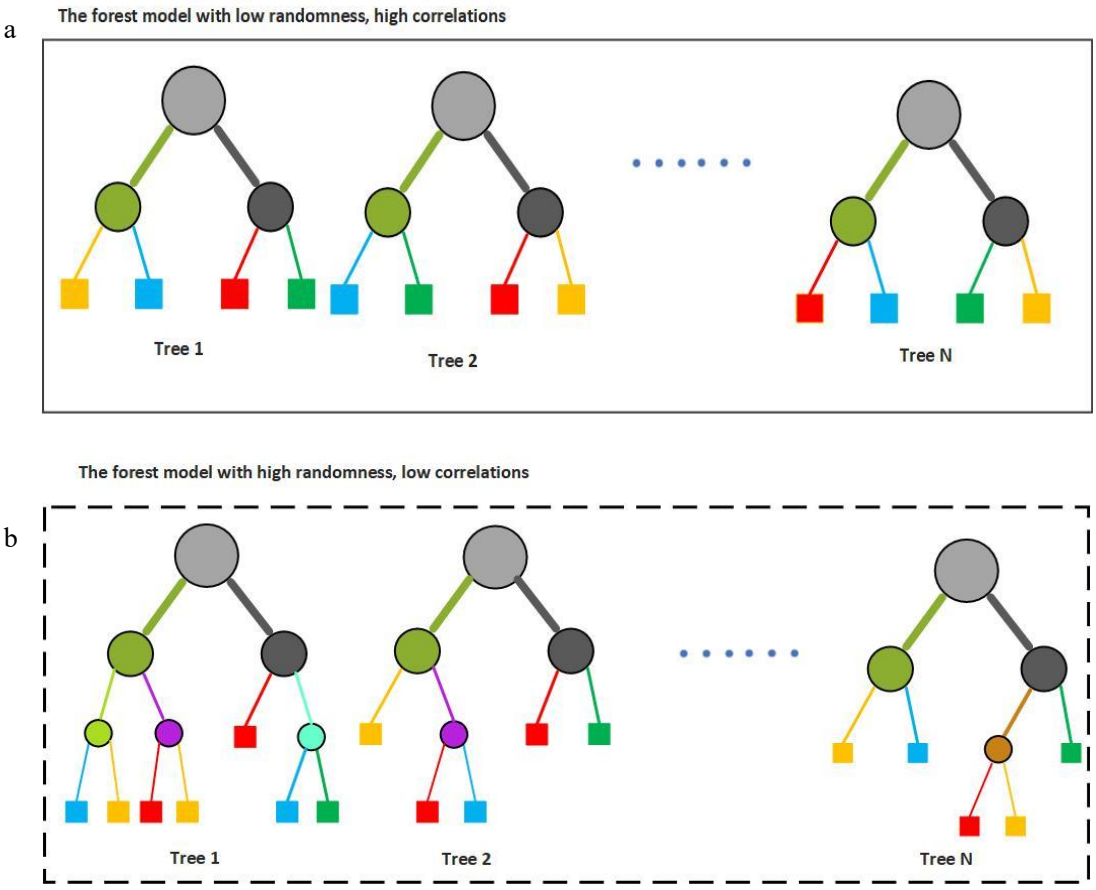


Figure 3.3 Examples of different forest models. a) The forest model with low randomness and high correlations. b) The forest architecture with low correlations and high randomness (Criminisi et al., 2013):

In the decision tree approaches, the mid-point in one dimension of the feature vector is selected as the threshold to split the input samples in the internal nodes. This approach is known as the mid-point thresholding approach; it is not an issue when choosing to use the decision tree as the classifier for the problem. However, if this method is employed to build forest models, the trees could be correlated. The randomised node optimisation approach is therefore employed in constructing the decision trees in order to decrease this correlation. The extremely randomised node optimisation is employed to build the trees to further reduce the tree correlations in the forest approaches. These three thresholding techniques are explained below.

Mid-point thresholding

In C4.5 and CART models, the training samples $\boldsymbol{X} = (x_1, x_2, \cdots, x_n)$ passing through the splitting nodes are separated into two child nodes, left and right. The threshold is selected as the mid-point between two adjacent points in the training samples. The mathematical expression of mid-point thresholding is defined by:

$$T_i = \frac{x_i + x_{i+1}}{2} \ (i = 1, \cdots, n-1) \tag{3.6}$$

where T_i is the threshold in the internal node in the forest model. x_i are the values of training samples at the same internal node of the model. n is the number of training samples.

Randomised thresholding approach

The mid-point thresholding approaches employed in C4.5 and CART models cannot increase the randomness in the forest models.

Randomised thresholding employs random values between two adjacent variables as the threshold to split the samples in the internal nodes of the tree models. Suppose that the training samples in an internal node of the tree model are $\boldsymbol{X} = (x_1, x_2, \cdots, x_n)$. The selected threshold in the randomised thresholding approach is defined by (Criminisi et al., 2013):

$$T_i \epsilon (x_i, x_{i+1}) \ (i = 1, \cdots, n-1) \tag{3.7}$$

where T_i is the threshold in that node, randomly selected from the (x_i, x_{i+1}) interval

The efficiency of building the decision tree using mid-point thresholding is the same as for the randomised thresholding approach. For example, in these two approaches, if the training data contains 1,000 variables, the number of thresholds for each dimension is 999.

Extremely randomised approach

The above two thresholding methods are not effective when building the decision

tree model. The extremely randomised approach is employed to construct the trees to increase the efficiency and decrease the tree correlations further. This approach randomly selects a number from uniform distribution between the maximum and minimum values. The mathematical expression is defined by (Geurts et al., 2006):

$$T_i \in (x_{min}, x_{max}), \quad (x_{min} = \min(\boldsymbol{X}), x_{max} = \max(\boldsymbol{X}) \ , \ i = 1,2,\cdots,m) \quad (3.8)$$

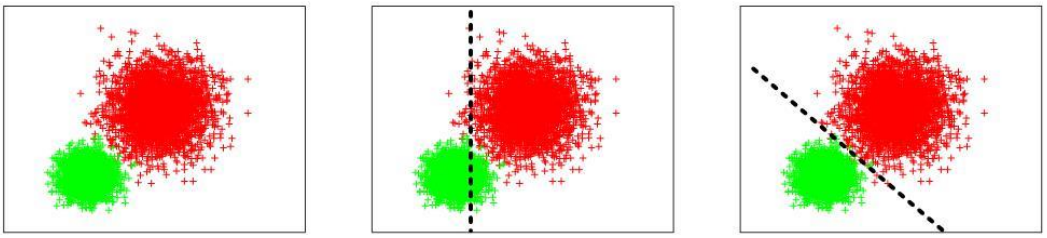
where $\boldsymbol{X}$ indicates the available training samples. x_{min} is the minimum value of the training samples $\boldsymbol{X}$ in the selected dimension, and x_{max} is the maximum value of the training samples in the same dimension, T_i is the randomly selected threshold with a condition defined in equation (3.8) and the total number of thresholds (defined by the user) is $m \ (m < n)$.

This approach is more efficient than the two previous techniques. For example, if the number of samples in one of the dimensions of the feature vector is 1000, the number of the thresholds for either mid-point thresholding or randomised mode is 999. However, in the extremely randomised thresholding approach, the number of the thresholds is determined by the user; if it is, says 50, the computation time for building the tree models will be much less than for the other two thresholding approaches.

3.2.3 Weak learners

The decision tree models ID3, C4.5 and CART all employ the axis-aligned weak learner. Another type is the oblique weak learner (Murthy et al., 1994). Both types are discussed below.

The classification results originated from using different types of weak learner could be different. Figure 3.4 indicates two different types of weak learner used to split the linear separable data, drawn from Gaussian distributions.



a b c

Figure 3.4 Examples of different weak learners. a) Data samples drawn from two Gaussian distributions. b) The splitting results are for using the decision tree with axis-aligned weak learner. c) The results for using the decision tree with oblique weak learner.

In Figure 3.4, samples drawn from different distributions are shown using different colours. Figures 3.4.b and 3.4.c represent the segmentation performance of the decision tree with axis-aligned and oblique weak learners respectively. The axis-aligned weak learners are a particular form of the oblique weak learner. Their splitting technique uses a line parallel to the x-axis or y-axis, and that for oblique weak learners employs lines at any orientation.

3.2.4 Termination conditions

Overfitting is one of the problems when using decision trees as classifier. To solve the overfitting problem, C4.5 (Quinlan, 1986) and CART (Breiman, 2017) employed different pruning techniques. The pruning technique cut the branches off the decision tree after it is fully grown.

Another technique to avoid the overfitting problem is to use pre-defined parameters in building the decision tree; if the tree model fits one of the parameters, the tree will stop growing. Many conditions have an impact on the growth of the decision tree. For example, it will stop growing when it meets the maximum depth of the model; when the nodes in the model contain too few data points; or when the samples in one node belong to one specific category of the data.

3.3 Random forest

The above discussion demonstrates many approaches for construction of decision trees. These decision tree models can be used in many practical applications, in which

the data could be noisy or contain outliers, to which the decision trees can be sensitive. To solve this problem, Breiman (2001) introduced the random forest technique, which combined the ensemble learning (this term is explained on page 19) and CART approaches (Breiman, 2017). Random forest models are classifiers which consist of hundreds or even thousands of decision trees to accomplish classification. The decision of the random forest models is voted by all the decision trees in the same model, and the class with the most votes is considered to be decision of the random forest model.

As mentioned in the previous section, the purpose and foremost advantage of forest models are that it fits well to multiclass classification problems. For gland segmentation, two and three target classes classification are designed in order to find the best way to describe morphological structure of gland objects for benign or malignant case. The details of two and three target classes in histology images have been discussed in Chapter 5, Sections 5.4.2.1 and 5.4.2.2.

Forest models can handle large databases. The technique has developed significantly over the last two decades, and these models have been applied to classification, regression, density estimation and semi-supervised learning problems. There are two main techniques for building forest models: bagging and random subspace.

3.3.1 Bagging

One of the approaches employed to construct forest model is Bootstrap Aggregation (Bagging) method introduced by Efron (1992). The algorithm for the Bagging method could be summarised as follow:

- Input: number of decision trees, M, in forest models; training data samples $\boldsymbol{X}$;
- Step 1: Create M random subsets of the training data from $\boldsymbol{X}$ using random selection with replacement method
- Step 2: For each random subset train corresponding decision tree in the forest model, with each decision tree fully grown without pruning
- Step 3: Determine the predictions for each decision tree, and the output of the forest model is the classes with the most votes from M decision trees.

The random selection of the data subsets and the subsequent voting scheme solves

the overfitting problem, present when using the decision trees. When using Bagging, the parameters to be set by a user include number of trees and number of data samples used.

3.3.2 Random subspace

Another technique used in construction of random forests is so called random subspace, which modifies the training data in the feature space. The method was demonstrated in detail in (Barandiaran, 1998), its advantages being that the computational time decreased significantly. The following briefly summarised the algorithm for random subspace forest construction:

- Input: number of the decision trees, M, in the forest model; training data samples $\boldsymbol{X}$ with $\boldsymbol{D}$ representing number of features;
- Step 1: For each decision tree, choose $\boldsymbol{X_i}$ ($\boldsymbol{X_i} \in \boldsymbol{X}$) samples from $\boldsymbol{X}$
- Step 2: Create a training data set by choosing $\boldsymbol{D_i}$ features from $\boldsymbol{D}$ ($\boldsymbol{D_i} < \boldsymbol{D}$) using sampling with replacement method, use the created training sets to train corresponding trees in the forest model with each decision tree fully grown without pruning.
- Step 3: Determine the predictions for each decision tree, and the output of the forest model is the classes with the most votes from M decision trees

3.3.3 Applications

The above sections briefly demonstrated the concepts of the decision tree and random forest models. In this part, applications of the random forest techniques are discussed, specifically image classification, image segmentation and regression.

Decision forest models have been applied to skin detection (Khan et al., 2010), and they used several classifiers in the tasks in order to find the best performance. The comparative results (has given by the experiment results) for this task are based on SVM, Adaboost, Naïve Bayes, Bayesian networks, RBF networks and random forest. Although based on visual inspection the performance of AdaBoost is similar to the results of the random forest technique, the classification accuracy, for the forest is 87.7%, what is better than AdaBoost's 79% accuracy. The number of trees applied in this problem was

ten, which is a small number. If the number of trees was increased in the forest model, a higher accuracy might have been reached.

Forest techniques are also used in face recognition (Kremic and Subasi, 2016); the performance of the random forest outperformed the results of the SVM algorithm. For the images used in that research, the background was simple and the face was displayed in the central part of the image. If the input images were changed to have more complex backgrounds, or input images contain multiple faces, the performance of the algorithm decreased.

This model has also been applied in pedestrian-detection problems, and the estimation results are based on the decision forest and linear SVM algorithm. González et al. (2015) demonstrated the comparative results of random forest and linear SVM algorithm based on HOG and LBP features. The HOG outperformed LBP when using the random forest model. Again, the databased used for that problem was relatively simple.

Random forest models have also been applied in regression applications, for example in solving protein fold prediction problems (Dehzangi et al., 2010). The study demonstrated that the random forest model is able to selected the best features from a very large population of features, when training with the Gini impurity criterion.

Fernández et al. (2015) used the HOG feature to describe facial images and to train the forest model to solve age prediction problem. However, no other features were tested, and maybe HOG is not the best feature to represent the local characteristics for those images. For further improvement, more types of features would have to be employed to represent facial images.

Regression forest is another forest model which has been applied in head pose estimation and has achieved excellent performance. Zhu et al. (2013) used several features to represent the local patterns of head images. The HOG feature turned out to be the best for the head pose estimation. The study also applied regression trees with four different kinds of data and compared the final results in order to find the best features to represent the essential patterns in the original images.

Rotation forest is a methodology which combines PCA (Principal Component Analysis) and random forest techniques. Rodriguez et al. (2006) demonstrated the

comparative results of rotation forest and other ensemble approaches, in which the rotation forest method outperformed other classification techniques. Several subsets of UCI databases were employed in the tests.

Mapping forest is an approach that performed well in facial expression recognition (Jampour et al., 2018). This technique achieved better performance than using linear mapping approaches, and provided computationally efficient implementation.

This section discussed the applications of the random forest, the forest model has been applied in many tasks, including image classification, object detection and image recognition. The commonality between these applications and gland segmentation is that all these tasks are treated as classification tasks. The reason for treating gland segmentation as classification tasks is that gland segmentation is a pixel-wise classification for each histology image.

3.4 Comparison of different forest models

From the above descriptions, different splitting criteria, different thresholding methods and different weak learners are employed to build different decision trees. Different random forest models consist of different decision tree architectures. Even for the same database, different random forest models will perform differently. It is necessary to choose one of the best forest models as the base model for the gland segmentation task.

Subsets of the UCI database (Dua and Karra Taniskidou, 2017), for which feature vectors are provided directly, are used to compare different forest models. The details of the datasets used to estimate performance of the tested forest are shown in Table 3.2.

Table 3.2 The details of the UCI datasets used to test different forest models

Name of database	Number of samples	Number of features
Congressional voting	435	16
Liver disorders	345	7
Connectionist Bench (Sonar, Mines vs. Rocks)	208	60
Ionosphere	351	34
Tic-Tac-Toe Endgame	958	9

3.4.1 Evaluation measures for the experiments

Each dataset in Table 3.2 represents a binary classification problem (this term is explained on page 19). The evaluation measures to validate the classification performance of different random forest models are: F1 score, precision, recall and classification accuracy.

In the binary classification problems, F1 score is a measure metric which is used to measure test's accuracy, and it is the harmonic average of precision and recall, it is defined by (Powers, 2011):

$$\mathrm{F1} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3.9}$$

Precision is a ratio of the relevant instances over the retrieved instances, and recall is the ratio of returned relevant instances over the total number of relevant instances. The mathematical expressions for these measures are given by (Powers, 2011):

$$Precision = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \tag{3.10}$$

$$Recall = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{3.11}$$

where $\boldsymbol{TP}$ represents true positive instances, $\boldsymbol{FP}$ indicates false positive instances, and $\boldsymbol{FN}$ represents false negative instances.

In classification problems, classification accuracy validates the performance of the classifier, it is defined by (Powers, 2011):

$$\mathrm{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+FP+FN+TN}} \tag{3.12}$$

where $\boldsymbol{TP}$ and $\boldsymbol{TN}$ are the two correctly predicted categories after classification; the denominator is the total number of instances in the entire training dataset.

3.4.2 Experimental setup and results

This section provides the details of the experiments for different types of the forest on UCI datasets, and the purpose of these experiments is to find one suitable forest

model applied to gland segmentation. All the experiments were done by using MATLAB 2015b, and the library with different forest models was custom built. All the forest consists of 10 decision trees, due to of the relative small size of the used datasets.

75% of the data randomly sampled from the datasets are used to train the classifier and the rest of the data are used to validate its performance. The average values of the evaluation metrics computed for 10 cross-validation experiments, for each of the datasets listed in Table 3.2 is used. To investigate the performance of different random forest models, this section compares following forest models:

R-O-GR: Randomised thresholding, with oblique weak learners and Gini ratio splitting criterion.

E-O-GR: Extremely randomised thresholding, with oblique weak learners and Gain ratio splitting criterion.

M-O-GR: Mid-point thresholding, with oblique weak learners and Gain ratio splitting criterion.

R-O-GI: Randomised thresholding, with oblique weak learners and Gini index splitting criterion.

E-O-GI: Extremely randomised thresholding, with oblique weak learners and Gini index splitting criterion.

M-O-GI: Mid-point thresholding, with oblique weak learners and Gini index splitting criterion.

R-O-IG: Randomised thresholding, with oblique weak learners and information gain splitting criterion.

E-O-IG: Extremely randomised thresholding, with oblique weak learners and information gain splitting criterion.

M-O-IG: Mid-point threshold, with oblique weak learners and information gain splitting criterion.

R-A-GR: Random thresholding, with axis-aligned weak learners and Gini ratio splitting criterion.

E-A-GR: Extremely randomised thresholding, with axis-aligned weak learners and Gain ratio splitting criterion.

M-A-GR: Mid-point thresholding, with axis-aligned weak learners and Gain ratio splitting criterion.

R-A-GI: Randomised thresholding, with axis-aligned weak learners and Gini index splitting criterion.

E-A-GI: Extremely randomised thresholding, with axis-aligned weak learners and Gini index splitting criterion.

M-A-GI: Mid-point thresholding, with axis-aligned weak learners and Gini index splitting criterion.

R-A-IG: Randomised thresholding, with axis-aligned weak learners and information gain splitting criterion.

E-A-IG: Extremely randomised thresholding, with axis-aligned weak learners and information gain splitting criterion.

M-A-IG: Mid-point thresholding, with axis-aligned weak learners and information gain splitting criterion.

The quantitative results of different forest models on **Liver** data are shown in Figures 3.5. The quantitative results for other datasets are shown in Appendix A. The reason for only providing the results of this datasets is that the results of different random forest models are similar to these two datasets. Based on the results shown here (see Figure 3.5) and in Appendix A, different forest models performed almost at the same level on these data, and the conclusion is that, for the same database, the forest with different splitting criteria, different thersholding and different types of weak learners is not affecting significantly the performance of forest models. Based on these, the mid-point threshold with axis-aligned weak learners and Gini impurity splitting criterion (**M-A-GI**) random forest method was selected for pixel level classification for gland segmentation.

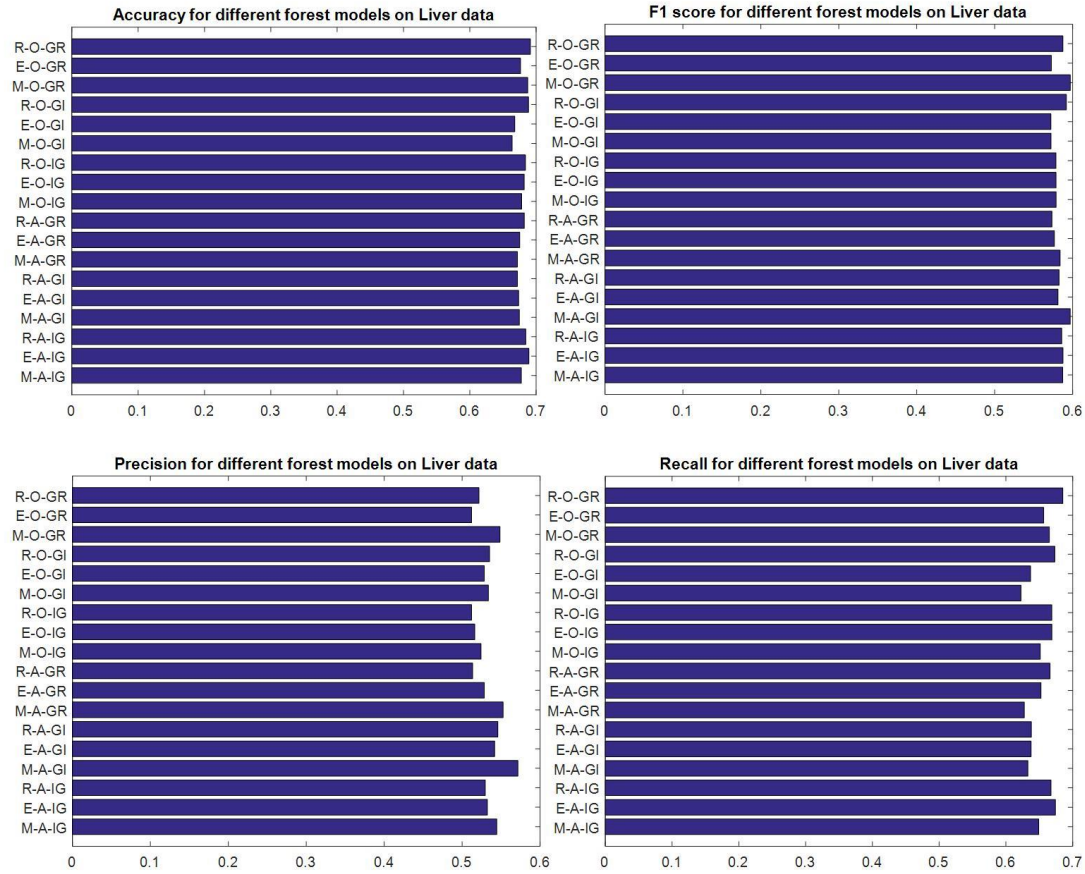


Figure 3.5 Validation parameters of different forest models on **Liver** database

3.5 K-means algorithm

The techniques described in previous sections are all examples of supervised learning. For classification, both supervised and unsupervised learning algorithms can be used to solve related problems, but unlike the supervised learning techniques, the training phase of unsupervised learning only needs the training samples without the ground truth. In this research, K-means algorithm (Lloyd, 1982) is used to estimate the discriminative properties of different features. There are many unsupervised learning methods which can be replaced and used to estimate the discriminative properties of these features in the work, such as Fuzzy C-means (Bezdek et al., 1984). The reason for choosing K-means is that this algorithm has been widely used, and this method can be easily interpreted.

The evaluation process is detailed in the following chapter. The K-means algorithm aims to divide a set of training samples $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ into $k(k \leq n)$ clusters.

The goal is to predict k centroids and labels c for each data point. The algorithm is summarised as:

- Input : Training samples $\boldsymbol{X}$ and k clusters.
- Step 1: Initialise the cluster centroids $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k$ randomly.
- Step 2: Associate each observation $\boldsymbol{x}_i$ in the training samples with the nearest centroid. This step will allocate the samples into k clusters.
- Step 3: Recalculate the cluster centroids $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k$
- Step 4: Repeat Steps 2 and 3 until convergence.
- Output: $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k$ cluster centroids (centroids are the centre of the cluster)

3.6 Summary

In this chapter, various machine learning techniques were first briefly discussed. Subsequently, decision tree methods were then introduced, including splitting criteria, types of weak learners and thresholding methods. Different strategies for avoiding overfitting problems in decision trees were also introduced. The methods used to build random forest models were then presented, and applications of random forest techniques discussed in Chapter 3 Section 3.3.3. Different datasets were used to test the different random forest models, concluding that the forests perform at a somewhat similar level for most data. Based on the performed experiments and taking into account relatively low computation complexity, random forest method with axis-aligned weak learners, mid-point thresholding and Gini impurity splitting criterion (**M-A-GI**) has been selected for the experiments on the gland data. The commonality between the subsets of UCI (University of California Irvine) database is that they are all real data after the feature extraction, and these data could be treated as classification tasks after feature extraction. The experiments aim to find the best random forest model for classification tasks. Gland segmentation could also be treated as a classification task after extracting the local patterns.

Finally, the unsupervised learning method, K-means clustering, was introduced in Chapter 3 section 3.5 as it is used to evaluate the discriminative properties of input

features in Chapter 4.

In the following chapter, the various features used to describe local patterns in histology images are described.

Chapter 4

Feature extraction

In this chapter, different feature extraction methods employed are discussed, together with the motivation for choosing these methods. After discussing their background, the feature extraction methods used in this work are described. Both deep learning features and hand-crafted features are used. Deep learning features are discussed in Chapter 4 Section 4.6, and hand-crafted features refer to the features determined by information involved in images. The K-means algorithm is also applied to cluster the generated feature vectors in order to ensure their discriminative properties.

4.1 Motivation

In classification problems, especially for supervised learning problems, different feature extraction approaches lead to different outcomes. Feature extraction plays an essential role in machine learning, including image classification, image segmentation and object detection. Feature extraction transforms the input data (refers to histology images in this work) into a set of features, dimensionality reduction projects the input data into a lower dimensional feature space. The image representations in the feature space are often called feature vectors and are used to train classifiers.

Feature vectors have applied to image classification, segmentation and object detection tasks. Haralick and Shanmugan (1973) surveyed two types of texture extraction technique: structural and statistical. They concluded that structural features are more suitable to represent the overall texture information of the whole image, and that statistical features achieve better performance in representing local patterns. Regarding the gland segmentation problem, Doyle et al. (2008) used spectral clustering and many texture features, including grey-level, Haralick and Gabor filter features, to classify the different degrees of breast cancer. This method achieved 95.8% accuracy in classifying the non-cancer and cancer images and 93.3% accuracy in separating the different grades of cancer. One possible development for this research is to use deep

learning techniques to classify these images.

In this research, both structural and statistical feature extraction approaches to the gland data are employed, the former for its better representation of global structural properties, and the latter to describe the local structural patterns.

The grey-level co-occurrence matrix (GLCM) is used to differentiate between the gland and the background. There are two reasons for choosing this feature: first, GLCM is one of the structural features that represent the overall gland's morphological structure; and secondly, GLCM was employed to solve and achieve good performance in the mitosis detection problem (Irshad et al., 2013), and the histology images are similar to the gland image used for mitosis detection.

The histogram is one of the intensity-based features representing the colour distribution of an RGB or a greyscale image. The intensity values of the pixels in the image build the intensity-based features. The reason for choosing the histogram feature is the possibility of using colour, and it has already been used in segmenting gland parts in histology images, achieving a performance F1 score of 0.54 (Ap et al., 2017).

Local binary pattern (LBP) is another type of intensity-based feature, used to describe the relationship between the central and the surrounding pixels. LBP is robust to slight changes of brightness in the images. Because there are both shape and colour pattern changes in the gland images, LBP features describe differences between the background and the gland.

Histogram of oriented gradient (HOG) is one of the gradient-based features, sensitive to the edges of the images. Gradient-based features are those generated by the image gradient. Because of contour variation of the gland in the histology images, HOG features are also employed here, and their gradient-based ability is useful in comparing the performance of the intensity-based features.

Deep features (such as GoogleNet and LeNet5 features in this work, this term is explained on page 19) learnt using deep learning techniques. There are two main reasons for choosing this technique. First, deep learning techniques were ranked top in recent competitions. Secondly, studies which focused on using hybrid methods combining deep learning and hand-crafted features have achieved good performance in

similar problems (Manivannan et al., 2018).

The above discussion explains the reasons for choosing these features; each one is discussed in more detail below.

4.2 Grey-level co-occurrence matrix

The grey level co-occurrence matrix (GLCM) describes the co-occurrence of pixel intensities in a given offset; the mathematical expression of the co-occurrence matrix $\boldsymbol{C}(\mathrm{i},\mathrm{j})$ is defined by (Eleyan and Demirel, 2011):

$$\boldsymbol{C}_{(\Delta \boldsymbol{p},\Delta \boldsymbol{q})}(\mathrm{i},\mathrm{j}) = \sum_{\boldsymbol{p}=1}^{n}\sum_{\boldsymbol{q}=1}^{m}\begin{cases}1, if\ \mathbf{I}(\mathbf{p},\mathbf{q}) = \mathbf{i}\ \mathbf{and}\ \mathbf{I}(\mathbf{p}+\Delta\boldsymbol{p},\mathbf{q}+\Delta\boldsymbol{q}) = \mathbf{j}\\ 0, otherwise\end{cases} \quad (4.1)$$

where $\boldsymbol{i}$ and $\mathbf{j}$ are pixel intensity, **p** and **q** indicate the geometrical position in the images $\mathbf{I}$. The offsets $(\Delta\boldsymbol{p},\Delta\boldsymbol{q})$ indicate the spatial relation for the co-occurrence matrix calculated. $\mathbf{I}(\mathbf{p},\mathbf{q})$ is the image intensity at position $(\mathbf{p},\mathbf{q})$ in the image.

There are many relevant features generated from this matrix, such as contrast, entropy, energy, correlation and homogeneity. The technique has been applied in mitosis detection, achieving excellent performance (Veta et al., 2015). The reason for using it here is that the background in the gland data is similar to the mitosis detection images. In this research, four experiments tested four different directions $\{0^o, 45^o, 90^o, 135^o\}$. For each direction, eight different offset matrices, [1 1] [2 2] [3 3] [4 4] [5 5] [6 6] [7 7] [8 8], were used to generate eight different co-occurrence matrices, and the final matrix $\boldsymbol{C}(i,j)$ was obtained as their mean. From each matrix $\boldsymbol{C}(i,j)$, five sets of texture information were calculated as defined in Table 4.1. $\boldsymbol{P_{ij}}$ is the value at (i,j) of the normalised symmetrical GLCM; $\boldsymbol{N}$ is the number of grey levels in the images; μ indicates the means of sum intensity in GLCM; and σ^2 indicates the variance of GLCM.

Table 4.1 Texture information generated from the co-occurrence matrix (Haralick and Shanmugan, 1973)

Texture feature name	Mathematical expression
$\boldsymbol{Entropy}$	$\boldsymbol{Entropy} = \sum_{i,j=0}^{N-1} -p_{ij}\log_2(p_{ij})$
$\boldsymbol{Energy}$	$\boldsymbol{Energy} = \sum_{i,j=0}^{N-1} (p_{ij})^2$
$\boldsymbol{Contrast}$	$\boldsymbol{Contrast} = \sum_{i,j=0}^{N-1} p_{ij} \cdot (i-j)^2$
$\boldsymbol{Correlation}$	$\boldsymbol{Correlation} = \sum_{i,j=0}^{N-1} p_{ij} \cdot \frac{(i-\mu)(j-\mu)}{\sigma^2}$
$\boldsymbol{Homogeneity}$	$\boldsymbol{Homogeneity} = \sum_{i,j=0}^{N-1} \frac{p_{ij}}{1+(i-j)^2}$

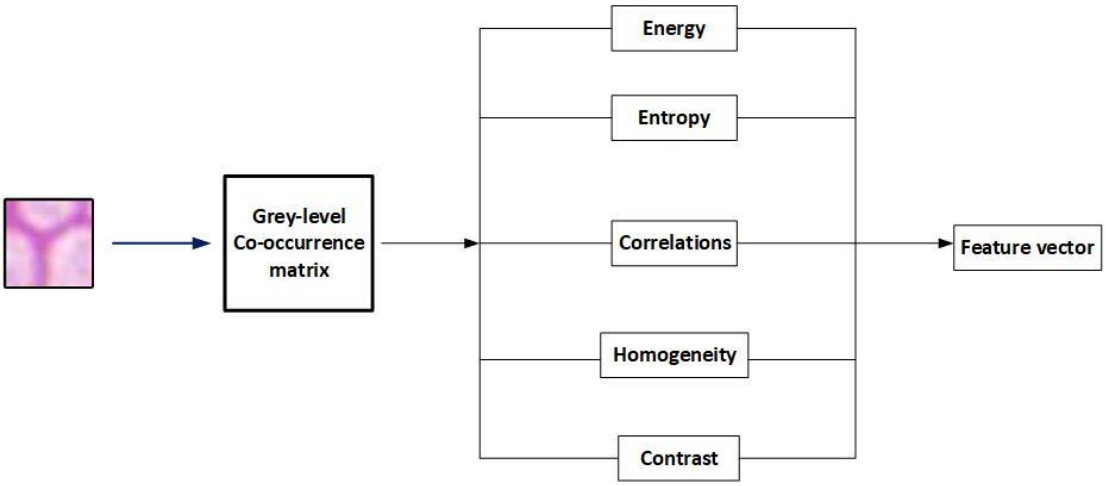


Figure 4.1 The process of generating the GLCM features (Eleyan and Demirel, 2011)

Figure 4.1 indicates the process of generating the features related to GLCM in the gland data. The left-hand part represents one of the examples in selected patches; the mid-left represents the GLCM generated from the selected patch; the mid-right indicates five sets of texture information generated from GLCM; and the right-hand part indicates the GLCM features describing the texture information in the selected patch.

4.3 Histogram

The image histogram feature represents the colour distribution of an image. Because of its computational efficiency it is frequently applied in various research fields,

including image classification, image segmentation, object detection, and object tracking.

Sergyan (2008) introduced a method which employed the histogram intersection as the kernel function applied in the SVM algorithm, and such an SVM model can achieve good results in image classification based on the colour information. Histogram features can be applied not only in image classification but also in image segmentation. Christ et al. (2017) demonstrated a deep learning technique, cascaded fully convolutional networks, to extract and predict the malignant part in testing images. Both histogram and deep learning techniques achieved good performance, although the deep learning approach slightly outperformed the histogram.

As the histogram feature can describe the characteristics of an object, it performs well in object detection. Schneiderman and Kanade (2000) employed the histogram feature to describe automatically local image features of human faces, with above 90% detection accuracy. Their input images were all greyscale, while a colour image could be treated as three individual greyscale images.

Despite the advantages outlined above, the original histogram feature could not distinguish between two images with the same intensity histogram but depicting objects with different geometry. For example, in the two images shown in Figure 4.2, the size of the area outlined in black is equal in (a) and (b). However, the original histogram feature was unable to detect the differences between the two images.

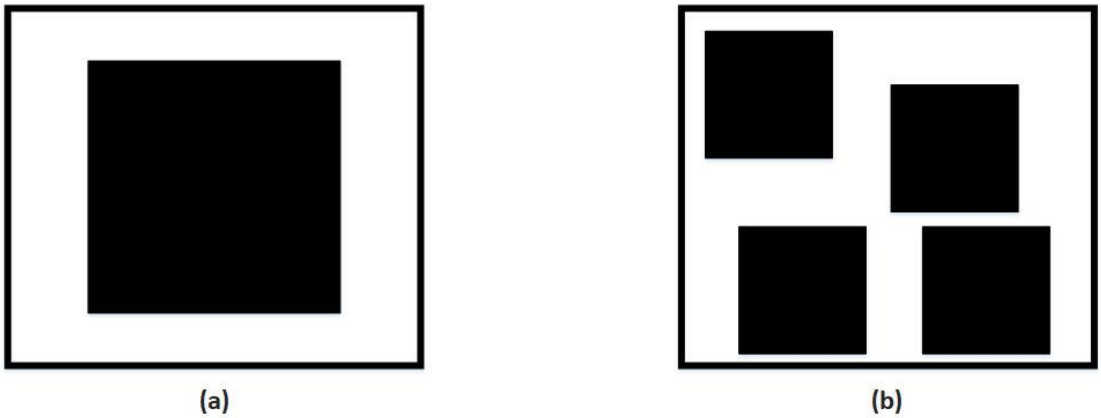


Figure 4.2 Different images with the same intensity distributions (Xiaoling, 2009)

To solve this problem, Xiaoling (2009) introduced an extended version of the histogram feature, where a set of different sized squares is used to extract the histograms for different parts of the region, therefore preserving some of the spatial

information. Figure 4.3 illustrates the results of using this extended histogram feature.

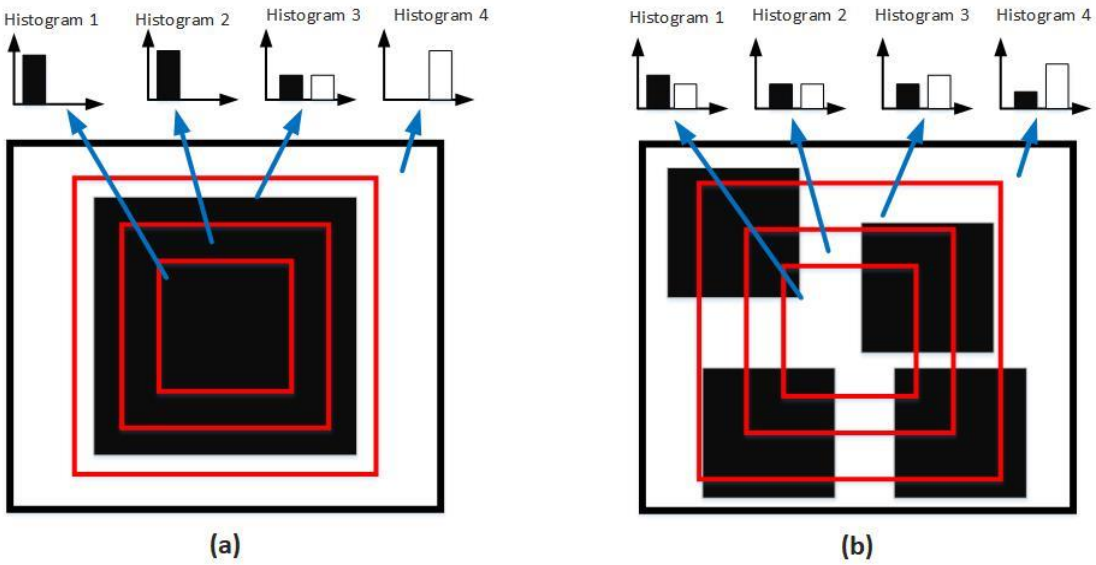


Figure 4.3 Examples of the extended histogram feature proposed by Xiaoling (2009)

This feature is named the ring histogram. Xiaoling employed “rings” with sides of three different lengths to extract the histogram features in the squares’ regions. Histograms 1 to 4 represent the corresponding regions. The final ring histogram combines these histogram features into a single feature vector.

In this work, the ring histogram feature is used to describe local information of the gland or background in the training images.

Although the ring histogram demonstrated by Xiaoling (2009) could solve the problem illustrated in Figure 4.2, it does not have rotation-invariant properties. In this research, a similar approach is employed to extract the histogram features of the training gland images, employing a set of circular shapes with different diameters to extract the histogram features from the gland images as shown in Figure 4.4.

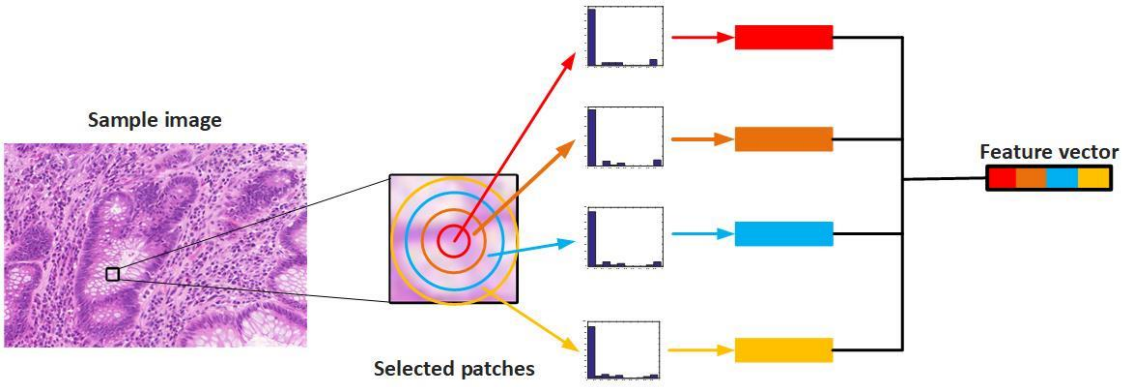


Figure 4.4 Extracting the ring histogram feature from one of the sample images

The central square indicates the patch selected from the sample image to its left. To its right, the different colours indicate histograms with different diameters. Finally, the right-hand part indicates the ring histogram feature vector representing the local patterns in the sample histology image.

In the segmentation without pre-classification, the experiment determined the best size of input patches; the number of rings in each patch is set in order to find the best parameters to generate ring histogram. An experiment to find the best parameter for random forest to provide the best segmentation results has also been set in this part. After identifying the best parameters for generating the ring histogram and for achieving the best results using random forest, these parameters are applied to segmentation with pre-classification. The details of these experiments are given in Chapter 6.

4.4 Local Binary Pattern

The LBP feature (Ojala et al., 1994; Ojala et al., 2002) is a type of texture feature, robust to small changes in the colour of the images. Based on visual inspection shown in Figure 2.3, the colour of gland objects in either benign or malignant tissue can look significantly different. LBP is robust to small changes in colour of the images, and the variation in colour of gland objects could be solve if the LBP has been used as the local image descriptor. LBP became famous because of its simple computation and excellent performance in human recognition (Wang et al., 2009). The original LBP feature did not possess the rotation-invariant property (this term is explained on page 20) , so that if the images were rotated the LBP features would not be the same. Subsequently, many extended versions of LBP features were introduced. Ojala et al. (2002) introduced an approach which employs the circular shift technique to shift the binary code until it matches the pre-selected rotation-invariant patterns. Zhao et al. (2012) combined the LBP feature and Fourier histogram and achieved good performance. Mehta and Egiazarian (2013) provided comparative results based on the different versions of LBP: the original LBP feature, the rotation-invariant LBP feature, the uniform LBP feature and the rotation-invariant uniform LBP feature. The last LBP feature outperformed the other extended versions.

Figure 4.5 is the example of generating the original LBP features described by Ojala et al. (1994, 2002).

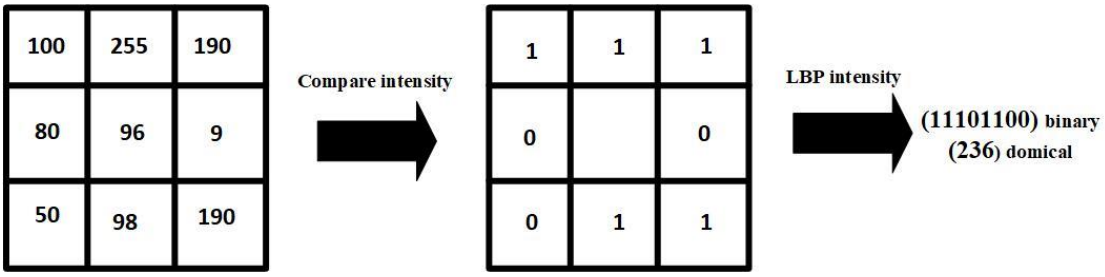


Figure 4.5 Creating the LBP features (Lahdenoja et al., 2013)

The left-hand matrix indicates the intensity of a 3-by-3 patch, and the right-hand one the values after comparing the intensity in this patch. The right-hand side indicates the LBP value used to build the LBP feature vector.

LBP is an intensity-based feature describing the relationship between the central and the surrounding pixels. Ojala et al.'s original version (1994) of LBP was limited by only being able to extract a fixed sized patch, 3-by-3, of the regions in the images. The new version (Ojala et al., 2002) employed a circle to capture the patterns in the images, and the mathematical expression for this type of LBP feature is defined by (Ojala et al., 2002):

$$LBP_{R,P} = \sum_{p=1}^{P} S(\boldsymbol{g}_p - \boldsymbol{g}_c) \cdot 2^{\boldsymbol{p}} \tag{4.2}$$

$$S(\boldsymbol{g}_p - \boldsymbol{g}_c) = \begin{cases} 0, & (\boldsymbol{g}_p - \boldsymbol{g}_c) < 0 \\ 1, & (\boldsymbol{g}_p - \boldsymbol{g}_c) \geq 0 \end{cases} \tag{4.3}$$

where $\boldsymbol{g}_c$ indicates the central pixel and $\boldsymbol{g}_p$ represents the surrounding pixels. $\boldsymbol{p}$ is the index of the neighbourhood around the central one, and $\boldsymbol{R}$ represents the radius of the neighbourhood; $\boldsymbol{P}$ is the number of neighbourhood pixels around the centre. When using a circle to extract the local patterns from the images, the coordinates of the surroundings are determined by the length of the radius of the captured circle and the number of neighbour pixels around the centre. If the coordinate of the neighbour pixels is not an integer, the method will employ bilinear interpolation to estimate the corresponding pixel value, using these values to generate the LBP vector.

With $\boldsymbol{R}$ radius and $\boldsymbol{P}$ the number of neighbour pixels, the number of discriminative patterns of the uniform LBP feature is 2^p. If P is increased, the numbers

of LBP patterns will increase significantly. To improve the statistical efficiency of LBP, the uniform LBP feature was introduced to extract the most fundamental structure. The LBP uniform pattern is defined as the pattern with at most two transitions between 0 and 1. For $\boldsymbol{P}$ surroundings of the uniform LBP feature, the number of patterns is $\boldsymbol{P}(\boldsymbol{P}-1)+3$.

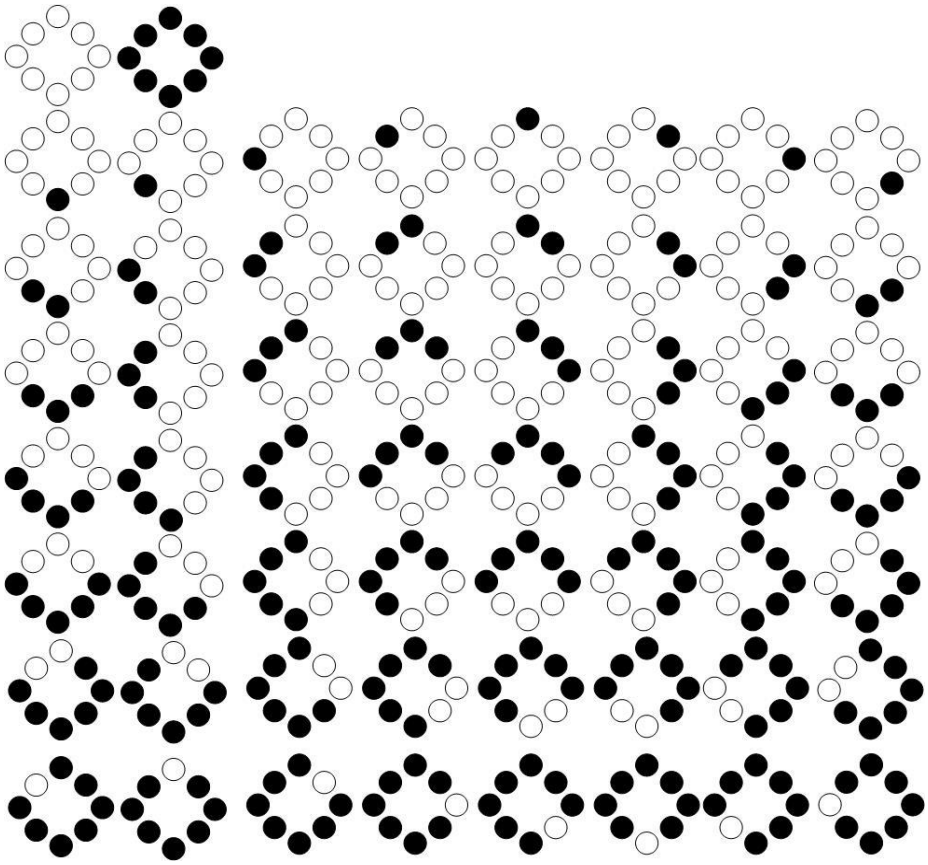


Figure 4.6 Example of eight neighbours uniform LBP features (Pietikäinen et al., 2011)

Figure 4.6 demonstrates the uniform LBP feature with eight neighbours. The white circles represent the pixel value of the neighbour bigger than the central one, and the black circles indicate the pixel intensity of the surrounding pixels smaller than the central one. The total number of uniform LBP feature with 8 neighbour pixels is 59 (8×7+3 =59, the formula ($\boldsymbol{P}(\boldsymbol{P}-1)+3$) has shown on page 72), all discriminative patterns of uniform LBP feature are generated come from original LBP pattern. 58 out of 59 patterns are uniform patterns (all these uniform patterns are shown in Figure 4.6) and all non-uniform patterns from original LBP patterns are constructing as one whole pattern for uniform LBP feature.

However, neither the uniform LBP feature nor the LBP feature has rotation-invariant

properties. Zhao et al. (2012) introduced an extended LBP feature containing the rotation-invariant property. Using the uniform patterns as an example, the process of generating the rotation-invariant uniform LBP pattern is discussed as follows. The number of discriminative patterns of rotation-invariant uniform LBP feature with 8 surrounding pixels is 10. The number of discriminative patterns of rotation-invariant uniform LBP feature is different if the number of the surrounding pixels is changed. In the first row two uniform patterns will build the first two patterns in uniform LBP features. From the second to the bottom rows shown in Figure 4.6, the uniform LBP features in each row will build as one whole pattern in rotation-invariant uniform LBP features. The rest of the non-uniform LBP patterns will build as a family pattern in rotation-invariant uniform LBP features. So, the total number of patterns of rotation-invariant uniform LBP features with 8 surrounding pixels is 10.

Another LBP feature also contains the rotation-invariant properties: rotation-invariant LBP feature. Ojala et al. (2000) introduced a version of the rotation-invariant LBP feature, whose mathematical expression is:

$$LBP_8 = min\{ROR(LBP_8, i) \mid i = 0,1,\cdots,7\} \tag{4.4}$$

where $ROR(x, i)$ is the circular bit-wise right shift on 8-bit number x i times. In Ojala et al. (2000) show that the total number of patterns of the rotation-invariant LBP feature is 36, as shown in Figure 4.7. The number of discriminative patterns of rotation invariant LBP is 36 if using 8-bit encoding and 3-by-3 matrix to capture the local pattern, and the total number of patterns (36) could be derived from equation 4.4. The number of discriminative patterns of rotation invariant LBP would be different if number of surrounding pixels has been changed.

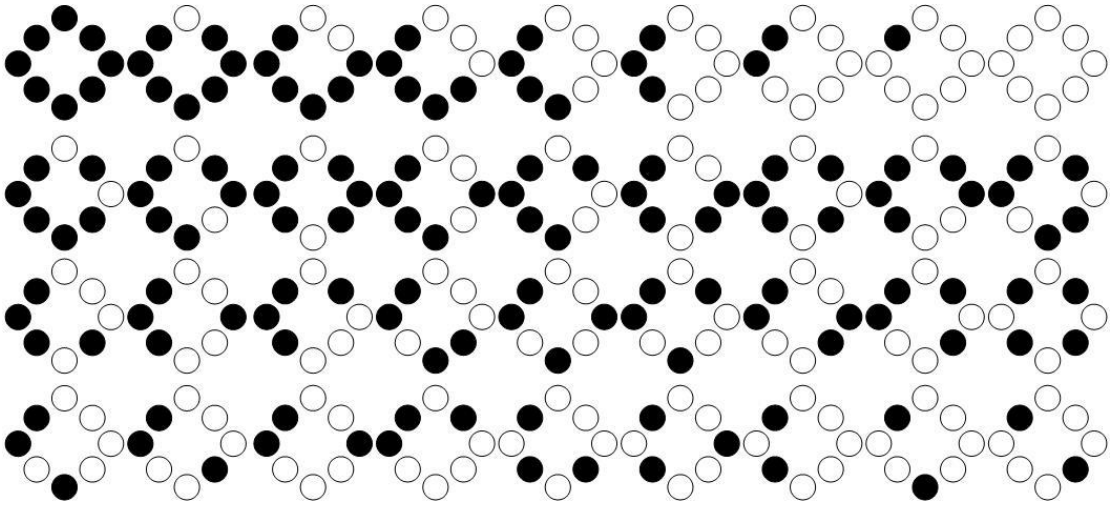


Figure 4.7 The patterns of rotation-invariant LBP feature (Ojala et al., 2000)

Two segmentation methods are employed to classify the gland and non-gland parts in histology images: segmentation with and without pre-classification. The details of these two methods are described in Chapter 5. For these two proposed methods, LBP feature extraction is applied in each selected patch from the histology images. Figure 4.8 shows an example of the process of extracting the original LBP features from a selected patch. The process of calculating the original LBP feature in gland data involves the following steps:

- Step 1: For each selected patch, divide it into four 9-by-9 cells. This input size is performed well in action recognition (Chen et al., 2017), and it has been adapted to gland segmentation in order to find if it is suitable for gland segmentation.
- Step 2: For each cell, compare the intensity of the centre and that of the surroundings. If the intensity of the centre is greater than that of the surrounding pixels, the value of this position is set to 1, otherwise, set to 0.
- Step 3: Calculate the intensity histogram for each cell, and normalise it.
- Step 4: Put all the intensity histograms together as a whole LBP feature for each selected patch. Put all intensity histograms for each selected patch together as a whole representing the local patterns from one image.

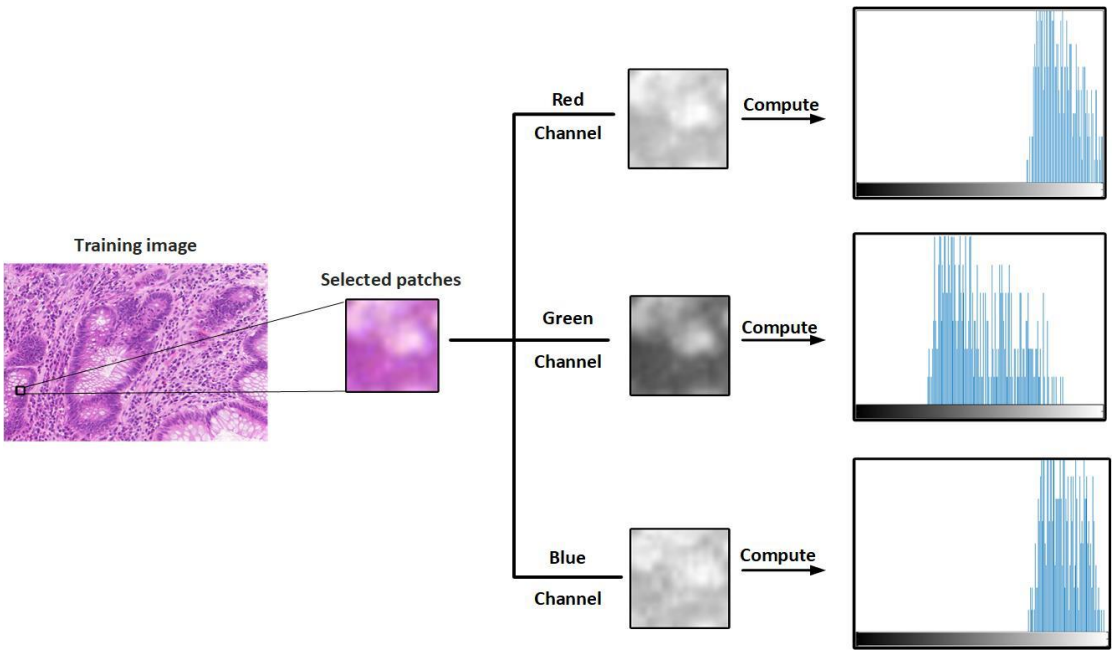


Figure 4.8 Creating original LBP feature for one of the gland images

Figure 4.8 indicates the process of extracting the original LBP feature from the gland image. The left-hand part of the figure indicates one sample training image from the gland segmentation database, with the selected patch to its right. The mid-right part indicates three greyscale images for the different colour channels. The right-hand side illustrates the intensity histogram for the corresponding channel. The number of patterns of the original LBP feature is 256, because the greyscale images contain intensities from 0 to 255. The reason for the number of output intensity is 256 ($2^8 = 256$) is that the original LBP feature is used 8-bit encoding, and the value of intensity will be different if different encoding scheme has been used. The x-axis in the LBP histogram uses a black to white bar to represent these 256 outputs intensity; the y-axis is the frequency of the corresponding pixels in the image. In this case, the total number of patterns from one selected patch is 256×3 = 768.

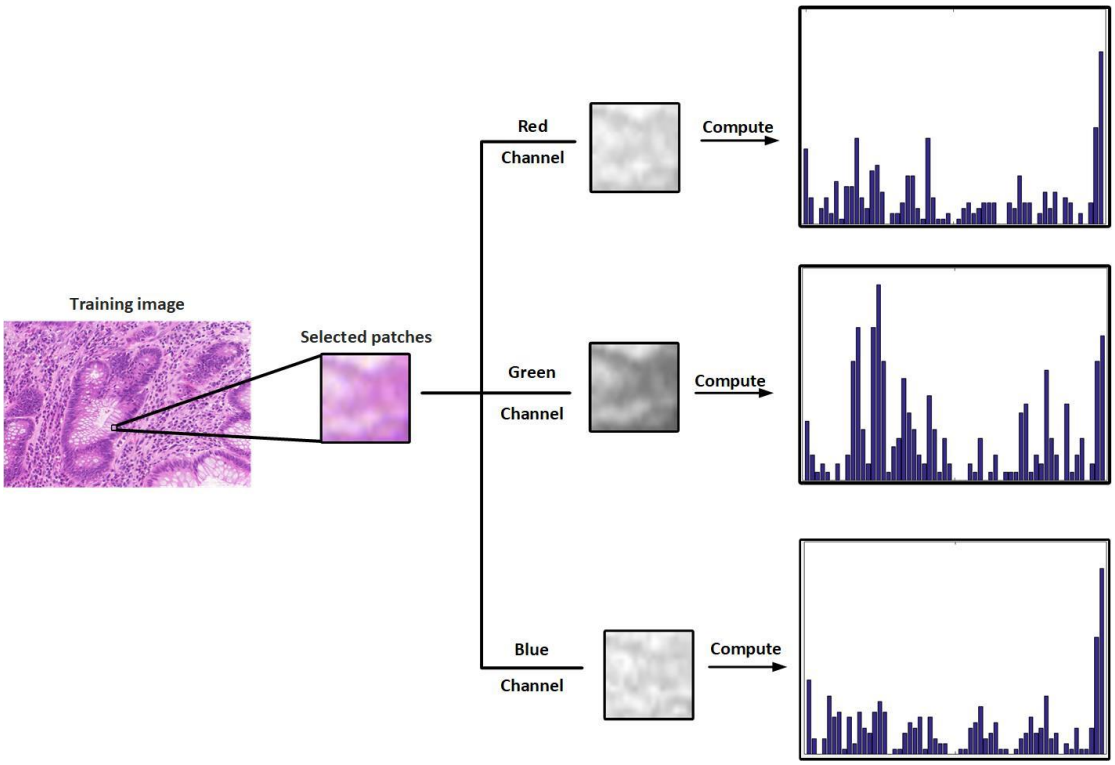


Figure 4.9 Creating uniform LBP feature for one of the gland images

Figure 4.9 shows the creation of the uniform LBP feature from histology images. The middle square presents one of the gland patches from the sample image to its left, and three different greyscale images for the different colour channels; the right side is the intensity histograms representing the uniform LBP feature. The number of patterns for each channel is 59 (the reason for number of patterns being 59 is explained on page 71), so the total number of patterns of each selected patch is 59×3 = 177.

Figure 4.10 shows the process of extracting rotation-invariant LBP features from the details of the rotation-invariant LBP feature given above.

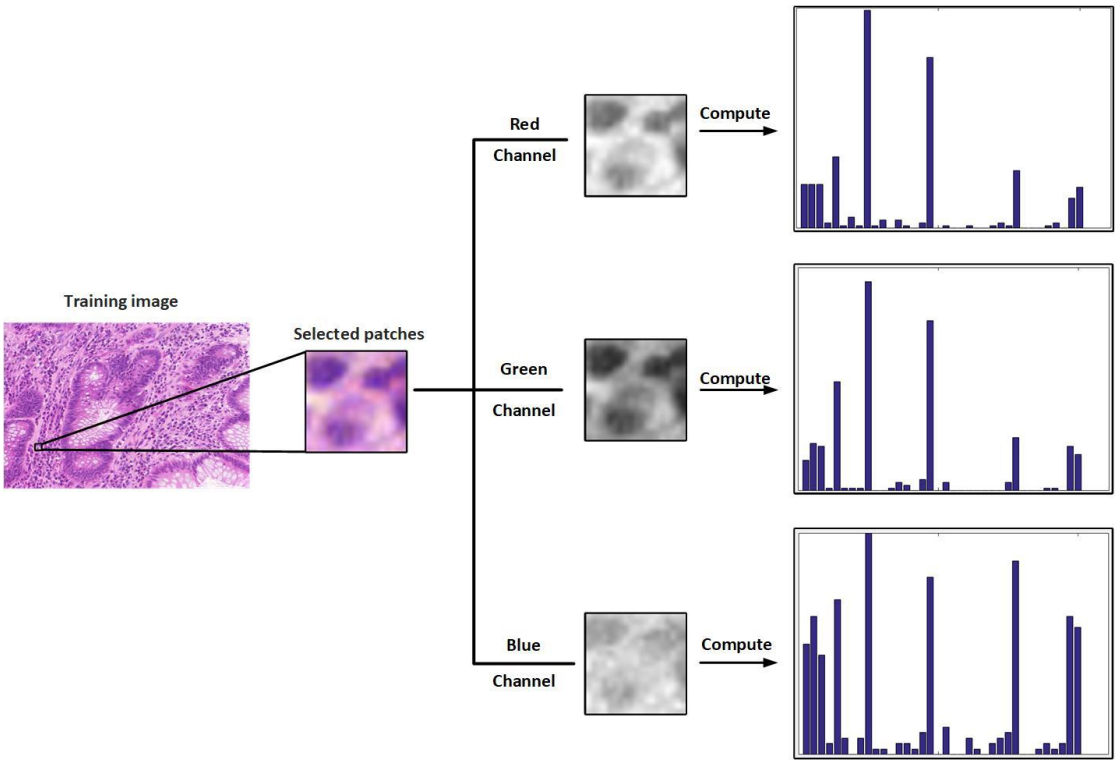


Figure 4.10 The process of extracting rotation-invariant LBP feature from histology image

For each greyscale image of the rotation-invariant LBP feature, the number of patterns is 36 (the reason for number of patterns is 36 is explained on page 73), and the total number of patterns of the selected patch is 36×3 = 108. Finally, extracting the rotation-invariant uniform LBP feature is shown in Figure 4.10.

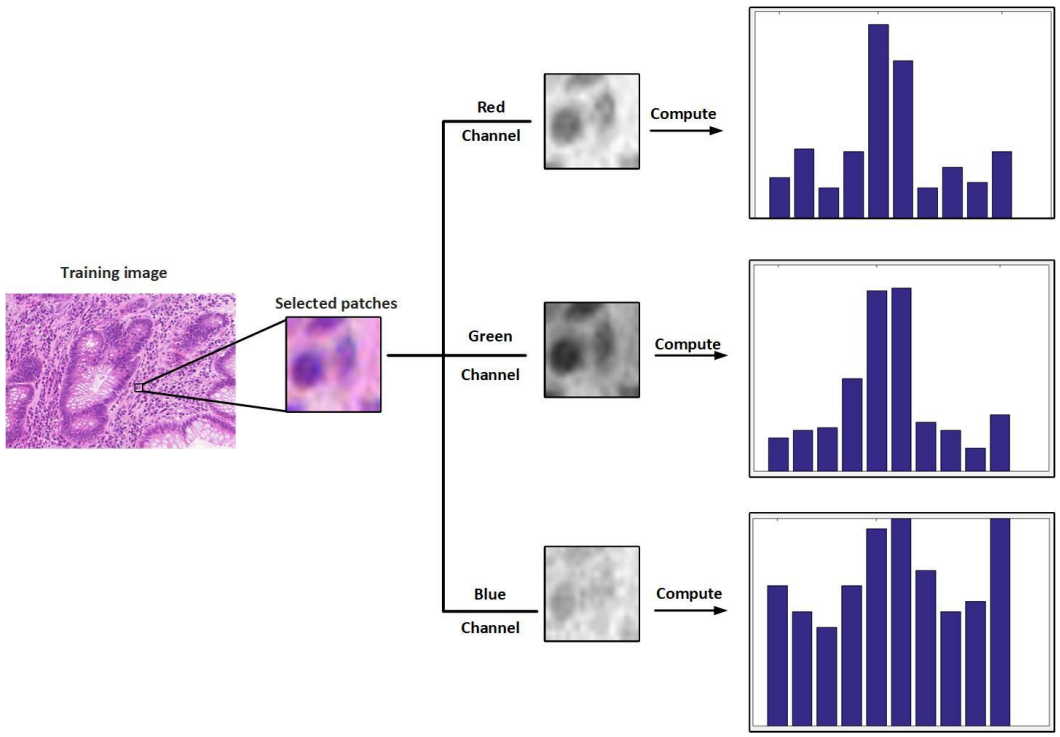


Figure 4.11 The process of extracting rotation-invariant uniform LBP feature from the gland image

Figure 4.11 illustrates the process of extracting the rotation invariant LBP feature from the same patches in the same image. The left-hand part is the same training image, with another background patch to its right, then three different greyscale images from the different colour channels in the selected patch. The right-hand side indicates the rotation-invariant uniform LBP feature. The number of patterns of each greyscale image is 10, as explained on page 73, and the total number of patterns of each selected patch is $10 \times 3 = 30$.

All of these four LBP features are used in the segmentation without pre-classification. The evaluation results of these four LBP features are described in detail in Chapter 6 Section 6.3.3.

4.5 Histogram of oriented gradients

HOG (Histogram of Oriented Gradient) is one type of gradient-based features, widely used in human face recognition and object tracking problems. The gradient in the images builds gradient-based features. Dalal and Triggs (2005) introduced two types of HOG, C-HOG (circular HOG) and R-HOG (rectangular HOG), to extract the local patterns achieving good performance for face recognition. The limitation of the images is that they are sampled in a short video, and most of the faces located in the centre of images, which is not always the case in the image. Neither of these two types of HOG feature has the rotation-invariant property, which was later developed by Skibbe and Reisert (2012). They simply involved mapping the original images into a Fourier domain to extract the local information. The Fourier HOG feature is represented in a similar way to the ordinary HOG feature.

For significant variations in the shape and size of the glands in the histopathological gland database, the HOG feature is employed to describe the difference between the gland and the background. In segmentation with and without pre-classification, original HOG and circular Fourier HOG features are used to represent the local patterns. The reason for the original HOG feature not working will be explained in Chapter 6. In this work, in the summary of segmentation with and without pre-classification, the results of circular HOG but not original HOG will be included. Figure 4.12 describes the process

of generation of original HOG features.

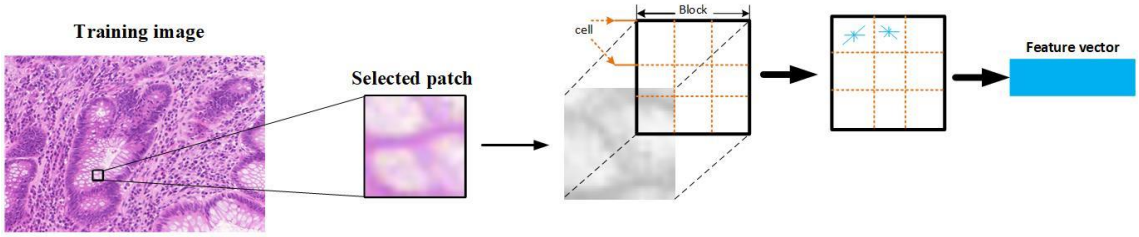


Figure 4.12 The process of generating the original HOG features

The middle part of the figure illustrates the generating process. Each patch is divided into four blocks and each block into nine cells. For each cell, eight bin histograms are extracted. The RGB patches are treated sequentially for each colour component (shown in the mid-right part in Figure 4.12). So, the final number of the original HOG feature is 8×4×9×3 = 864, which is the size of the feature vector shown in the right-hand part of the figure. This version of the HOG feature was only used in the segmentation without pre-classification in this work.

The magnitude of the gradient and the gradient orientation as determined by the original HOG feature are defined by (Gonzalez and Woods, 2002):

Gradient magnitude: $$\mathrm{G(x,y)} = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{4.5}$$

Gradient orientation: $$\alpha = \arctan\frac{G_y(x,y)}{G_x(x,y)} \tag{4.6}$$

where $G_x(x,y)$ and $G_y(x,y)$ are the horizontal and vertical gradients of the pixel (x,y) in the image. The equations (5.7) and (5.8) indicate the expression of these two concepts are defined by (Gonzalez and Woods, 2002):

$$G_x(x,y) = \mathrm{H}(x+1,y) - \mathrm{H}(x-1,y) \tag{4.7}$$

$$G_y(x,y) = \mathrm{H}(x,y+1) - \mathrm{H}(x,y-1) \tag{4.8}$$

where (x,y) indicates the position index of one pixel in the image. $\mathrm{H}(x,y)$ is the pixel intensity of the pixel (x,y).

In the segmentation with pre-classification approach, the circular Fourier HOG feature is employed to describe the different categories in these models. This feature

incorporates rotation-invariant properties by converting the images into the Fourier domain and extracting the HOG feature from it. Fourier HOG also uses the gamma correction in order to handle the non-linear illumination and contrast changes. The expression is defined by (Skibbe and Reisert, 2012):

$$\mathrm{g}_{\gamma} := \|\mathrm{g}\|^{\gamma}\hat{\mathrm{g}} \quad \gamma \in (0,1] \tag{4.9}$$

where g is the gradient field of the image, $\hat{\mathrm{g}} := \frac{\mathrm{g}}{\|\mathrm{g}\|}$, and $\hat{\mathrm{g}}$ is the gradient orientation field. $\|\mathrm{g}\|$ is the gradient magnitude.

The circular Fourier HOG feature applies the 2D Gaussian function to capture the local patterns of the images. The mathematical expression is defined by (Skibbe and Reisert, 2012):

$$\mathrm{CHOG}\{\mathrm{f}\}_{\omega}(x,n) = \int \|\mathrm{g}(\mathrm{r})\| \delta_n(\hat{\mathrm{g}}\,(r))\omega(x-r)dr \tag{4.10}$$

where $\mathrm{CHOG}\{\mathrm{f}\}$ represents the dense field of the Fourier HOG over the whole image. g is the gradient field of the image f. n indicates the current histogram entry. δ_n indicates the Dirac delta function on the circle that selected the gradient out of g with orientation n.

Figure 4.13 illustrates the process of extracting the local circular Fourier HOG feature from the gland images. All the parameters used in the experiments are estimated based on multiple experiments.

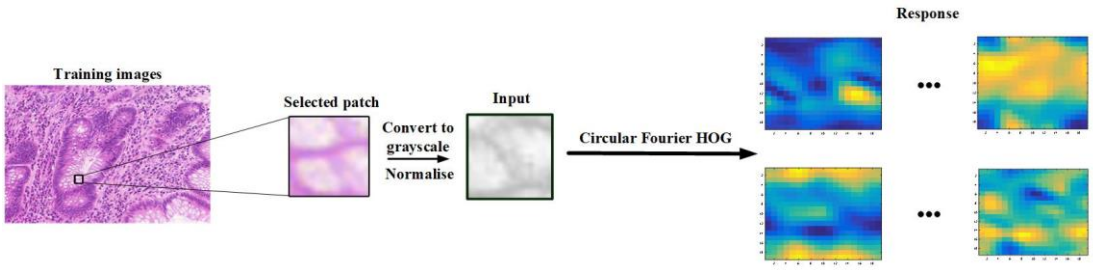


Figure 4.13 The process of extracting the local circular Fourier HOG from histology image

The output of patch is the input to generate the response of circular Fourier HOG showing in the right-hand part. Blue represents low response and yellow high response in the right-hand side in the figure.

4.6 Deep learning features

Deep learning techniques are used not only for the segmentation problem, described in detail in Chapter 3, but also for feature extraction. The features extracted from deep learning techniques are known as deep learning features or deep features. Some models could be used as feature extractors applied to this problem. LeNet-5 architecture, GoogleNet architecture and AlexNet architectures were achieved excellent performance in recognition tasks. He et al. (2016) introduced the ResNet architecture, and achieved better performance on ImageNet classification. Recently, DenseNet (Huang et al., 2017) was introduced and performed better than ResNet in the same object recognition task. There are also other deep learning techniques that could be used as feature extractors.

In this work, two deep learning architectures, LeNet-5 and GoogleNet, are employed to extract the local patterns from histology images from gland data, discussed below. The reason for choosing LeNet-5 architecture is that this architecture is often used nowadays, and this architecture (LeNet-5) still can achieve better performance. The reason for choosing GoogleNet architecture is that GoogleNet is a more up-to-date deep learning architecture compared with LeNet-5 architecture, and GoogleNet is a powerful tool compared with LeNet-5.

4.6.1 LeNet-5 architecture

This deep learning model is the first well-known architecture in handwriting recognition. Pattern recognition is related to feature extraction, and its aims to classify the objects into a set of given categories or classes. Features or feature vectors are used in order to describe the characteristic of the objects (Theodoridis and Koutroumbas, 2014). Feature extraction methods are applied to the given training data to extract the local representation of that data. The architecture (shown in Figure 4.14) has been used to adapted in gland segmentation to extract the local patterns in histology images.

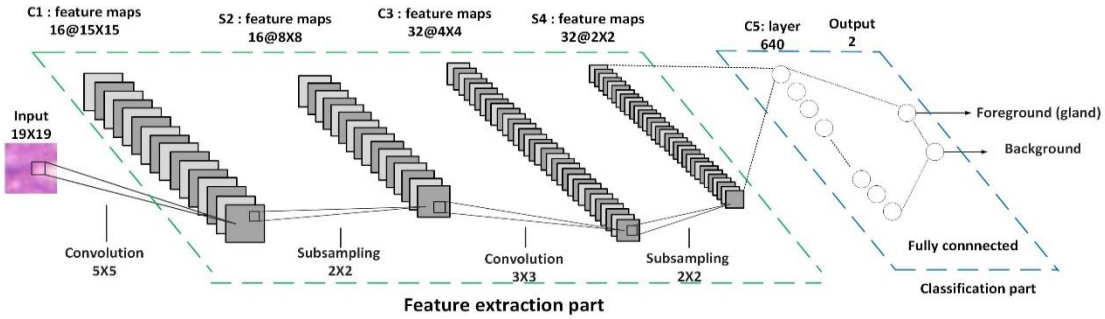


Figure 4.14 LeNet5 architecture adapted in gland segmentation

The LeNet5 architecture was used to extract the local patterns of the gland and the background. Extracting the local patterns and gland segmentation are not the same but related. As mentioned in Chapter 3 (on page 41), this work focused on using supervised machine learning methods to deal with gland segmentation. For supervised learning, it contains training and testing phases. In both training and testing phases, local patterns (or features) of gland and background in histology images are necessary to train the classifier. The experiments which used these architectures (shown on Figures 4.14 and 4.15) as feature extractor are discussed in Chapter 6, Section 6.3.2.

The green bounding box is the feature extraction part of this architecture, and the blue bounding box is the classification section of the network. The deep feature learnt by this model is extracted before the input of the 5th layer. C1 and C3 represent the 1st and 3rd convolution layers in the network respectively, and S2 and S4 indicate the 2nd and 4th sub-sampling layers in the network model. The symbols below the corresponding layers are the number of feature maps and their size in feature extraction part. For example, the symbol C1: feature maps is 16@15×15 means that there are 16 15-by-15 feature maps of the 1st convolution layer in the network. The symbols in the classification part are the names of the layers and the output for the corresponding layer.

The architecture shown in Figure 4.14 is used in both segmentation with and without pre-classification. For the 2-classes pixel-level classification problem, this architecture has been used in the benign and malignant category. For 3-classes pixel-level classification problem in segmentation with pre-classification, a similar LeNet-5 architecture was used to extract the local patterns, as shown in Figure 4.15. The difference between these two architectures (shown in Figures 4.14 and 4.15) is the number of the output. The 2-classes and 3-classes classification are to find the best way

to describe the morphological structure of gland objects in benign and malignant tissue, and the details for generating the ground truth for these two classifications are discussed in Chapter 5, Sections 5.4.2.1 and 5.4.2.2.

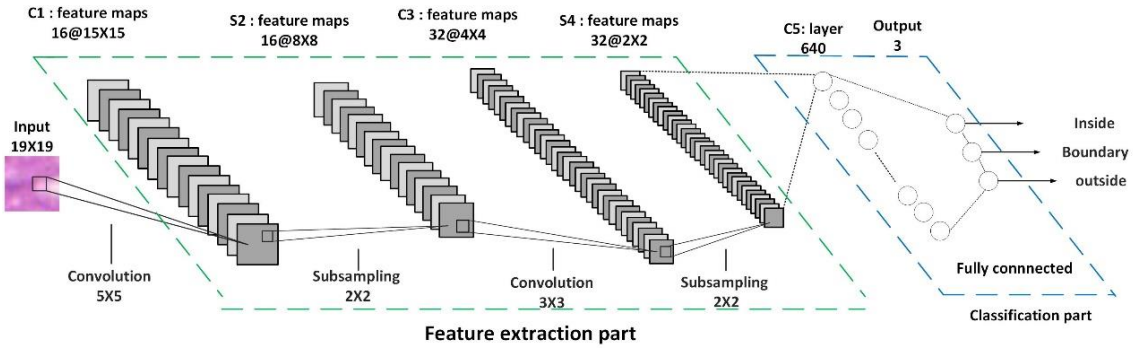


Figure 4.15 The architecture used in segmentation with a pre-classification method for three-classes pixel-level classification method

The parameters in each layer in these architectures, as shown in Figures 4.14 and 4.15, were used without optimising. All these architectures have been adapted in extracting the local pattern in gland segmentation.

4.6.2 GoogleNet architecture

GoogleNet (Szegedy et al., 2015) is a 22-layer deep learning architecture; in this research, it is employed to extract the local patterns in the gland images. Any size of input image could be used to train the GoogleNet architecture, but the size of all these image will change to 224-by-224 in order to fit the requirement of training GoogleNet (Szegedy et al., 2015). The size and number of the input images for GoogleNet architectures will affect the segmentation results. The experiments set to find the best parameters for GoogleNet will be detailed in the segmentation without pre-classification in Chapter 6 Section 6.3.3. For segmentation with pre-classification, the experiments using GoogleNet architecture are shown in Appendix E. The architecture shown in Figure 4.16 is the same as the original architecture, and it has been adapted to gland segmentation used as feature extraction.

For the segmentation with pre-classification method, the architecture used for 2-classes pixel-level classification for the images with one category tissue (benign or malignant) is the same as the architecture shown in Figure 4.16. For 3-classes pixel-level classification in segmentation with pre-classification for one specific category cases, the

features were extracted using the similar GoogleNet architecture but with 3 outputs (inside, outside and boundary of that category cases).

The segmentation results of features from GoogleNet architecture in segmentation with and without pre-classification are discussed in Chapter 6, Sections 6.3.3 and 6.4.2.

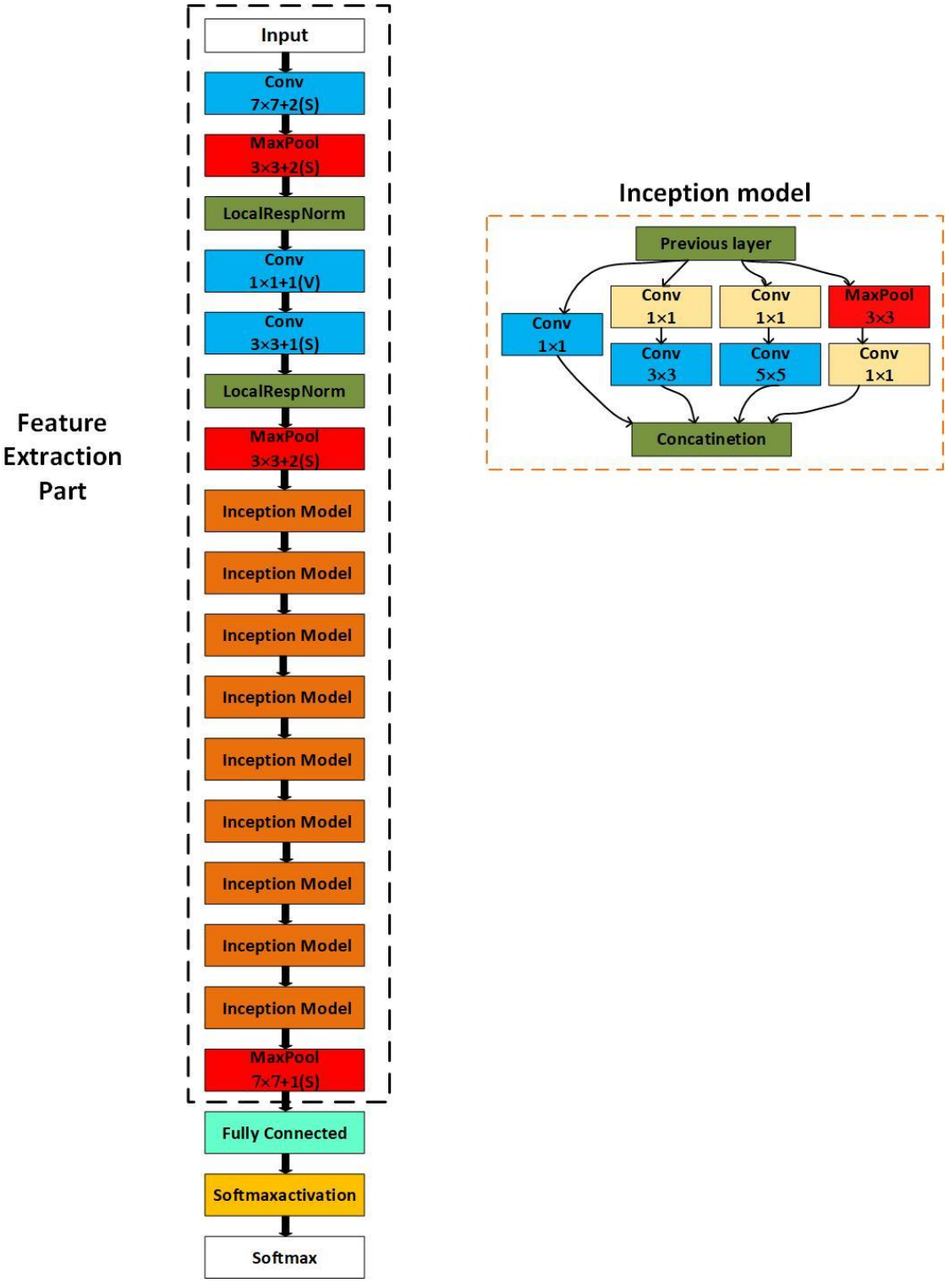


Figure 4.16 GoogleNet architecture to extract the local patterns in gland segmentation

The GoogleNet architecture is described by Szegedy et al. (2015), and the bounding box with a black dashed line is the feature extraction part in GoogleNet. Deep learning features are extracted after the bounding box before the fully connected layer.

4.7 Features discriminative analysis

Before using these feature vectors, it is useful to investigate their discriminative properties. One of the techniques is to use K-means clustering (Lloyd, 1982) in order to check if the features could cluster into distinctive categories as in the training data. The reason for choosing K-means algorithm is that it is often used, and it can be easily interpreted (discussed in Chapter 3, Section 3.5).

The K-means algorithm was used to estimate the discriminative properties of different feature vectors extracted from histology images. The estimation process is detailed in this section, and the reason for choosing this technique. The K-means algorithm uses Euclidean distance when performing the clustering.

The estimation process used K-means algorithm in this research follows:

- Input: The feature $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ extracted from histology images, and the number of distinctive clusters $\boldsymbol{K}$.
- Step 1: Using K-means algorithm to partition the features into $\boldsymbol{K}$ different clusters.
- Step 2: Using the original category of the features $\boldsymbol{X}$ to label the new $\boldsymbol{K}$ clusters; the label of the cluster will be the label with the maximum samples of that category.
- Step 3: Calculate the accuracy between $\boldsymbol{K}$ clusters and the original feature vectors.

The number of clusters is different for different feature vectors. For example, the value of $\boldsymbol{K}$ in the segmentation without a pre-classification method is 2, and the value of $\boldsymbol{K}$ in segmentation with a pre-classification approach is either 2 or 3, depending on the number of the classes. The evaluation measure selected to describe the discriminative properties is accuracy. The label for corresponding cluster is return to the maximum classes contained in that cluster.

All the experiments which evaluating the discriminative properties are detailed in Appendix E. From the results shown in Appendix, the features generated from GLCM have the worst discriminative properties (only achieving 50% accuracy), and the GoogleNet deep learning features have the best discriminative properties with above 90% accuracy. Some other data features from histology images have the results somewhere in between these two.

4.8 Summary

This chapter discussed the feature extraction methods used in gland segmentation. The motivation for using each of these methods was discussed, followed by application of the feature extraction approaches themselves. Both hand-crafted and deep learning features were considered. For the hand-crafted features, ring histograms, LBP and HOG all achieved good performance for local patterns in the images. The features generated from the GLCM performed well for whole texture information in the images. All deep learning techniques could be used as feature extractors, and two deep learning models were employed to extract the local patterns.

Most of the related studies focused on demonstrating only the algorithm to solve the problem, ignoring the discriminative properties of the feature vectors before the training process takes place. However, in this research, after extracting all the features from the histology images, the unsupervised learning technique, K-means, was used to investigate their discriminative properties, a significant difference with other studies. In summary, the estimation results of all feature vectors extracted from histology images show that the discriminative properties and could be used in further testing in segmentation.

The following chapter demonstrates two segmentation approaches, with and without pre-classification method. The pre-processing method to process the colour in the histology images, and the post-processing method processed the probability map will be discussed.

Chapter 5

Segmentation method and pre-/post-processing

Chapter 4 discussed the feature extraction for the histopathological gland segmentation data, and the estimation of the discriminative properties of features extracted from the histology images. In this chapter, two types of segmentation, with and without pre-classification, are employed to classify the gland and non-gland parts in testing images. Due to the variation in colour, histogram correction is used to convert the colour in all the images to the same level. The output of the pixel-level (this term is discussed on page 20) classifier, random forest, is the probability maps, and two post-processing methods are used in order to produce better segmentation results.

5.1 Image classification and segmentation

Image classification applications use machine learning techniques to learn the characteristics of a set of labelled data, and to predict a set of unlabelled data using the trained classifier. Image classification and image segmentation are related but not the same, and these two terms are discussed in 'Thesaurus' (starts from page 19). Image classification and segmentation use some of feature vectors to represent the characteristics of the different regions in the image to train the classifier, and the prediction will be made based on the trained classifier.

In this work, the histopathological segmentation problem is involved in distinguishing between the background and the foreground (gland) in testing gland images, using the two types of the segmentation technique, as described follows.

5.2 Segmentation without pre-classification

The segmentation without pre-classification is the typical method for gland segmentation. This method (Segmentation without pre-classification) is designed based

on the aim and objectives of the work. The segmentation without pre-classification employs the sliding window technique to extract local patterns from the training gland images. The aim of the gland segmentation is to separate the background and the gland parts in histology images. The local patterns of gland and non-gland parts are extracted based on the provided ground truth. Chapter 4 explained the methods to extract the local patterns from selected patches in these images.

Figures 5.1 and 5.2 show the traditional way of solving this problem; Figure 5.1 shows the training phase of the classifier, and Figure 5.2 illustrates the testing phase of the gland segmentation process.

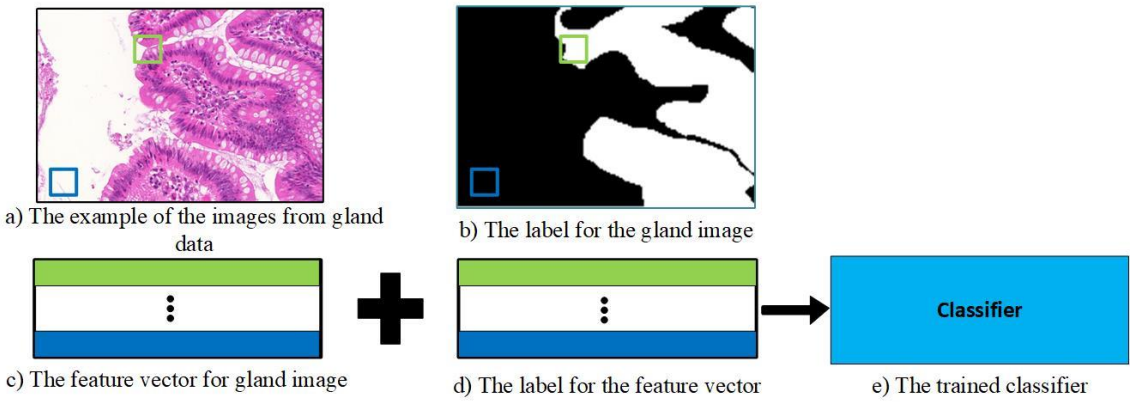


Figure 5.1 The process of the training phase of segmentation without pre-classification method (Manivannan et al., 2017). (a) sample image from the gland dataset. (b) represents the label for the same sample image. (c) represents the feature vector generated from the gland image, and (d) the label vector generated from the label image. (e) represents the trained random forest model

Figure 5.1 indicates the segmentation without pre-classification approach to train the forest models using the sliding window technique. Figure 5.1.a is the sample image from the gland database, and the blue and green squares are extracted from it. Figure 5.1.b represents the label for the same sample, the white part being the gland label, and the black part the background. The blue and the green squares are the background and gland categories for the image. Figure 5.1.c demonstrates the feature vector generated from the selected patches, with the blue and the green parts the features for the gland and the background respectively. Figure 5.1.d is the label for the feature vector, the colours corresponding to the same categories. Figure 5.1.e establishes the trained classifier; in this work, the random forest is used as the primary classifier, using the feature vector and its label to train the models.

The testing process follows, using the trained classifier to predict the testing images. Figure 5.2 shows this process for an image from the gland database.

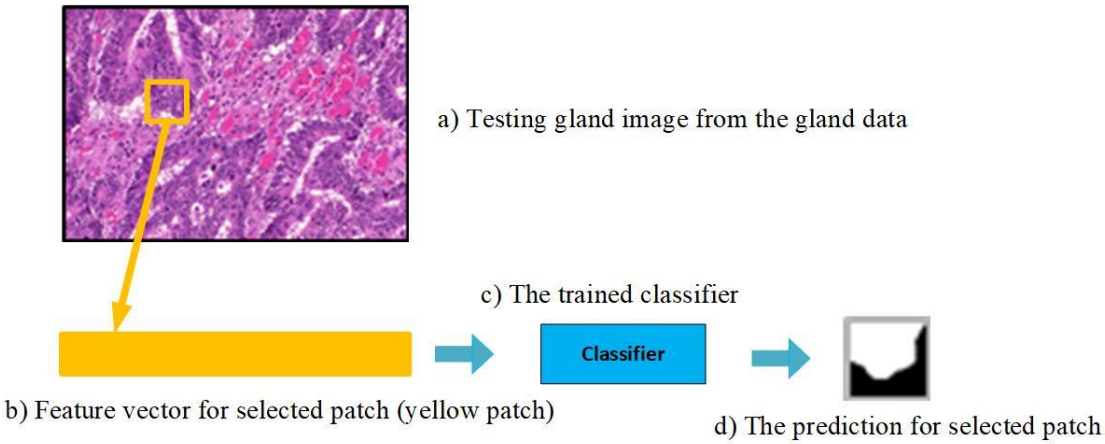


Figure 5.2 Testing process of segmentation without pre-classification (Manivannan et al., 2017). (a) sample testing gland image from the gland database. (b) feature vector for the selected area. (c) the trained classifier (the same classifier as shown in 5.1.e). (d) prediction for the selected patches from the testing image.

Figure 5.2 indicates the testing process for the segmentation without pre-classification. Figure 5.2.a illustrates the sample testing image and Figure 5.2.b the feature vector generated from the yellow square in the histology image. Figure 5.2.c represents the classifier previously trained (see Figure 5.1), and Figure 5.2.d represents the prediction for the feature vector generated by the selected patches from the testing image.

Figures 5.1 and 5.2 show the segmentation without pre-classification approach to identifying the gland and the background in histology images, referred to the traditional or usual way to solve the gland segmentation. These two figures (Figure 5.1 on page 88 and Figure 5.2 on page 89) are referred to the typical method that would be used in this work.

5.3 Limitations of segmentation without pre-classification

Chapter 2 discussed the various components in benign and malignant cases, and gave examples of images from the gland dataset. From visual inspection of images with benign tissue and those with malignant tissue in the gland dataset, the morphological structure in the images with benign tissue is seen to be different from that in images

with malignant tissue. Figure 5.3 shows a sample of training images from the gland dataset, the top row being malignant cases, and the bottom row is benign cases.

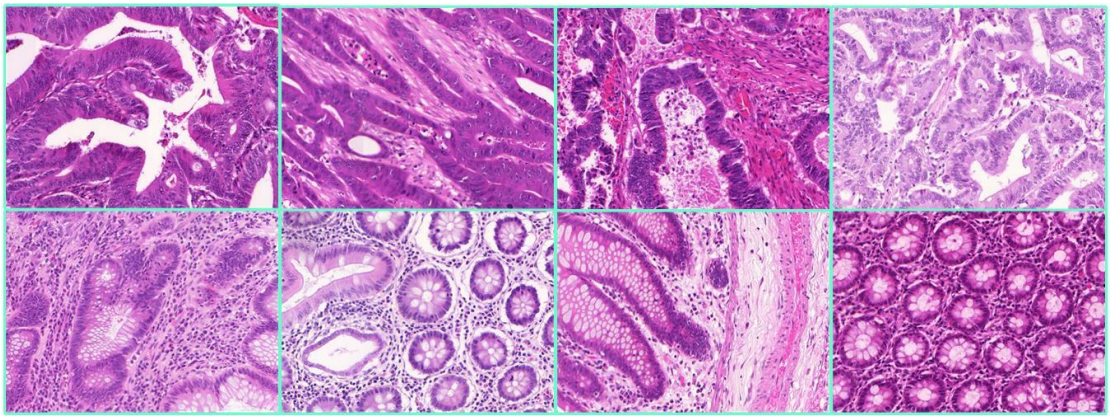


Figure 5.3 Sample images from the gland dataset

From Figure 5.3, it is easy to observe that the patterns in the malignant cases and that in benign cases are different. In the images with benign glands, lumen, cytoplasm and epithelial cell comprise the gland parts in corresponding histology images, and others are the background parts in histology images. In the images with malignant glands, cytoplasm and epithelial cells comprise the gland parts, and others are the background part. The differences in texture between these two types are the glandular tube in images with malignant tissue, which are broader than that in images with benign tissue, and there is no bubble structure in the malignant images that occurs in images with benign tissue. These two examples demonstrate the most significant differences in distinguishing between these images with two different types (benign and malignant) tissue.

By inputting all extracted features from histology images into random forest, the forest model will learn the differences between the gland and the background. It is harder for the classifier to learn the significant differences between the images with benign tissue and those with malignant tissue, and the segmentation results are inadequate based on the performance of three segmentation methods shown in Chapter 6, Section 6.6. The solution for further improving the performance is to separate the images first (image-level classification) and then deal with gland segmentation (pixel-level classification). Segmentation with pre-classification is designed in order to force the forest model to learn the characteristics between the gland and the background in images with two different types of tissue (benign and malignant tissue).

Segmentation with pre-classification approach was therefore employed in this research in order to help the classifier distinguish between the characteristics of the gland and the background in these two types of image. This approach originated with the CVML group described by Sirinukunwattana et al. (2017), and employs three categories of pattern in the gland images to apply a deep learning technique.

Unlike the original CVML, segmentation with pre-classification employs two or three categories of pattern in one of the types of gland image to train and to predict the gland parts in that type of the image. In this case, the patterns in images with benign or malignant tissue are manually separated and train the forest model. The classifier seems to learn the differences between different classes in image with benign or malignant tissue improving the segmentation results. The following section discusses the details of segmentation with pre-classification.

5.4 Segmentation with pre-classification approach

Figure 5.4 illustrates the process of the segmentation with pre-classification method used in this research.

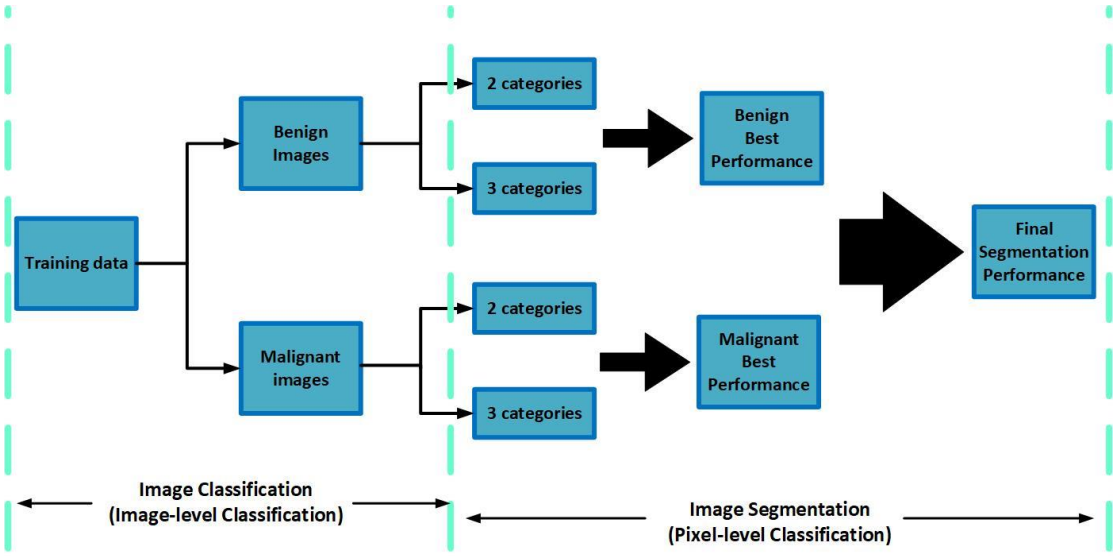


Figure 5.4 The process of segmentation with the pre-classification approach

From Figure 5.4, it is easy to see that the process of segmentation with pre-classification contains two-level classification, image-level and pixel-level. The meaning of these two terms are discussed in 'Thesaurus' (start from page 19). In the image-level classification part, both traditional approaches and deep learning techniques are

employed to distinguish benign or malignant images in gland dataset. In the pixel-level classification step, different feature extraction approaches demonstrated in Chapter 4 are employed to extract the corresponding patterns in the gland images. For the pixel-level segmentation problem, two or three categories of patterns of the gland images are extracted by applying the sliding window technique (see Figure 5.4). The two or three categories are designed in order to find the best way to describe the morphological structure of the gland in either benign or malignant tissue. The methods used in image-level and pixel-level classification are discussed in the following sections.

5.4.1 Image-level classification

The previous section introduced segmentation with pre-classification applied to the gland segmentation problem. In this section, the traditional hybrid method which combines the HMAX model with random forest and deep learning techniques is discussed and applied to the image-level classification problem. This step aims to separate the images in the gland dataset into benign or malignant.

HMAX model

Theriault et al. (2013) introduced the HMAX model and achieved excellent performance in image classification problems. There are both similarities and differences between the HMAX model and the deep learning techniques (also known as a convolutional neural network or CNN). The crucial difference between these two models is that the back-propagation approach is employed in deep learning but not within HMAX. The HMAX model employs a bank of Gabor and convolving filters to achieve the feature extraction. The reason for choosing HMAX and forest model is that this method can be easily interpreted, whereas the LeNet is hard to understand the details of generating these classification results.

In this work, the HMAX model is adopted, although only to extract features from the benign and malignant images, using them to train the forest models to classify histology images in testing data benign or malignant. Figure 5.5 shows the structure of the HMAX model; each layer is then discussed.

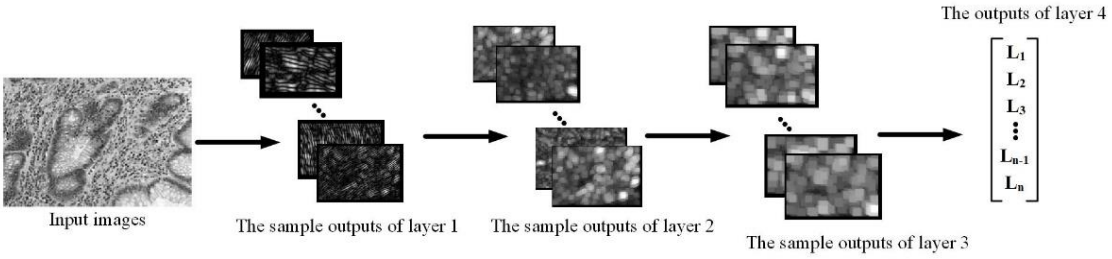


Figure 5.5 Structure of the HMAX model (Theriault et al., 2013)

The structure of the HMAX model is similar to that of the network shown in Figure 4.12. The input is a greyscale image generated by converting one of the original gland images from the gland dataset. The figures next to the input images are sample outputs from layer 1 in the HMAX model, generated by convolving the set of Gabor filters with the input image. The sample outputs from layer 2 are determined by applying max-pooling sampling, which stores the most useful information in a fixed section in the images, by storing the maximum value in a fixed region of the original images.

The output of layer 1 in the model is determined by the results of the convolution operation between the input images and a set of Gabor filters; the size and direction of the Gabor filters are defined by the following formulae in the reference (Theriault et al., 2013):

$$\mathrm{G}_{\sigma,\theta}(x,y) = \exp\left(\frac{x_0^2+\gamma\cdot y_0^2}{2\sigma}\right)\cdot\cos(\frac{2\pi}{\lambda}x_0) \tag{5.1}$$

$$x_0 = x\cos\theta + y\sin\theta \tag{5.2}$$

$$y_0 = y\cos\theta - x\sin\theta \tag{5.3}$$

where θ, λ and σ respectively represent the direction, wavelength and size of the Gabor filter; (x,y) indicates the coordinate index of the pixels in the input images, and γ is the aspect ratio of the Gabor filter. The output from layer 1 is the absolute value of the results of the convolution operations between the input images $I(x,y)$ and the Gabor filters, and the mathematical expression is defined by (Theriault et al., 2013):

$$\mathrm{L1}_{\sigma,\theta} = \left|G_{\sigma,\theta} * I(x,y)\right| \tag{5.4}$$

where $\mathrm{L1}_{\sigma,\theta}$ indicates the output of layer 1; $I(x,y)$ and $G_{\sigma,\theta}$ respectively represent

the images input into layer 1 in the HMAX model and the Gabor filters used in layer 1, as introduced above.

Layer 2 in both the HMAX model and the convolutional neural network is similar, using the selected feature map from layer 1 for dimensionality reduction, and selecting one small neighbour in the feature map of layer 1 and the maximum value of each position (x, y) in $H_{\mathrm{x,y}}$ as the output of this layer. The mathematical expression is given by (Theriault et al., 2013):

$$\mathrm{L2}_{\sigma,\theta}(x, y) = \max_{H_{\mathrm{x,y}} \in \mathrm{L1}_{\sigma,\theta}} H_{\mathrm{x,y}} \tag{5.5}$$

Layer 3 in the HMAX model is composed of one convolving filter. This filter combines the low-level features from the Gabor filter and the regional mid-level features in the image. The corresponding mathematical expression is defined by (Theriault et al., 2013):

$$L3_{\sigma}^{m} = \alpha^{m} * L2_{\sigma} \tag{5.6}$$

where α^{m} and $L3_{\sigma}^{m}$ respectively represent the convolving filter of layer 3 and the output of this layer.

The essence of layers 4 and 2 in the HMAX model is the same. The maximum value of $L3_{\sigma}^{m}$ in layer 3 is selected to form the output of layer 4. That is, the output of layer 4 is a feature vector composed of the maximum values of features selected by a series of convolving filters.

After extraction from the histology images, these features are input into the random forest model to train the classifier. The types of gland image in testing will use the trained forest model for prediction.

AlexNet

Most of the deep learning techniques (i.e. LeNet5, AlexNet and GoogleNet) can be used in the image classification problem, and they have achieved better performance compared with typical machine learning methods (i.e. SVM, decision trees and random forest).

Krizhevsky et al. (2012) introduced the AlexNet architecture and achieved a top-5

classification error rate of 15.3% significantly better than the second best method with 26.2%. Compared with LeNet-5, AlexNet is deeper and more powerful. The original contribution of this architecture was to introduce the ReLU function and dropout. The reasons for introducing ReLU are: (1) AlexNet is not efficient if the activation functions were using a sigmoid function, but with ReLU function the computation time is reduced; (2) for a deeper convolution network, the gradient will decrease close to 0 when the sigmoid function is employed in training. In this case, it is difficult to complete the training of the corresponding deep learning model. The reason for using dropout in AlexNet is that this technique could reduce the number of correlations with other neurons in the architecture, helping the network to learn more robust features.

GoogleNet and ResNet

GoogleNet and ResNet-50 are two up-to-date deep learning methods for image classification. The reason for choosing these methods is that the excellent performance in related tasks. GoogleNet has performed well on food image classification (Singla et al., 2016) and medical image classification (Khan et al., 2019; Bayramoglu et al., 2016). There are several ResNet architectures, such as ResNet-34 and ResNet-50. The ResNet-based network achieved better results in vehicle classification (Jung et al., 2017) and the classification of medical image (Yu et al., 2017). The reason for discussing deep learning methods here is that they have been used for different purposes. Deep learning methods discussed in Chapter 4, Section 4.6 are used for feature extraction. These architectures (GoogleNet, ResNet and AlexNet) have been used in image-level classification.

The above three deep learning models are used as part of the image-level classifier. The details of creating the training data, the validation data and testing data will be provided in Chapter 6, and the training process and parameters used in these models will also be discussed.

Image augmentation

The overfitting could be an issue when training deep learning models. Image augmentation methods are used to avoid overfitting when training deep learning models.

There are many possible image augmentation methods used to increase the number of training images. The local image deformation and colour jitter methods are proposed here to improve training of the deep architectures.

Local image deformation is to change the appearance of the image by mapping the points in the images to new positions without changing the colours. It should only slightly change the morphological structures in the images. It reflects a possible variations in tissue morphology during the preperation of the histology tissue sample. Figure 5.6 shows a sample of the local image deformation method applied in histology images from gland data.

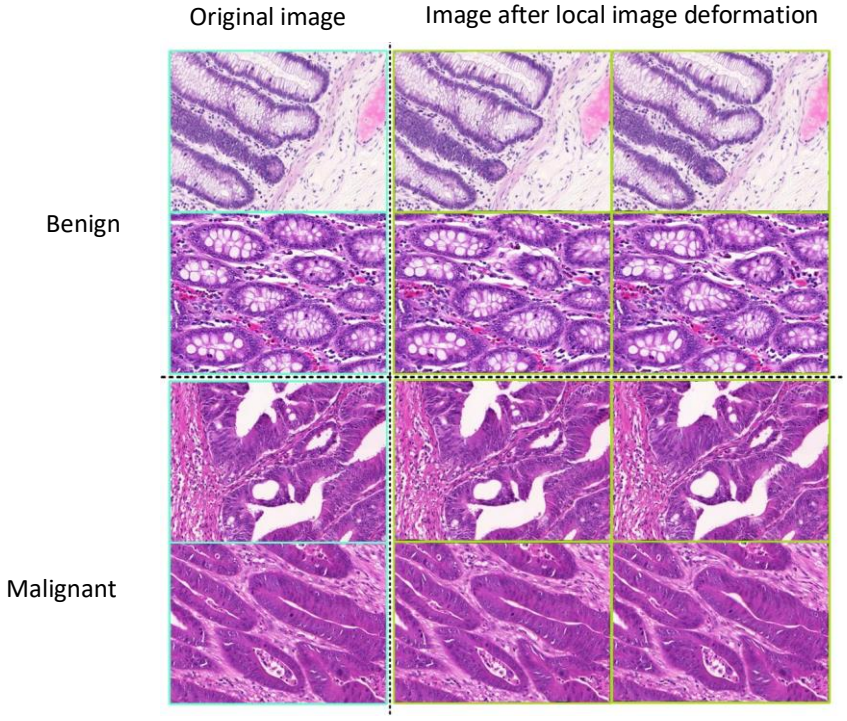


Figure 5.6 Original images and images after local image deformation

The images with a bright blue boundary are examples of the images with benign tissue and images with malignant tissue from the gland dataset (Nasir, 2015). Those outlined in bright green are the results after using local image deformation. From the visual inspection, it is readily seen that the contours of the glands in the original images are different from those after using local image deformation methods. Slight changes in the contours of the gland in benign and malignant cases could help the classifier to learn more information to improve the performance.

Colour jittering is another type of data augmentation widely used in deep learning

to change the value of the pixels in order to change the colour of the images without changing the texture information. The following steps are required:

- Step 1: Convert the original RGB image into HSV colour space.
- Step 2: Adding an random number in each dimension of HSV colour space, using:

$$Image_{i\epsilon(h,s,v)} = Image_{i\epsilon(h,s,v)} + n \tag{5.7}$$

where $Image_{i\epsilon(h,s,v)}$ represents each channel of the image. n is the random number, and the range of this value is between 0 and 1. The range of n is different, and will be different again if different data is selected.

- Step 3: Convert the image, $Image$, from HSV colour space into the original RGB colour space.

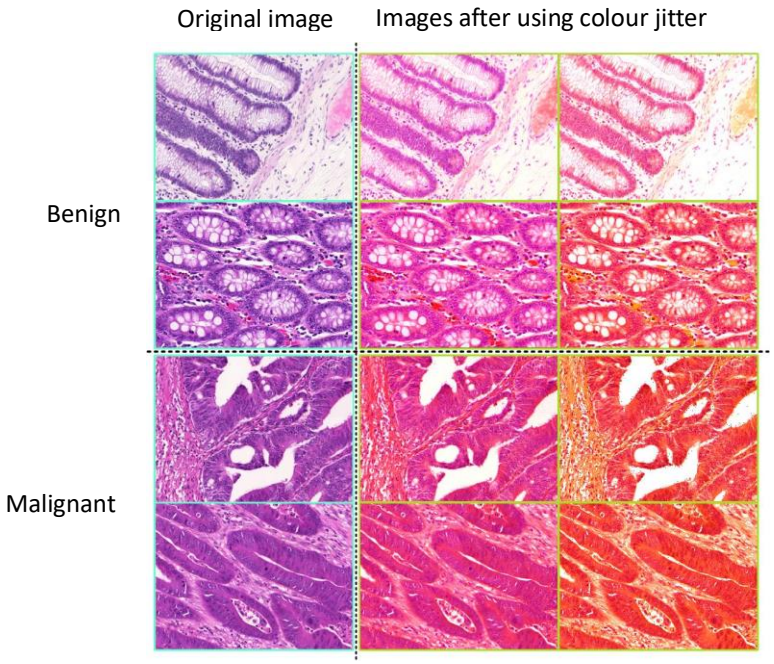


Figure 5.7 Original images and images after colour jitter

Figure 5.7 shows original images from the gland data, and the resulting after using the colour jittering approach. From visual inspection, the colour jittering changed the colour in the original images and without changing structure of the glands in the images.

5.4.2 Pixel-level classification

The pixel-level classification in segmentation with pre-classification aims to segment the background and the gland in the testing images, using the sliding window technique. The morphological structure of the gland generates the different categories

of benign or malignant image in the pixel-level classification part. For example, the segmentation with pre-classification method two benign classes are used in extracting the gland and the background features in benign training images and in predicting these sections in the testing images.

5.4.2.1 Two-categories pixel-level classification problem

In this section, data for the two classes of the pixel-level classification problem, benign or malignant, are considered as one category of gland images to extract the features and train the classifier, and then make the prediction for the corresponding category of gland images. For example, the two benign classes in segmentation with pre-classification enter the features extracted from the images with benign tissue into the forest models, and predict the benign testing images based on the trained forest model. A similar technique is employed for the two malignant classes in the improved segmentation problem.

For these 2-category labels, the original background and foreground (gland) labels are provided in the gland segmentation data. The features extracted from the benign or malignant images employ the feature extraction approaches demonstrated in the Chapter 4. The experiments' details of two target classes classification using are shown in Appendix E. For the benign cases, the results are shown in Appendix E, Section E.1. The results of malignant cases are shown in Appendix E, Section E.3.

5.4.2.2 Three-categories pixel-level classification problem

From Figure 5.3, it is clear that the patterns of the gland and the background in the benign and malignant gland images are different. For both benign and malignant images, three different categories were created based on the morphological structure of the gland in the histopathological images: the inside region, the boundary region and the outside part of the gland respectively. The labels for these three classes were not provided in the original database, and so have to be created before extracting the patterns of the three classes. The following section indicates the process of creating the labels for these three categories, for benign and malignant gland images respectively. The details of three target categories pixel-level classification are shown in Appendix E.

The results of three target classes classification of benign cases are shown in Appendix E, Section E.2, and the results of malignant cases using three target classes are shown in Appendix E, Section E.4.

Benign categories

This section discusses the process of creating the labels for the three categories for benign gland images. Figure 5.8 indicates the sample images from the benign category in the training database, and the corresponding label image is located on the right-hand side of each image. Different colours represent different gland objects in corresponding images, and the black area in the label image indicates the label of the background.

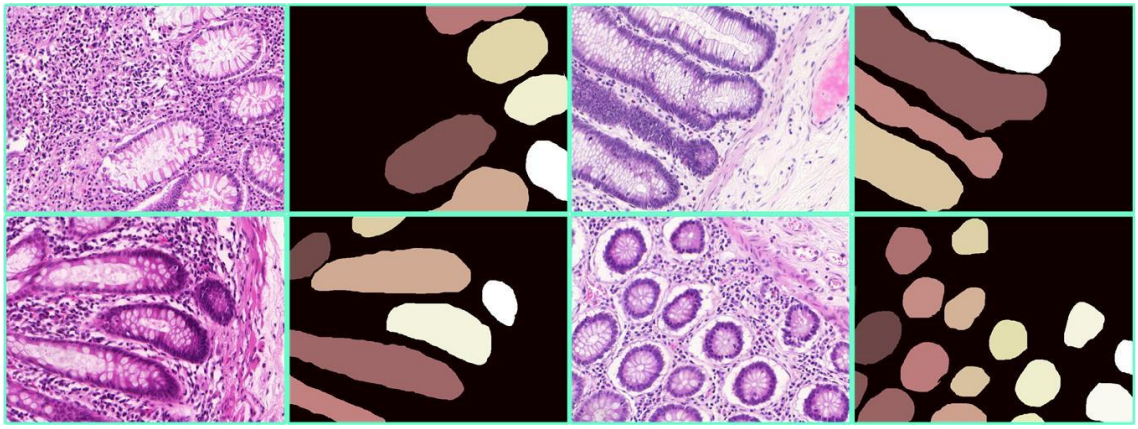


Figure 5.8 Sample images and corresponding ground truth from gland dataset

This segmentation problem is to classify glands by the inside region, the boundary or the outside part in the testing images. The inside region labels are generated from those provided for the original training images. A 15-iteration image erosion operation (Van Den Boomgaard and Van Balen, 1992) is employed to get only the inside of benign glands. The reason for choosing 15 iterations is that the value comes from the experiments; the significant pattern of the inside gland in images with benign tissue is that there are white bubbles. If there are no white bubbles in the original gland labels, they are manually labelled as boundary labels.

Figure 5.9 indicates the sample images and the corresponding labels for the inside gland class, and the black area in the label image indicates the background. The label for the background in benign category images is same as in the original gland dataset; different colours in the image represent different inside gland objects.

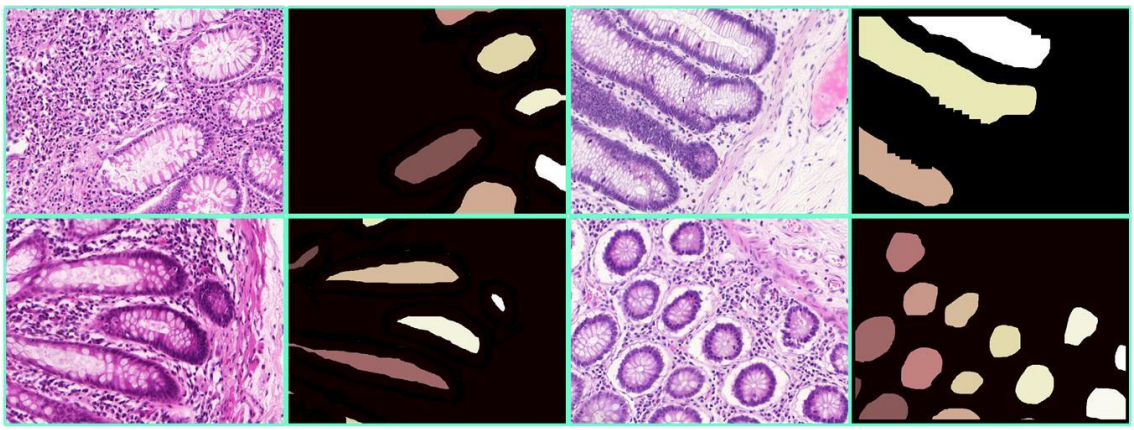


Figure 5.9 Sample images and the inside gland labels of corresponding images

The outside labels in the histology images are the same as the background labels provided in the histology images, indicated by the black area shown in the ground truth in Figure 5.8. The boundary label for benign gland images is generated by performing an XOR operation (Davies, 2002) between the outside label and the inside gland labels. Figure 5.10 illustrates the sample image and the boundary labels; similarly different colours indicate different gland boundaries.

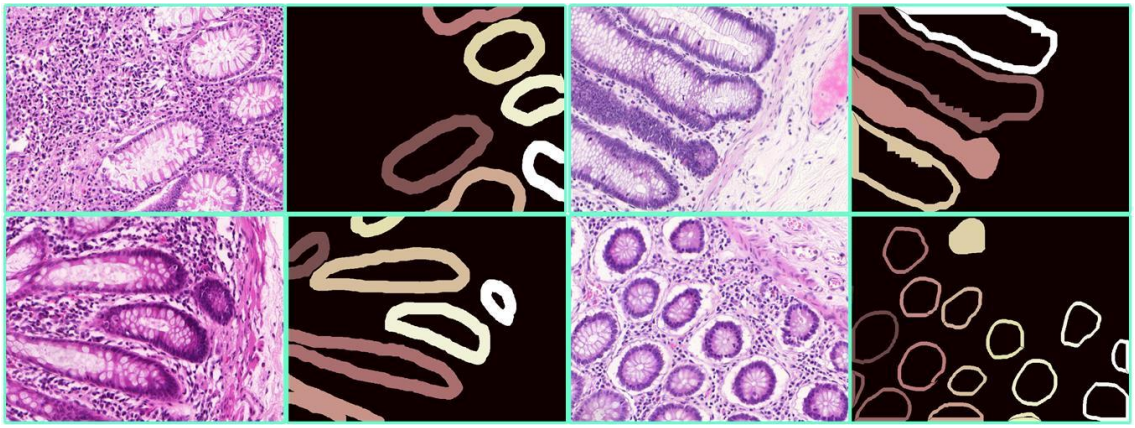


Figure 5.10 Sample image and the boundary label for corresponding image

Malignant categories

For the three malignant categories segmentation approach, the features were extracted from these three classes in the malignant gland images. The morphological structure of the gland in malignant images is different from that in images with benign tissue, but the three labels are generated in a similar way.

The inside labels are determined using 20 image erosion (Van Den Boomgaard and Van Balen, 1992) iterations of the original gland labels; the cell walls in malignant images

are broad, and this figure of 20 fits the entire range. Figure 5.11 indicates sample images and corresponding inside labels, and different colours illustrate different gland inside regions.

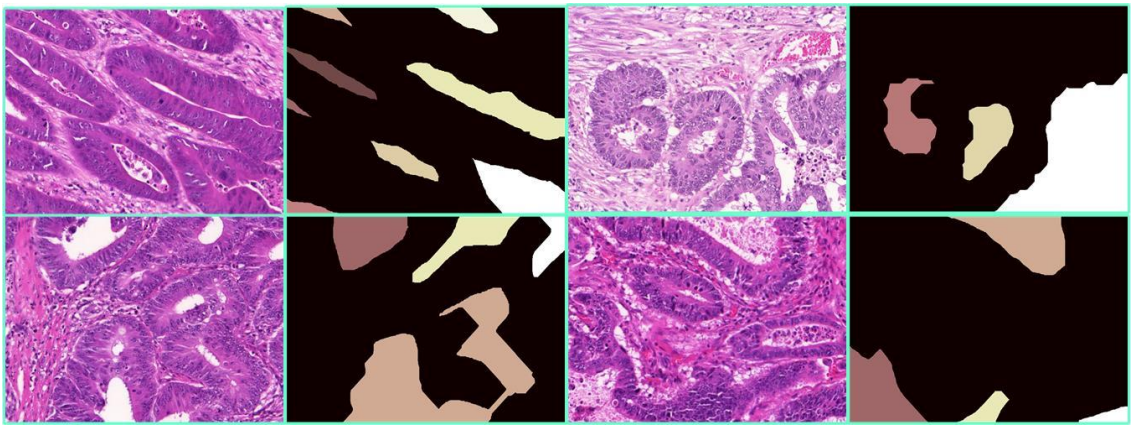


Figure 5.11 Sample images and inside labels for malignant category

A similar approach was used to generate the boundary label for malignant category gland images, an XOR operation between the inside and outside labels. Figure 5.12 illustrates the sample images and corresponding boundary labels. The different colours illustrate the boundary for different gland objects.

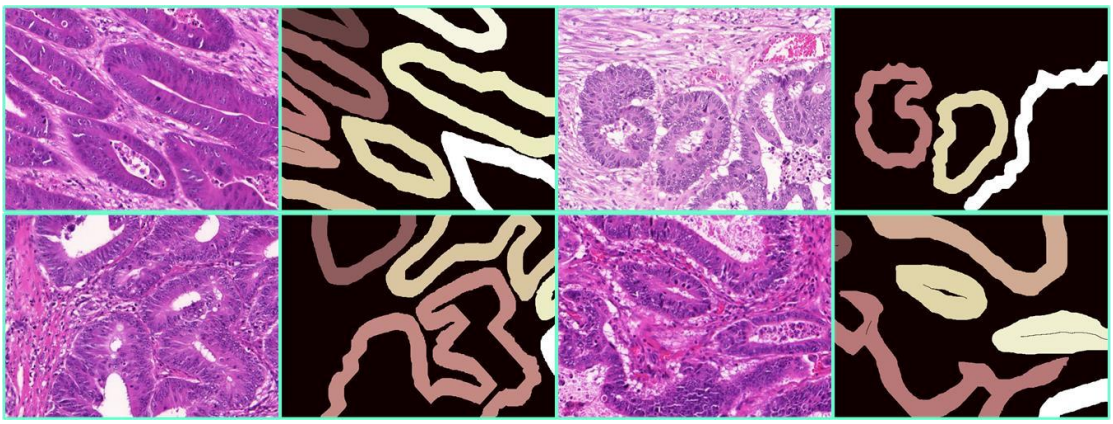


Figure 5.12 Sample images and boundary labels for malignant category

The features for these three categories of malignant images are generated by applying the feature extraction methods demonstrated in Chapter 4. The forest models use these features to train and to make the predictions.

5.4.2.3 Segmentation with pre-classification at different levels

Segmentation with pre-classification method is described at the beginning of Chapter 5 Section 5.4. That method could be used with both hand-crafted and deep

learning features. In this section, two segmentation methods with pre-classification at different level are introduced. Method 1 (on page 103) shown in Figure 5.13 is the typical method for solving gland segmentation, and this method is named as segmentation without pre-classification in this work. As shown in Figure 5.13, the difference between methods 2 and 3 is that for the method 2 extracted local patterns separately to train the feature extraction methods for the benign and malignant gland images, whereas for the method 3 the features are learnt for all images and the separation between the benign and malignant is preformed of the feature level with two separate random forest models.

Both methods 2 and 3 are using GoogleNet for extracting the local patterns. These two variants of the segmentation with pre-classification should only be used with deep features, as there is no difference between them when using hand-crafted features.

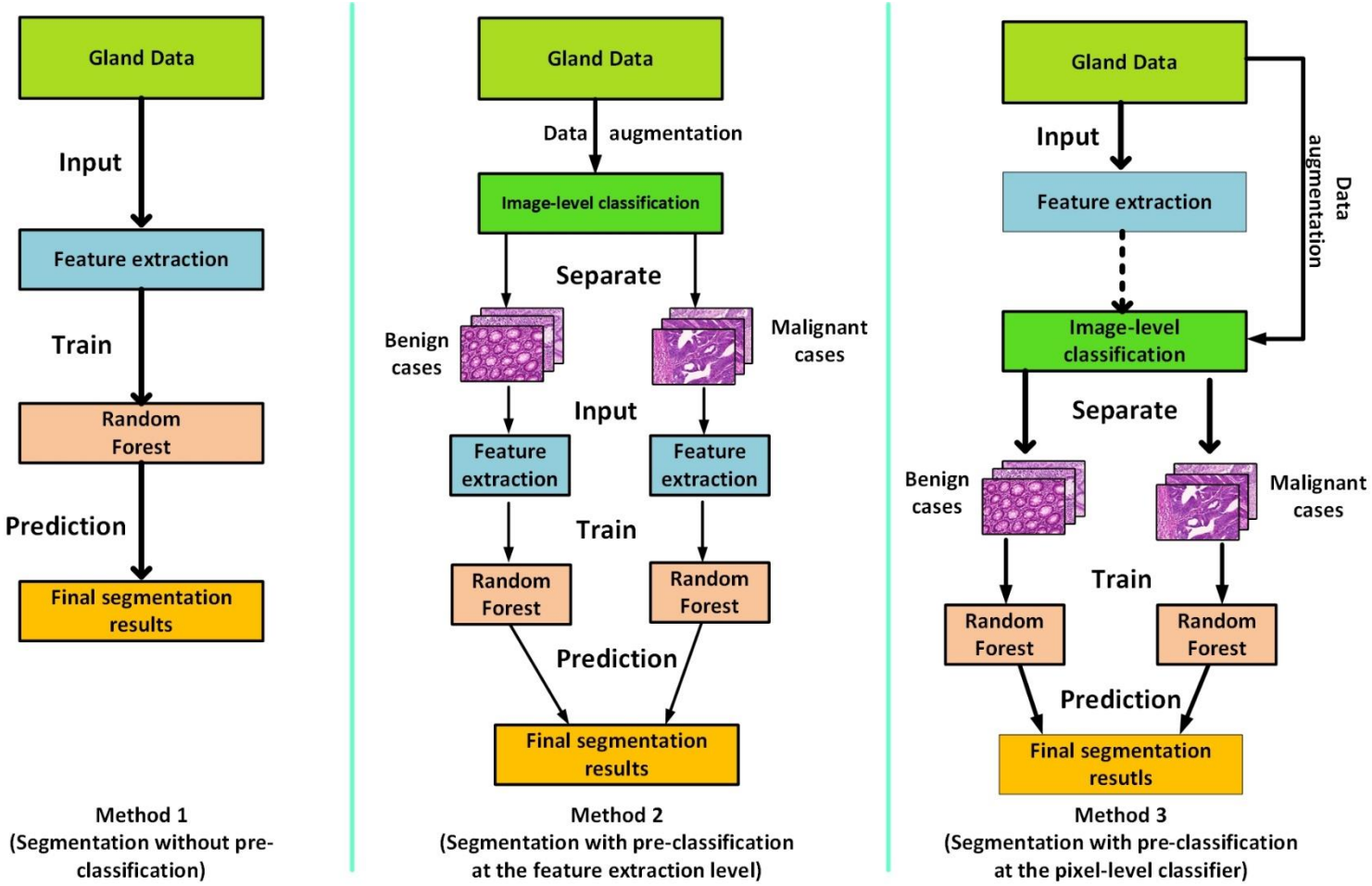


Figure 5.13 The details of three segmentation methods used for gland instance segmentation

5.5 Pre-processing

The details of preparation of the histology image were provided in Chapter 2. Different individuals stained the images in different way so that the colours would not be at the same level. From the sample images from the gland dataset shown in Chapter 2, the images are varying in colours and tones throughout the whole dataset. The pre-processing method is important and necessary in order to reduce the need for pre-categorisation colour harmonisation. Histogram matching is a widely used method for dealing with the issue of colour disparity (Veta et al., 2015).

Histogram matching refers to matching the histogram of the target image to one reference histogram according to certain rules. Histogram equalisation (Hum et al., 2014) is a special form of histogram matching using discrete distribution. This method is effective for images whose histogram is densely distributed. In this work, the reference histogram is the mean of all the images in the training part. Using the mean histogram as reference, the histogram of the target image will be matched against it. The output of this process is the results of the pre-classification work. From the various components discussed in these histology images, the cytoplasm in malignant images is one of the texture features. To preserve this texture information in the results, large white areas are not retained, but are replaced by black pixels. Figure 5.14 shows an example of the white area in histology images needing or not needing to be removed.

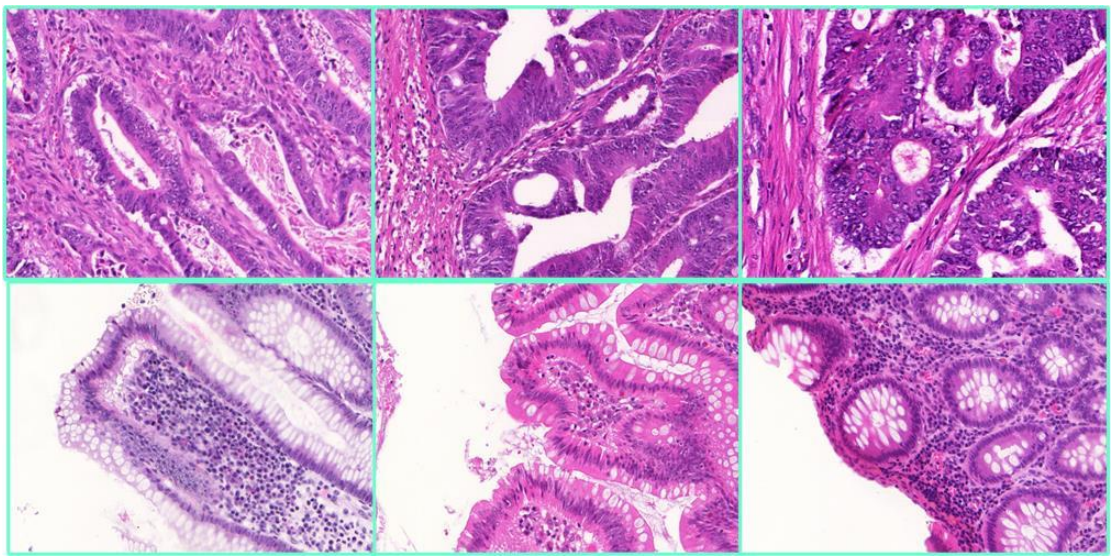


Figure 5.14 Sample histology images with white areas that need/don't need to be removed

Top row images in Figure 5.14 contain white areas but do not need to be replaced by the black pixels; to the bottom row images in contain large white areas that do need to be replaced by black pixels. The method used for removing unwanted extensive white areas in the images was by setting the thresholds in the green channel of RGB images. Figure 5.15 shows the results of removing the white areas identified in the previous figure.

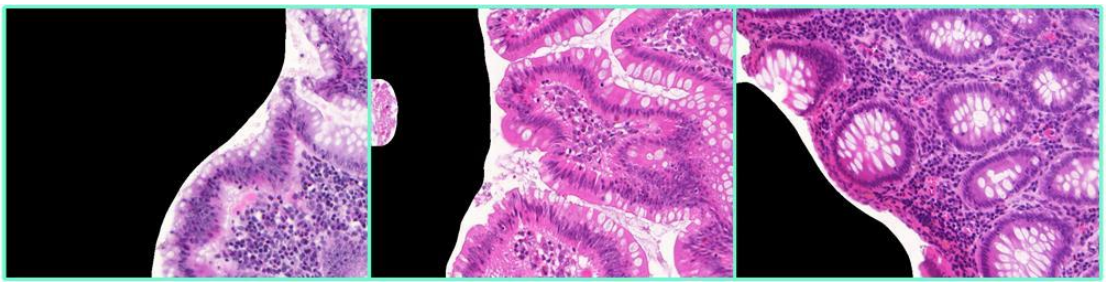


Figure 5.15 The results of removing unwanted white areas in histology images

Figure 5.16 shows samples of output after using the histogram matching method. The top row images are benign after histogram correction, and the images in bottom row are malignant after correction.

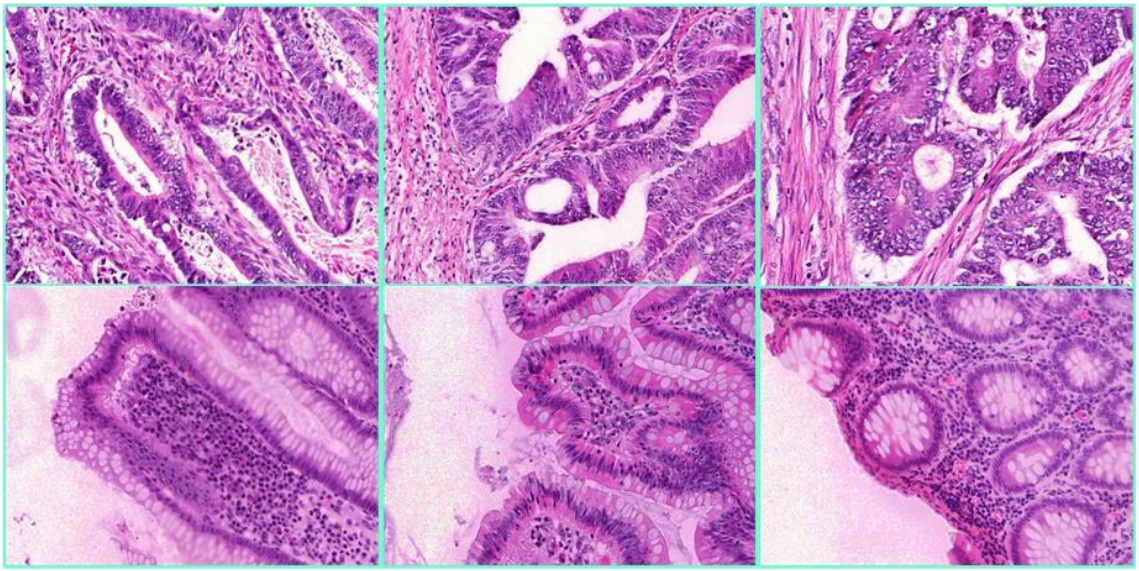


Figure 5.16 Sample images after using histogram correction

Comparing the sample images shown in Figure 5.16 and those in Figure 5.14, the colours and tones in the latter are at the same level. To identify the effect of the variation in the colour and the tones of gland objects, the experiment results of methods with/without colour correcting are shown in Appendix E , Section E.7.

5.6 Post-processing

The output for machine learning and deep learning techniques, and also for the random forest techniques, is probability maps. The output of the image segmentation problem aims to categories each pixel in the testing images in a class that is already known in training data based on learning the distribution. Figure 5.17 presents the examples of the probability maps generated by random forest using different features in the segmentation without pre-classification method. The probability maps generated by random forest using different features in segmentation with pre-classification are not shown.

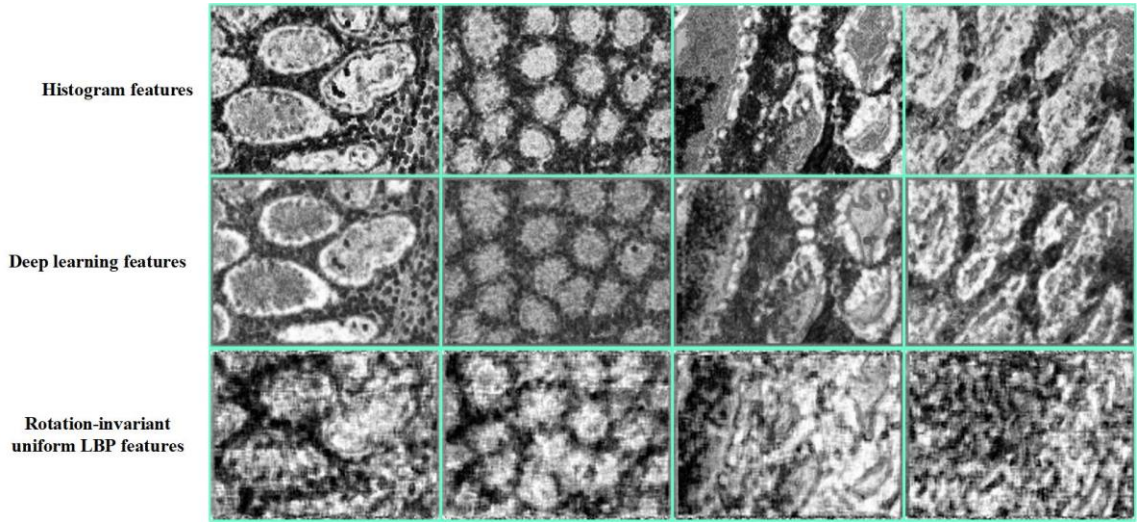


Figure 5.17 Sample of the probability maps in segmentation without pre-classification

The closer the colour to white, the higher the probability of the pixel belonging to a gland. To achieve better segmentation results, probability maps used as segmentation results would not provide a good quantitative result. If the segmentation results (generated by using the morphological post processing method) are choosing as the results, the quantitative results would improve. In this case, post-processing involves a set of operations to remove the imperfections. Two post-processing methods were used in this work to transform the probability maps into segmentation results (label image): morphological post-processing and the level set approach, discussed in detail in the following sections.

5.6.1 Morphological post-processing

Morphological post-processing is a set of non-linear operations relevant to the shape or the morphology of an image. It aims to remove imperfections by determining the form and structure of the images, with methods including dilation, erosion, opening and closing. In this research, a set of morphological post-processing methods were used to process the probability maps generated by random forest using different types of the features. From the above discussion, two proposed segmentation methods were used to solve gland segmentation.

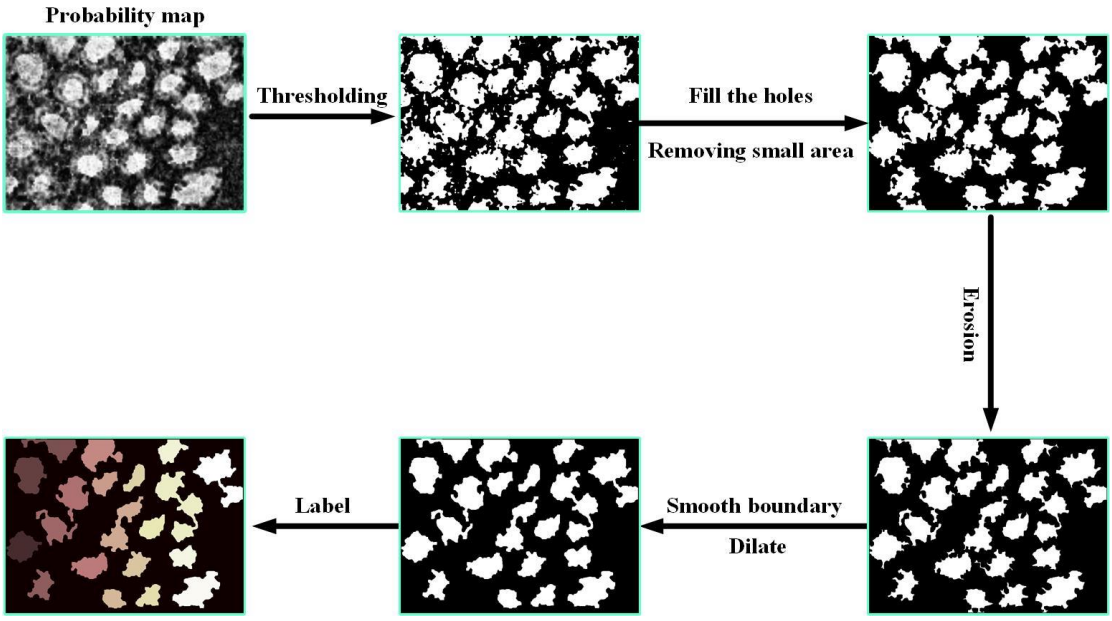


Figure 5.18 Morphological post-processing of processing probability maps

Figure 5.18 illustrates the details of morphological post-processing of the probability maps. The top left image illustrates one of the probability maps generated from random forest, and the top middle image illustrates the binary image after thresholding. The top right image illustrates filling the holes in the binary image and removing the small objects. The bottom right image shows the results of eroding objects in the previous image. The bottom middle image shows the results of smoothing the gland objects, and the bottom left image shows the segmentation results after labelling different gland objects.

To find the best segmentation performance of different segmentation methods with different features, several experiments with different value of the parameters are

designed. The best performance of different methods using different features are provided by using different value of parameters in the experiments. For different experiments, the values of the parameters are different for different segmentation method. The source code of these experiments could be found in accompanying software.

In segmentation with pre-classification, post-processing method used a pixel-level classification method. For both benign and malignant images, the steps in processing the probability maps are the same as those shown in Figure 5.15. However, the parameters used in each step were different from those used in above. Again, all the best segmentation results were tested based on multiple experiments and the best segmentation results based on the evaluation measures, described in detail in Chapter 6.

5.6.2 Level set algorithm

This approach is also widely used in medical image processing, and the active contour models for segmentation have become increasingly popular. The first active contour model was introduced by Kass et al. (1988), has developed dramatically over the last two decades. It iteratively involves the initial curve of the boundary of the target objects. The level set algorithm could handle topological changes of the object contour, and is easily adapted to any dimensional segmentation problem (Zhang et al., 2008). The method was also used in this work to process the probability maps, combined with the geodesic active contour model (Caselles et al., 1997) for boundary-based segmentation and the Chan-Vese model (Chan and Vese, 2001) for region-based segmentation.

Before discussing the details of the level set algorithm, some of the concepts are introduced. The function ϕ is used to indicate the active contour $\boldsymbol{C} = \{x \mid \phi(x) = 0\}$ and the points inside or outside this contour have either positive or negative values. The direction of the curve normal $\vec{\boldsymbol{N}}$ determines the point of propagation of the contour. Figure 5.19 illustrates the conventions used in the level set algorithm.

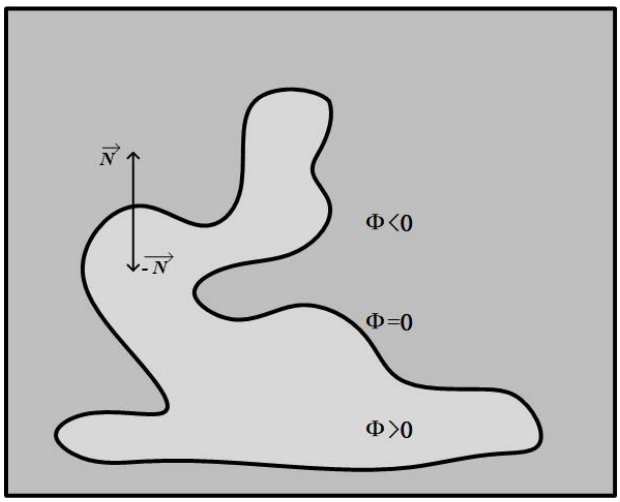


Figure 5.19 Conventions used in the level set algorithm (Zhang et al., 2008)

Most of the active contour algorithm is to minimise the function which defines the contour close to the target boundary with small values. As one of the active contour algorithms, the level set algorithm minimises the function as described by (Zhang et al., 2008):

$$\varepsilon(\phi) = -\alpha \int_{\Omega} (\boldsymbol{I} - \mu) H(\phi)\, d\Omega + \beta \int_{\Omega} g|\nabla H(\phi)|\, d\Omega \qquad (5.7)$$

where I is the binary image after applying simple thresholding to the probability maps generated by the random forest technique $g = g|\nabla \boldsymbol{I}|$ is the boundary feature of the maps relevant to the image, Ω indicates the image domain. α is a pre-defined parameter to represent the low boundary of the target object. The first term of the function is to confine the contours to the region with a grey level bigger than μ. The assumption for this process is that the grey-level value of the target contour is high; otherwise, a grey-level mapping approach will be used. The second term is the geodesic active contour function, used to control the contour to attach it to the region with a high gradient.

The partial differential equation (PDE) can be determined from the directional derivative applied to the function (5.7) given by the equation (Zhang et al., 2008):

$$\phi_t = |\nabla \phi| \left[\alpha(\boldsymbol{I} - \mu) + \beta \mathrm{div}(g \frac{\nabla \phi}{|\nabla \phi|})\right] \qquad (5.8)$$

Using an identity of $\mathrm{div}(g\vec{f}) = \langle \nabla g | \vec{f} \rangle + g\mathrm{div}(\vec{f})$, equation (5.8) could be written as described in the reference (Zhang et al., 2008):

$$\phi_t = |\nabla\phi| \left[\alpha(\boldsymbol{I} - \mu) + \beta \langle \nabla \mathrm{g}, \frac{\nabla\phi}{|\nabla\phi|} \rangle + \beta g \mathrm{div}(\frac{\nabla\phi}{|\nabla\phi|})\right] \quad (5.9)$$

This equation can be rewritten as follows because in Osher et al. (1988) the $C_t = \gamma \vec{\boldsymbol{N}}$ and $\phi_t = \gamma|\nabla\phi|$ represent the same curve envolution, it is given by (Zhang et al., 2008):

$$C_t = \alpha(\boldsymbol{I} - \mu)\vec{\boldsymbol{N}} - \beta\langle \nabla \mathrm{g}, \vec{\boldsymbol{N}} \rangle \vec{\boldsymbol{N}} + \beta g \, \kappa \vec{\boldsymbol{N}} \quad (5.10)$$

where $\vec{\boldsymbol{N}}$ is the normal vector and $\vec{\boldsymbol{N}} = \frac{-\nabla\phi}{|\nabla\phi|}$, and $\kappa = \mathrm{div}(\frac{\nabla\phi}{|\nabla\phi|})$ is the curveture. The first term in the above equation is to control the propagation movement of the curve. The second term is to move the curve to the target object contour. The third term is to control the smoothness of the curve. If $|\nabla\phi| = 1$, the equation (5.8) could be simplified as given by (Zhang et al., 2008):

$$\phi_t = \alpha(\boldsymbol{I} - \mu) + \beta \, \mathrm{div}(\mathrm{g} \, \nabla\phi) \quad (5.11)$$

All these equations have been used in the provided post-processing (level-set) method in this work. The reason for listing these equations is to help understand the development of level-set proposed by Zhang et al. (2008).

Figure 5.20 indicates the steps in using the level set algorithm to process the probability maps generated from the pixel-level classifier, random forest, using different features extracted from the histology images.

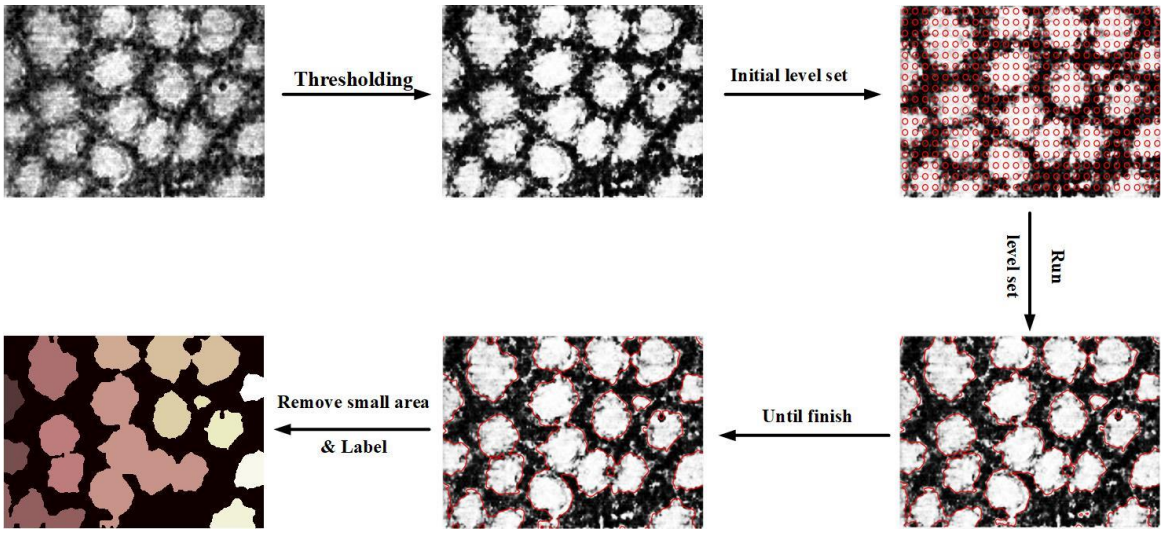


Figure 5.20 Steps in the level set post-processing method

The top left image is one of the probability maps generated by random forest, the middle image is the output after applying the probability map, and the top right image is the initial level set parameters. The right bottom image is the first iteration of the level set results, and the middle image is the output of the level set algorithm. The bottom right image is the final segmentation result for this image, using different colours to represent different gland objects.

For different values of the propagation and smooth parameters in this algorithm, the segmentation results are also affected. The value of the propagation parameter is 15 and the value of the smooth parameter is 90; these values were selected after comparing the values of the corresponding best results.

5.7 Summary

This chapter discussed gland segmentation without pre-classification and its limitation, that there is no understanding of the texture features in the images in the gland dataset. To overcome this limitation, an improved version of the method, segmentation with pre-classification, was introduced. The reason for using this method is that the morphological structure in benign and malignant images is different, and this method helps the random forest classifier to learn the specific rule for differentiating the gland and the background in benign or malignant images. Segmentation with pre-classification has two steps: image-level and pixel-level classification. For the image-level

classification step, the histology images are partitioned into benign and malignant category. In the pixel-level classification step, the segmentation results are generated from processing the probability maps using morphological post-processing; the probability maps were predicted by random forest. The random forest classifier was trained by the features extracted from only benign or malignant images.

With the various colours and tones in the original histology images, the pre-processing method used to achieve colour harmonisation. The probability maps were generated from random forest, and they could not be used directly as segmentation results. Two post-processing methods were used to process the probability maps to obtain the segmentation results, as discussed at the end of the chapter.

In the next chapter, the results of these two proposed segmentation methods are evaluated.

Chapter 6

Results

Chapter 6 focuses on examining different evaluation metrics for instance segmentation, and performance evaluation of different segmentation methods on the gland database. Random forest model is used as the primary classifier in this work. The reasons behind choosing random forest were discussed in the previous chapter (Chapter 3). The various data features and random forest design parameters are also discussed. The results of segmentation with and without pre-classification are then presented. Finally, the best results of segmentation with and without pre-classification are stated and compared using the introduced evaluation metrics.

The gland instance segmentation problem, including: clinical problem justification (e.g. explanations of the clinical significance of the gland segmentation), clinical data collection and baseline assessment (including ground truth segmentation results) were defined by clinical pathologists from Warwick University involved in development of the dataset. The reported work is wholly focused on solving the segmentation problem. Additional work will be needed to translate the proposed segmentation technologies into clinical practice. This though was considered to be outside the scope of the research.

6.1 Evaluation metrics

The key objective of this section is to describe different metrics used for evaluation of instance segmentation methods. Two types of metrics are described: region-based and contour-based.

6.1.1 Region-based evaluation metrics

In gland segmentation problems, two region-based evaluation measures are frequently used: **F1 score** and **object-level Dice index**. **F1 score** is used to evaluate the detection accuracy of gland objects. **Object-level Dice index** is used for assessing the overall area-based segmentation accuracy of individual gland object. These evaluation

measures were introduced in Sirinukunwattana et al. (2015).

The **F1 score** is employed to estimate the detection accuracy of glands. A segmented gland structure which has more than 50% overlap with its ground truth objects is treated as a true positive (**TP**); otherwise, it is treated as a false positive (**FP**). The difference between the number of ground truth and true positive objects is treated as the number of false negatives (**FN**). The mathematical expression for the F1 score is defined by (Sirinukunwattana et al., 2017):

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{6.1}$$

where the mathematical expressions for precision and recall are given by:

$$Precision = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \tag{6.2}$$

$$Recall = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{6.3}$$

where $\boldsymbol{TP}$ indicates the number of true positives, $\boldsymbol{FP}$ represents the number of false positives, and $\boldsymbol{FN}$ represents the number of false negatives.

Suppose $\boldsymbol{s}$ represents a set of segmented glands, and $\boldsymbol{g}$ indicates a corresponding set of ground truth. A function $\boldsymbol{G_*} : \boldsymbol{s} \rightarrow \boldsymbol{g}$ is introduced for each segmented gland $S \in \boldsymbol{s}$, $\boldsymbol{G_*}(S) = G \in \boldsymbol{g}$ where G has the largest overlapping area with S. If there is no overlap between the segmented gland and the ground truth, the value of this function is an empty set. A similar function $\boldsymbol{S_*} : \boldsymbol{g} \rightarrow \boldsymbol{s}$ is used, for each $G \in \boldsymbol{g}$, $\boldsymbol{S_*}(G) = S \in \boldsymbol{s}$ where S has the largest overlapping area with the ground truth G.

The **object-level Dice index** evaluates the segmentation accuracy of the results. Given $\boldsymbol{G}$ as a set of pixels belonging to the ground truth of the gland object, and $\boldsymbol{S}$ representing a set of pixels belonging to the segmented gland, the mathematical expression for the Dice index is defined by (Dice, 1945):

$$Dice = \frac{2\ (|\boldsymbol{G} \cap \boldsymbol{S}|)}{|\boldsymbol{G}| + |\boldsymbol{S}|} \tag{6.4}$$

However, the Dice index in (6.4) is defined on the pixel level, not the object level. In this work, an object-level Dice index is employed to evaluate the segmentation

performance (Sirinukunwattana et al., 2015). The object-level Dice index is introduced to distinguish between different instances of the gland. Suppose n_g indicates the number of ground truth occurrences for gland objects in the test image, manually labelled by pathology experts and n_s is the number of segmented objects. G_i is the ground truth for the ith gland object, and S_p represents the segmentation results for the pth segmented object. The object-level Dice index is defined by (Sirinukunwattana et al., 2017):

$$Dice_{\text{obj}}(\boldsymbol{g},\boldsymbol{s}) = \frac{1}{2}\left[\sum_{i=1}^{n_g}\omega_i Dice\left(G_i,\boldsymbol{S}_*(G_i)\right) + \sum_{i=1}^{n_s}\widetilde{\omega_q}\, Dice\left(\mathbf{G}_*(S_p),S_p\right)\right] \quad (6.5)$$

where

$$\omega_i = {|G_i|}\Big/{\sum_{j=1}^{n_g}|G_j|}\ ,\ \widetilde{\omega_q} = {|S_q|}\Big/{\sum_{p=1}^{n_s}|S_p|} \quad (6.6)$$

$Dice\left(G_i,\boldsymbol{S}_*(G_i)\right)$ in equation (6.5) is to evaluate the overlapping area between each ground truth and the corresponding segmented object, and $Dice\left(\mathbf{G}_*(S_p),S_p\right)$ is to evaluate the overlapping area between each segmented object and its corresponding ground truth. Both of the terms are weighted by the corresponding area of glands, providing less significance to small segmented gland and small ground truth objects. The functions $\boldsymbol{G}_*$ and $\boldsymbol{S}_*$ are explained in previous paragraph.

6.1.2 Contour-based evaluation metrics

In gland instance segmentation problem, object-level Hausdorff distance is one of the contour-based evaluation measures used. This method was used as one of the evaluation measures in the MICCAI 2015 Gland Segmentation Challenge. However, another evaluation contour-based measure used in this work is the Boundary Jaccard (BJ) metric, introduced by Fernandez-Moral et al. (2018). The contour-based evaluation measures are generated based on the distance between two contours. The reason for the object-level BJ index being better than the object-level Hausdorff distance in evaluation of the shape similarity for gland segmentation is discussed in Chapter 6, Section 6.1.3. The following part describes these two contour-based evaluation measures.

Object-level Hausdorff distance is often employed to evaluate the shape similarity between the ground truth and the segmented object. The general mathematical expression of the Hausdorff distance between ground truth G and segmentation results S is defined by (Beauchemin et al., 1998):

$$\mathbf{H}(\mathrm{G},\mathrm{S}) = \max\{sup_{x\in G} inf_{y\in S}\|x-y\|, sup_{y\in S} inf_{x\in G}\|x-y\|\} \tag{6.7}$$

where $\|\cdot\|$ indicates the Euclidean distance between pixels $x \in G$ and $y \in S$.

Such defined measure is though not suitable for evaluation of instance segmentation method. In case of instance segmentation, the object-level Hausdorff distance needs to be used. The object-level Hausdorff distance is defined by (Sirinukunwattana et al., 2017):

$$\mathbf{H}_{\mathrm{obj}}(\boldsymbol{g},\boldsymbol{s}) = \frac{1}{2}\left[\sum_{i=1}^{n_g} \omega_i \mathbf{H}\big(G_i, \boldsymbol{S}_*(G_i)\big) + \sum_{i=1}^{n_s} \widetilde{\omega_q}\, \mathbf{H}\big(\boldsymbol{G}_*(S_p), S_p\big)\right] \tag{6.8}$$

where the meaning of the mathematical symbols used in equation 6.8 is the same as in equation 6.6. If a ground truth G has no corresponding segmented gland (where $\boldsymbol{S}_*(G) = \emptyset$), the Hausdorff distance is calculated between the ground truth G and the nearest segmented gland $S\,\epsilon\,\boldsymbol{s}$ instead. A similar approach applies for a segmented object which has no corresponding ground truth object.

BJ metric was introduced by Fernandez-Moral et al. (2018), and this evaluation measure is sensitive to infra-segmentation and over-segmentation results. This metric is developed from the **BF metric,** which was described in Csurka et al. (2013), which has two main disadvantages. If the distance between the two contours is greater than the threshold θ, the value of the BF metric is 0. Otherwise, the value of this metric will be close to 1; in short, this metric is not continuous. Second, the value of this measure will be the same if the same number of boundary pixels are in the distance θ.

The **BJ measure** can handle these two drawbacks, and this value is used to calculate the distance from the boundary in ground truth to the boundary in segmentation results $B_{gt}^c \rightarrow S_{ps}^c$ for the class c, to obtain the number of true positives ($\boldsymbol{TP}_{B_{gt}}^c$) and false negatives ($\boldsymbol{FN}^c$). Similarly, the distance from the boundary in segmentation results to

the boundary in ground truth $B_{ps}^{c} \rightarrow S_{gt}^{c}$ for the same class c, to determine the number of true positives ($\boldsymbol{TP}_{B_{ps}}^{c}$) and false positives ($\boldsymbol{FP}^{c}$). The total number of true positives is defined as ($\boldsymbol{TP}^{c} = \boldsymbol{TP}_{B_{ps}}^{c} + \boldsymbol{TP}_{B_{gt}}^{c}$). The values of these parameters are defined by (Fernandez-Moral et al., 2018):

$$\boldsymbol{TP}_{B_{gt}}^{c} = \sum_{x \in B_{gt}^{c}} z \quad \text{with} \quad z = \begin{cases} 1 - \left(\frac{d(x, S_{ps}^{c})}{\theta}\right)^{2} & \text{if } d\left(x, S_{ps}^{c}\right) < \theta \\ 0 & \text{otherwise} \end{cases} \tag{6.9}$$

$$\boldsymbol{FN}^{c} = \left|B_{gt}^{c}\right| - \boldsymbol{TP}_{B_{gt}}^{c} \tag{6.10}$$

$$\boldsymbol{TP}_{B_{ps}}^{c} = \sum_{x \in B_{ps}^{c}} z \quad \text{with} \quad z = \begin{cases} 1 - \left(\frac{d(x, S_{gt}^{c})}{\theta}\right)^{2} & \text{if } d\left(x, S_{gt}^{c}\right) < \theta \\ 0 & \text{otherwise} \end{cases} \tag{6.11}$$

$$\boldsymbol{FP}^{c} = \left|B_{ps}^{c}\right| - \boldsymbol{TP}_{B_{ps}}^{c} \tag{6.12}$$

The **Boundary Jaccard metric** is defined according to the BJ index described by (Fernandez-Moral et al., 2018):

$$\boldsymbol{BJ} = \frac{\boldsymbol{TP}^{c}}{\boldsymbol{TP}^{c} + \boldsymbol{FN}^{c} + \boldsymbol{FP}^{c}} \tag{6.13}$$

The BJ metric described above is used to evaluate semantic segmentation methods. In this work, this Boundary Jaccard (BJ) index has been extended for evaluation of instance segmentation. In a way similar to the extensions proposed for Dice index and Hausdorff distance. The object-level BJ index proposed in this work is an extension of BJ. It is used for evaluation of gland instance segmentation, and is defined by:

$$\boldsymbol{BJ_{ins}}(g, s) = \frac{1}{2}\left[\sum_{i=1}^{n_g} \omega_i \mathbf{BJ}\left(G_i, S_*(G_i)\right) + \sum_{\mathrm{p}=1}^{n_s} \widetilde{\omega_q}\, \mathbf{BJ}\left(G_*\left(S_p\right), S_p\right)\right] \tag{6.14}$$

where $\mathbf{BJ}\left(G_i, S_*(G_i)\right)$ is to determine the BJ Index between the ground truth objects and the corresponding segmented objects. $\mathbf{BJ}\left(G_*\left(S_p\right), S_p\right)$ is to evaluate the BJ index between the segmented objects and the corresponding ground truth.

The object-level BJ index is an extension of BJ, and it has the similar properties as BJ. However, in this work, the experiment results only show that the object-level BJ

index is not sensitive to segmentation outliers, which is not described in the reference (Fernandez-Moral et al., 2018).

The object-level Hausdorff distance was proposed before for evaluation of the gland instance segmentation methods. In this work, the object-level BJ metric is also used as the fourth evaluation measure to assess the instance segmentation results. Both BJ and Hausdorff metrics are used to measure the shape similarity. However, object-level Hausdorff distance is sensitive to the segmentation outliers, whereas BJ is not sensitive to such outliers. The evidence is shown in the following section.

6.1.3 Analysis of evaluation metrics

The following examples help to understand the meaning and significance of different evaluation measures introduced in the previous sections.

Figure 6.1 illustrates the meaning and significance of F1 score.

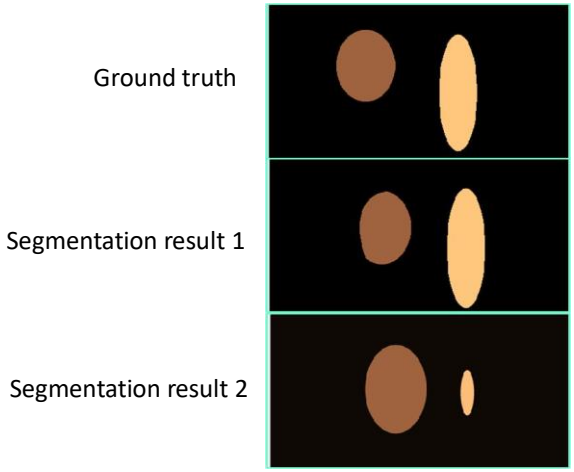


Figure 6.1 F1 score evaluation example

Following the description in Chapter 6 Section 6.1.1, the F1 score used in this work is a region-based evaluation measure. Both segmented gland objects in segmentation result 1 (shown in Figure 6.1) have more than 50% overlapping area with the corresponding ground truth glands, so both of them are treated as **TP**; there is no **FN** or **FP**. The brown object in segmentation result 2 has more than 50% overlap with the ground truth and is treated as **TP** (TP =1). The yellow gland object has less than 50% overlap and so is treated as **FP** (FP = 1). Based on the definition of **FN** in Chapter 6, Section 6.1.1, **FN** is calculated as the difference between the number of ground truth objects and the number of true positives, and therefore in this case FN = 1. Once **TP**, **FP**

and **FN** have been computed, the value of the F1 score is calculated based on equations (6.1-6.3). Table 6.1 shows the F1 score for the two results. As discussed in Section 6.1.1, The range of F1 scores is between 0 and 1, and the closer the value of F1 is to 1 the better are the segmentation results.

Table 6.1 F1 score for the segmentation results in Figure 6.1

Segmentation methods	F1 score
1	1
2	0.5

Figure 6.2 illustrates the meaning and significance of object-level Dice index.

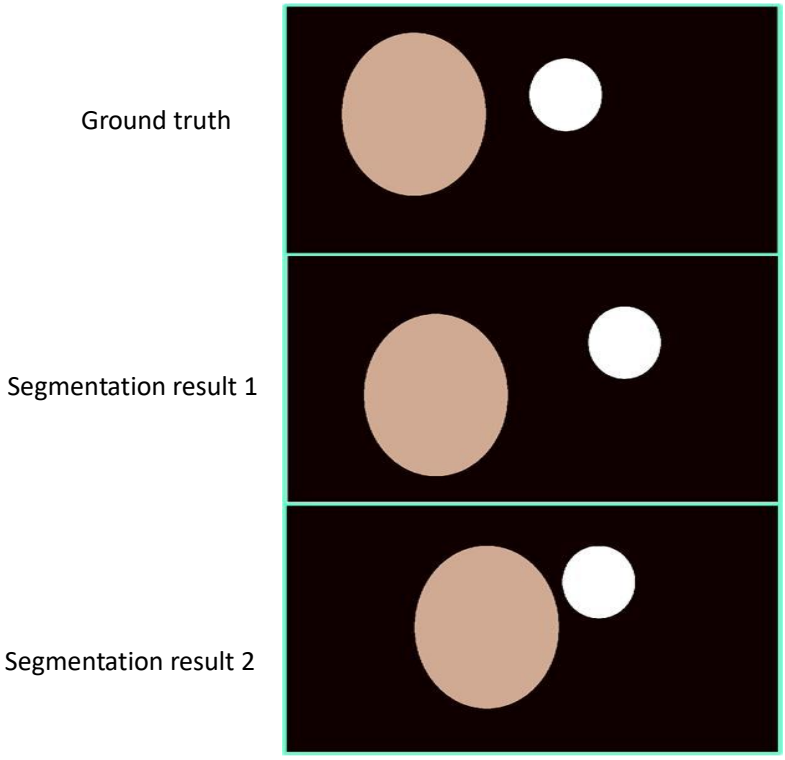


Figure 6.2 Object-level Dice index evaluation example

From visual inspection, segmentation result 1 has more overlap with the ground truth than does segmentation result 2. Table 6.2 confirms this observation showing, the object-level Dice index for results 1 and 2, as respectively 0.595 and 0.389. The closer the value of this parameter is to 1, the more overlap there is between the ground truth and the segmentation results; in this case, segmentation result 1 is better than result 2.

Table 6.2 Object-level Dice index for the results in Figure 6.2

segmentation result	Object-level Dice index
1	0.595
2	0.389

Another evaluation measure used in this research is object-level Hausdorff distance, which measures the shape similarity between segmentation results and ground truth, as illustrated in Figure 6.3.

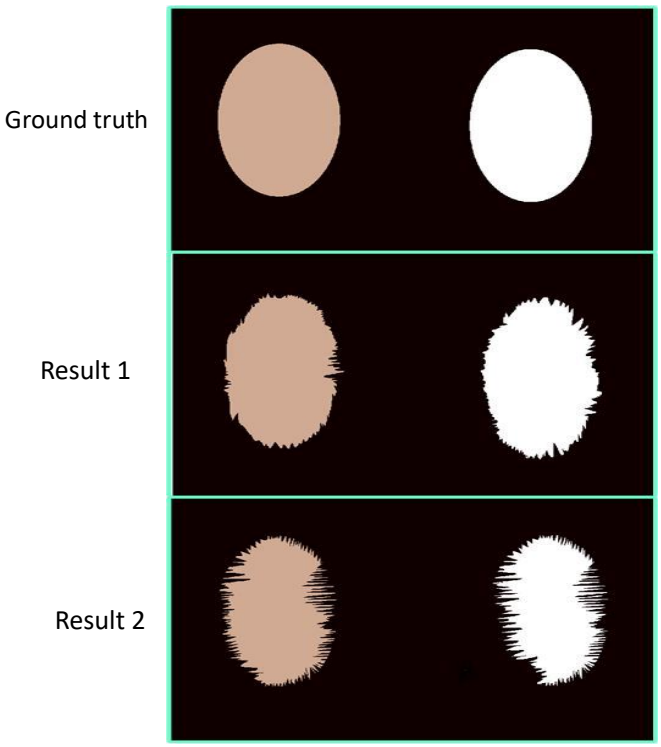


Figure 6.3 Object-level Hausdorff distance evaluation measure

Based on visual inspection, the shapes of the two gland objects in segmentation result 1 are closer to the ground truth than those in segmentation result 2. Table 6.3 shows the object-level Hausdorff distance calculated for the two results. In this case, the smaller the value, the better is the result. The results in Table 6.3 confirm the conclusion from the visual inspection.

Table 6.3 Shape similarity for the segmentation results shown in Figure 6.3

Segmentation result	Object-level Hausdorff distance
1	17.899
2	57.608

The last evaluation measure used in this work is object-level Boundary Jaccard index. Figure 6.4 shows the value of Boundary Jaccard index for infra-segmentation and over-segmentation for two different gland objects.

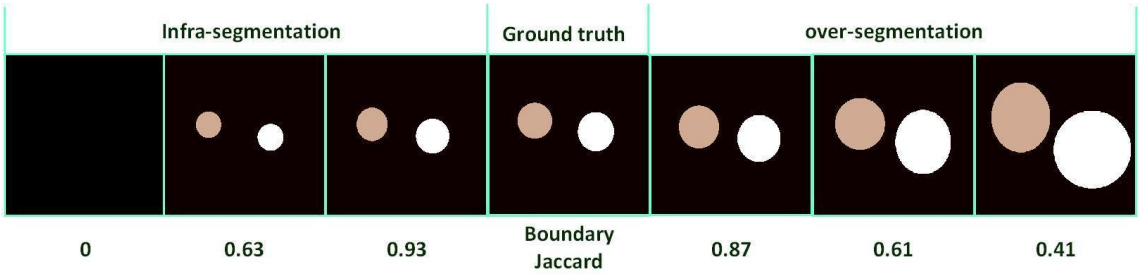


Figure 6.4 Example of infra-segmentation and over-segmentation of two different gland objects

The centre of Figure 6.4 shows the ground truth for the two different gland objects; the right part and the left part of the diagram show over-segmentation and infra-segmentation of the same gland objects respectively. The number below each segmentation result is the corresponding Boundary Jaccard index. As with the F1 score, the range of the boundary Jaccard index is between 0 and 1. The closer this index is to 1, the better the segmentation results are.

Figures 6.1, 6.2, 6.3 and 6.4 provide illustrative examples for an intuitive interpretation (the explanation is based on the simulated examples) of the evaluation measures used in this work. The ground truth and the segmentation results shown in Figures 6.1-6.7 are simulated results, not the real segmented results generated using gland data. These simple examples help to explain the meaning and significance of the metrics for evaluation of gland segmentation results. The following examples (Figures 6.5-6.7) and short discussion illustrate the reasons for using different measures rather than a single measure for evaluation of the gland segmentation results; the reasons why the object-level BJ index is considered to be better than the object-level Hausdorff distance are discussed.

Figure 6.5 illustrates a situation where the values of the F1 score and the object-level Dice index are practically the same, but the value of object-level Hausdorff distance is significantly different between results 1 and 2. Table 6.4 shows the values of the three evaluation measures for the corresponding results.

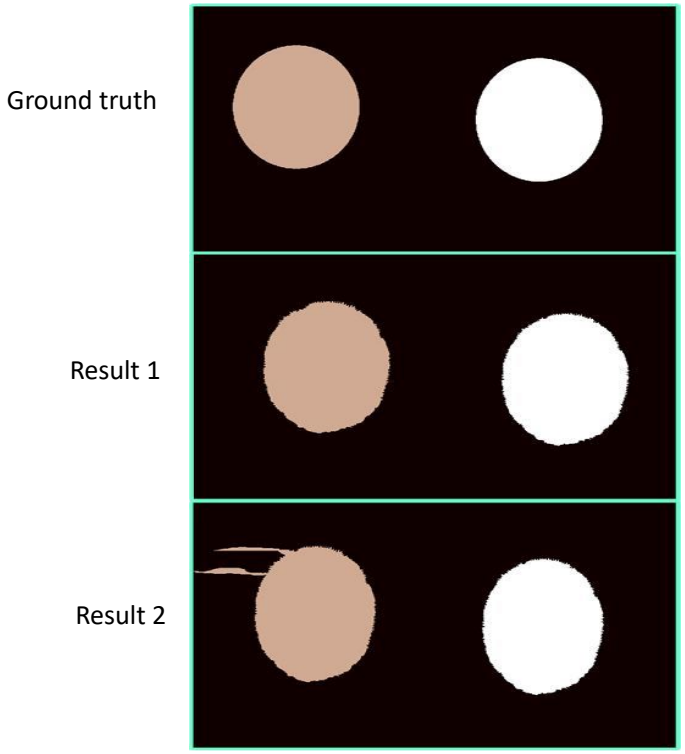


Figure 6.5 Significance of the object-level Hausdorff distance measure

Table 6.4 Evaluation measures for the segmentation results shown in Figure 6.5

Segmentation result	F1 score	Object-level Dice index	Object-level Hausdorff distance
1	1	0.833	**44.640**
2	1	0.833	68.370

If only two evaluation measures, F1 score and object-level Dice index, are used, there is no apparent difference between segmentation results 1 and 2. However, based on visual inspection, the shape of segmentation result 1 is closer to the shape of the ground truth than that of result 2. The shape of the left gland object in result 2 has a small number of outliers. It is discussed before the smaller the value of object-level Hausdorff distance, the better the segmentation is. Table 6.4, thus, indicates that result 1 is better than segmentation result 2.

The following example shows the importance of including F1 score in the evaluation of the gland segmentation.

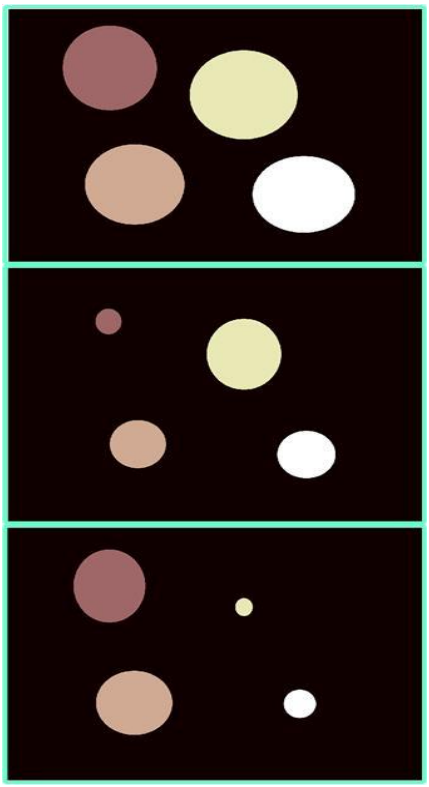


Figure 6.6 Significance of the F1 score

Table 6.5 Evaluation measures for the segmentation results shown in Figure 6.6

Segmentation result	F1 score	Object-level Dice index	Object-level Hausdorff distance
1	0.380	0.531	**64.546**
2	**0.647**	**0.550**	70.012

This time, there are four different gland objects and it is difficult to indentify the best results from visual inspection. If the object-level Dice index and Hausdorff distance are used for evaluation, result 1 has less overlap with ground truth than result 2, but more shape similarity with the ground truth than result 2. If only one evaluation measure is used, the result of the segmentation ranking could be different. Using two measures, results 1 and 2 achieve the same ranking (see Chapter 6 Section 6.2). In this case, it is important to have a third evaluation measure. From detection accuracy, that is the F1 score. The F1 score indicates that result 2 is better than 1 breaking the tie.

The above two examples represent the importance of using both object-level Hausdorff distance and the F1 score in evaluating the segmentation results. The following example shows the importance of using the object-level Dice index: (see Figure 6.7 and Table 6.6)

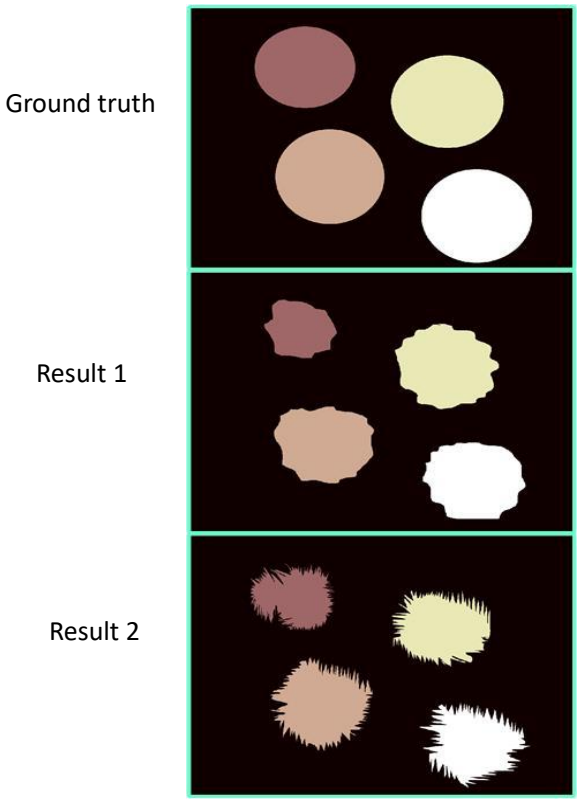


Figure 6.7 Significance of the object-level Dice index

Table 6.6 Evaluation measures for the segmentation result shown in Figure 6.6

Segmentation result	F1 score	Object-level Dice index	Object-level Hausdorff distance
1	0.837	**0.815**	**38.104**
2	**1**	0.767	47.482

For results 1 and 2 shown in Figure 6.7, the object-level Hausdorff distance indicates the result 1 to be better, but using the F1 score, the result 2 seems to be better. If only used one measure, the ranking of these two results could be different. If both of these measures used, results 1 and 2 achieved the same ranking. It is therefore useful to use a third measure, in this case that the best segmentation results based on ranking score is results 1.

The above figures explain why a single evaluation measure for gland segmentation is inappropriate. Figure 6.8 shows why the object-level Boundary Jaccard index is better than object-level Hausdorff distance for evaluation of segmentation performance.

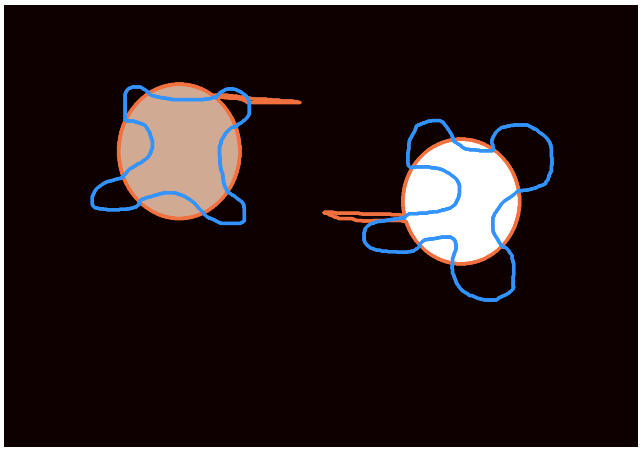


Figure 6.8 Object-level Boundary Jaccard index VS object-level Hausdorff distance

Table 6.7 Evaluation measures for the segmentation result shown in Figure 6.8

Segmentation result	Object-level Boundary Jaccard index	Object-level Hausdorff distance
1 (orange contour)	**0.985**	148.323
2 (blue contour)	0.653	**87.891**

The orange and blue contours are the shapes of two segmentation results. The numbers in bold shown in Table 6.7 are the best results when using different evaluation metrics. Based on a visual inspection, the orange contours for both objects are closer to their ground truth than blue contours. These two metrics have different properties. The object-level BJ index is not sensitive to outliers and is bounded with the possible values between 0 and 1, whereas the object-level Hausdorff distance is sensitive to the and is unbounded. For high-quality evaluation of the segmentation results, the ranking of the segmentation should not change significantly when only a small number of pixels are misclassified. From this point of view the object-level BJ measure is better for evaluation of gland segmentation than object-level Hausdorff distance.

To assess gland instance segmentation results, these four evaluation measures: F1 score, object-level Dice index and object-level Hausdorff distance and object-level Boundary Jaccard index have been introduced, explained and tested. It has been shown that these measures convey complementary information about quality of segmentation results. It is therefore important to use at least a subset of these measures (if not all) when evaluating quality of gland segmentation.

6.2 Ranking strategy

Different feature extraction methods (including ring histogram, rotation invariant uniform LBP, circular Fourier HOG, LeNet5 and GoogleNet) used in this research were discussed in Chapter 4. For each feature extraction method, the corresponding segmentation performance is assigned one ranking score per evaluated metric and test data partition: with three evaluated metrics F1 score, object-level Dice index and object-level Hausdorff distance and two testing subsets of gland images.

Figure 6.9 shows the ground truth and three segmented results (they are all results using histogram feature not simulated data) using the gland data. Table 6.8 shows the values of measures obtained three for each segmentation result.

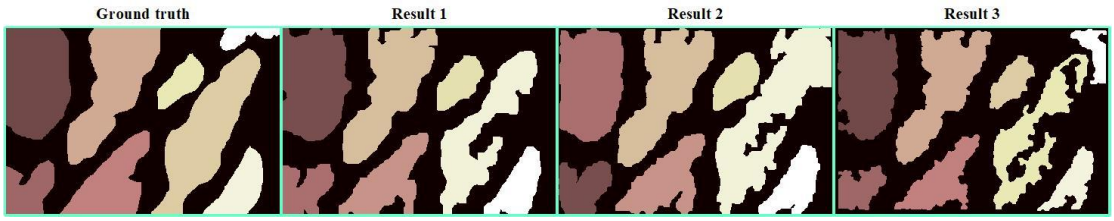


Figure 6.9 Example of the ground truth and three different segmentation results

Table 6.8 Evaluation measures for the segmentation result shown in Figure 6.9

Segmentation Results	F1 Score		Object-level Dice index		Object-level Hausdorff distance		Rank sum
	Value	Rank	Value	Rank	Value	Rank	
1	0.91	1	0.91	1	26.01	1	3
2	0.91	1	0.84	3	30.28	2	6
3	0.91	1	0.88	2	43.32	3	6

From Table 6.8, results 1, 2 and 3 have an object-level Dice index of 0.91, 0.88 and 0.84 respectively. In Chapter 6 Section 6.1.4, it has been explained that the bigger the value of object-level Dice index, the better the segmentation performance is. The ranking

scores are therefore 1, 2 and 3 respectively. The rankings of the object-level Hausdorff distance and F1 score for the different segmentation results are determined similarly.

The aim of this work is to find a stable method for gland segmentation, but using only the mean value is not sufficient to evaluate a method stability. Therefore, both mean and the standard deviation are used in order to evaluate different configuration of the segmentation methods. The numbers in brackets listed next to mean values correspond to computed standard deviation obtained for each corresponding metric on the test sample used in the experiments. The methods with low standard deviation and high mean value are considered to perform well.

The lowest value of the rank sum score indicates the best performing configuration in each test. For example, for the experiment reported in Table 6.9, result obtained with 500 trees is considered the best as it has achieved the lowest rank sum for both the mean and combined mean and standard deviation (reported in brackets).

6.3 Results for segmentation without pre-classification

The evaluation measures used for gland segmentation and the ranking strategy for these results were detailed in Chapter 6, Sections 6.1 and 6.2. In this section, the results of the segmentation method without pre-classification with different features are presented. Different feature extraction methods are used to extract local patterns in selected patches, and random forest was used as the classifier.

6.3.1 Segmentation results for histogram features

This section describes the evaluation measures for the histogram features in segmentation without pre-classification. The details of the ring histogram feature used in gland segmentation are discussed in Chapter 4, Section 4.3. Sliding window techniques are used to extract the local features. The values of the design parameters (size of input patch, number of rings per patch and number of trees in the forest) affected the segmentation results. The experiments described in this section are designed to find the significance of different design parameters on the segmentation performance.

The first experiment aims to test the significance of a number of trees in the forest.

To compare the effect of the number of trees, the values of other parameters are fixed.

For the reported results, the 19-by-19 pixel input patches, 8 different rings per patch and 85,000 training patches are used. The reason for choosing these values for this experiment, are discussed below. Table 6.9 indicates the results from using different numbers of trees in the random forest model; 100, 300 and 500 trees were used to produce separate probability maps, and the final results were determined by using a set of morphological post-processing operations (discussed in Chapter 5, Section 5.6.1).

Table 6.9 Comparison of results when using different numbers of trees in random forest

Number of trees	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	
100	0.50 (0.24)	2 (1)	0.44±0.24	3 (2)	0.54 (0.21)	3 (3)	0.58 (0.19)	3 (3)	261.60 (179.21)	3 (3)	244.31 (136.45)	1 (3)	15 (30)
300	0.48 (0.24)	3 (1)	0.59±0.26	1 (3)	0.57 (0.19)	2 (2)	0.61 (0.16)	2 (1)	221.15 (170.40)	2 (2)	272.17 (120.77)	3 (1)	13 (23)
500	0.51 (0.26)	1 (3)	0.57±0.23	2 (1)	0.58 (0.18)	1 (1)	0.63 (0.16)	1 (1)	185.69 (157.00)	1 (1)	263.31 (125.57)	2 (2)	8 (17)

The best performance, highlighted in blue, was provided by the random forest with 500 decision trees. From this experiment, the conclusion is drawn that increasing the number of trees could help to improve the segmentation performance. Although the performance of random forest is likely to continue to improve as the number of decision trees increases, that improvement is bounded. Based on the experimental evidences shown in (Probst and Boulesteix, 2017; Latinne et al., 2001), the performance of the random forest will stop improving at a certain point, even if the number of decision trees in the forest model will continue to increase. For a forest with 2000 decision trees, 29.2 GB RAM is needed for training, whereas for the forest with 500 trees only 15.8 GB RAM is needed for training. The increasing size of the forest model leads to higher storage and computational demands, therefore as a compromise it has been decided to use 500 forest trees for all subsequent experiments. All these experiments run on a computer with Intel Core™ i7-4720HQ CPU and 16 GB RAM using MATLAB 2015b.

The next experiment is designed to estimate the significance of the size of the input patch. A small input patch contains limited information for the classifier to identify the differences between classes. Any input size that is too big will also confuse the classifier in recognising different regions in images. As it is impractical to test all possible sizes of the input patch in gland segmentation, requiring a large number of experiment and resulting in only small changes in segmentation performance, four different sizes (15x15,

19x19, 27x27 and 35x35) have been selected for test. The other parameters used to extract the ring histogram are fixed.

Figure 6.10 shows the segmentation results using different sizes of input patch; again, these results were generated using morphological post-processing; different colours indicate different gland objects.

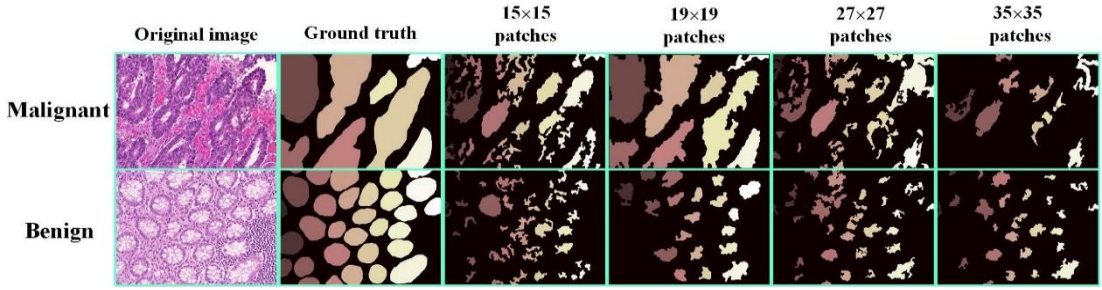


Figure 6.10 Comparison of the results of ring histograms with different sizes of input patches in segmentation without pre-classification. Different colours indicate different gland objects.

Table 6.10 shows the evaluation measures for the corresponding results, with the best segmentation performance highlighted in blue.

Table 6.10 Comparison results when using different patch size

Input Patch Size	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	
15 x 15	0.45 (0.23)	2 (2)	0.40±0.21	3 (2)	0.50 (0.20)	2 (3)	0.44±0.21	3 (4)	227.93 (143.01)	3 (2)	295.02 (158.13)	4 (3)	17 (33)
19 x 19	0.51 (0.26)	1 (4)	0.57±0.23	1 (4)	0.58 (0.18)	1 (2)	0.63±0.16	1 (1)	185.69 (157.00)	2 (3)	263.31 (125.57)	2 (1)	8 (23)
27 x 27	0.43 (0.25)	3 (3)	0.52±0.21	2 (2)	0.48 (0.22)	3 (4)	0.61±0.18	2 (2)	266.88 (169.87)	4 (4)	243.20 (130.33)	1 (2)	15 (32)
35 x 35	0.22 (0.13)	4 (1)	0.19±0.15	4 (1)	0.45 (0.17)	4 (1)	0.40±0.19	4 (3)	178.98 (95.80)	1 (1)	289.51 (165.14)	3 (4)	20 (31)

Using the evaluation measures to rank segmentaion performance, 19-by-19 was found to give the best segmentation results, therefore 19-by-19 patches have been selected for subsequent experiments.

Based on these two experiments, 500 trees and 19-by-19 patches are selected, this is considered a good choice for the available hardware configuration (see the description on page 128). The following experiment is designed to find the best value of number of rings per patch. As before, the values of the other design parameters are fixed. Table 6.11 shows the comparative results from using a different number of rings in each selected patch.

Table 6.11 Comparison results when using different number of rings

Different Number of Ring per Patch	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	
7	0.45 (0.23)	3 (1)	0.53 (0.23)	3 (1)	0.54±0.16	3 (1)	0.61±0.20	2 (2)	175.62±134.35	3 (1)	277.51±145.82	3 (2)	17 (25)
8	0.51 (0.26)	1 (2)	0.57 (0.23)	2 (1)	0.58±0.18	1 (2)	0.63±0.16	1 (1)	185.69±157.00	2 (3)	263.31±125.57	1 (1)	8 (18)
9	0.50 (0.26)	2 (2)	0.61 (0.24	1 (3)	0.58±0.18	1 (2)	0.55±0.22	2 (3)	185.13±156.48	1 (2)	270.07±164.05	2 (3)	9 (24)

Eight rings per patch deliver the best results. After comparing all the ring histogram features design parameters, the following experiment is designed to test the significance of post-processing method. Table 6.12 indicates the results of processing the probability maps generated from 19-by-19 patches with eight different rings.

Table 6.12 Comparison results when using different post-processing methods

Different Post-processing	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	
Level set	0.47 (0.23)	2 (1)	0.51 (0.23)	1 (1)	0.53 (0.18)	2 (1)	0.62 (0.20)	2 (2)	225.63 (150.63)	2 (1)	272.78 (145.82)	2 (2)	12 (19)
Morphological	0.51 (0.26)	1 (2)	0.57 (0.23)	1 (1)	0.58 (0.18)	1 (1)	0.63 (0.16)	1 (1)	185.69 (157.00)	1 (2)	263.31 (125.57)	1 (1)	6 (14)

Figure 6.11 is an example of segmentation results using the post-processing methods whose evaluation is shown in Table 6.12.

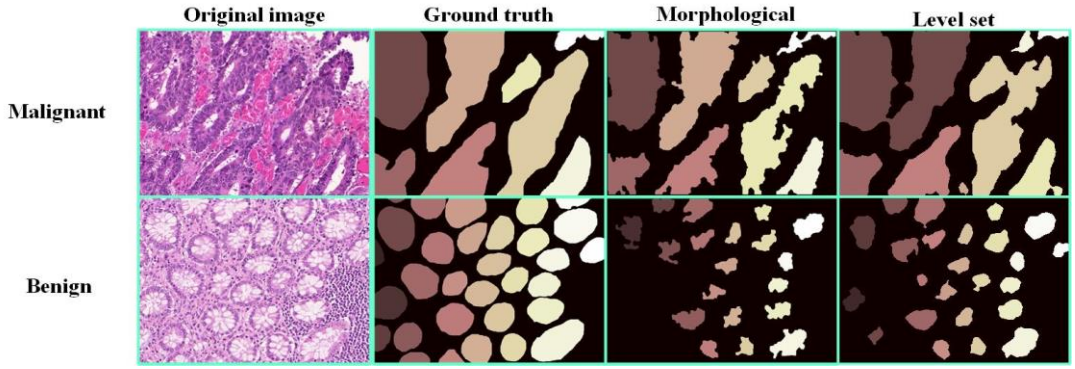


Figure 6.11 Comparison of results of different post-processing method

The segmentation results using level set post-processing method are worse than those using morphological post-processing. For the benign case segmentation results, it seems that the level set algorithm provided smooth and accurate results, but for the malignant result, although the results are smooth, a bridge connecting two close glands seriously affects the Hausdorff distance. Chapter 2 indicated that there are more images with malignant tissue (43 testing images) than images with benign tissue (37 testing images) in the whole dataset. The level set algorithm provides poorer object-level Hausdorff distance performance in the images with malignant tissue, which strongly affected the overall results of this method. For this reason, the following experiments use the morphological post-processing method rather than the level set algorithm in order to achieve better numerical scores.

6.3.2 Segmentation results with deep learning features

The LeNet-5 architecture (introduced in Chapter 4, Section 4.6) is used as a feature extraction method adapted for gland segmentation. The following experiments are designed to find values of the design parameters of the LeNet5 deep learning features. The hardware and software environments are the same as described in Chapter 6 Section 6.3.1. The forest with 500 trees is used for the pixel-level classification.

Table 6.13 shows the results of using different sized input patches on deep features. A sample of the segmentation results is shown in Figure 6.12. The number of input patches to train LeNet5 should not be too small, as this could cause overfitting, nor can it be too big due to the limitations of the hardware (see Chapter 6, Section 6.3.1). Although the number of chosen patches per image could be different, in this experiment 1,000 patches per image have been selected (there are 85 images in training dataset, 85,000 patches in total) to describe the local patterns.

Table 6.13 The results of deep learning features using LeNet-5 with different size of input patches

Different Input Size	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	rank	
15 X 15	0.42 (0.21)	2 (1)	0.46 (0.30)	2 (3)	0.53 (0.19)	2 (2)	0.47 (0.19)	3 (3)	191.87 (146.81)	1 (1)	399.67 (144.33)	3 (2)	13 (25)
19 X 19	0.48 (0.25)	1 (3)	0.52 (0.20)	1 (1)	0.56 (0.18)	1 (1)	0.51 (0.18)	2 (2)	210.75 (158.51)	3 (3)	319.64 (147.83)	2 (3)	10 (23)
27 X 27	0.32 (0.21)	3 (1)	0.43 (0.21)	3 (2)	0.48 (0.20)	3 (3)	0.58 (0.17)	1 (1)	210.44 (153.76)	2 (2)	278.23 (141.41)	1 (1)	13 (24)

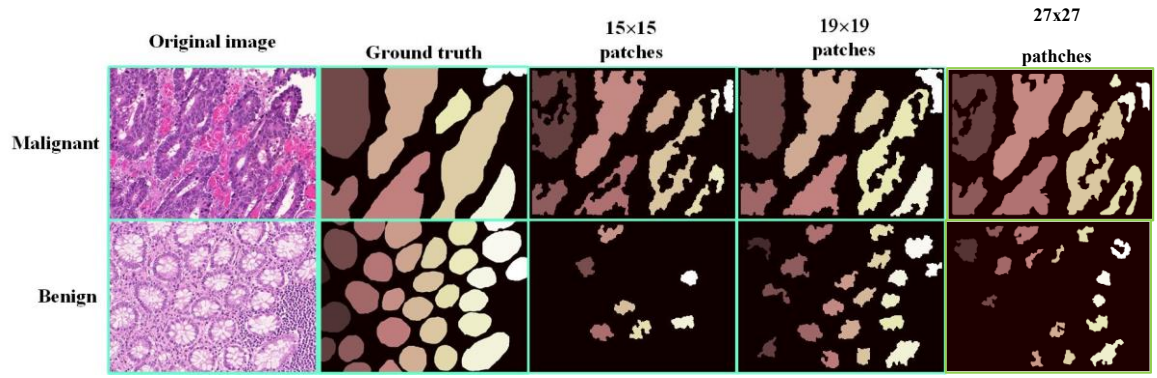


Figure 6.12 Segmentation results of deep learning using LeNet-5 with different size of input patches

From both the qualitative visual inspection and the quantitative results, segmentation using 19-by-19 patches detected more gland objects, giving the most accurate results.

The last experiment is designed to find the significance of the number of input patches, with other parameters being fixed. Table 6.14 shows the results of deep

learning features with different numbers of training input patch, and a sample of segmentation results is shown in Figure 6.13.

Table 6.14 Results of deep feature from LeNet-5 with different number of input patches

Number of input patches	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	rank	
85000	0.48 (0.25)	2 (1)	0.52 (0.21)	2 (1)	0.56 (0.19)	2 (1)	0.51 (0.18)	2 (1)	210.75 (158.51)	1 (1)	319.64 (147.83)	2 (1)	11 (17)
170000	0.60 (0.25)	1 (1)	0.63 (0.22)	1 (2)	0.62 (0.20)	1 (2)	0.60 (0.20)	1 (2)	212.59 (163.29)	2 (2)	256.21 (149.51)	1 (2)	7 (17)

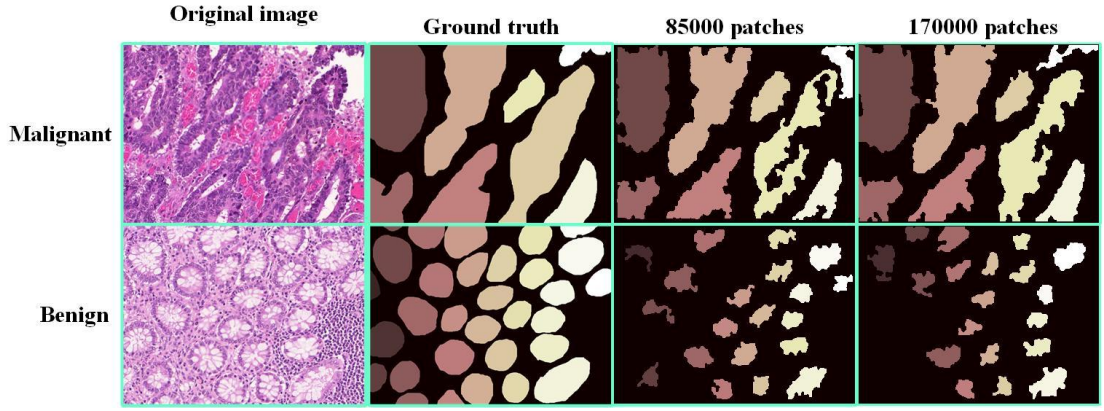


Figure 6.13 A sample of segmentation results with LeNet-5 features as function of different number of training input patches.

This experiment shows that increasing the number of patches does not significantly improve the performance, indeed it does not improve if the standard deviation is also taken into account. If high performance alone is the priority, more patches (170,000 in this experiment) are used.

The above experiments show the effect of different parameters when LeNet5 features are used. GoogleNet architecture is also used as feature extractor for gland segmentation (the details of GoogleNet feature were discussed in Chapter 4, Section 4.6.2). The parameters used to generate GoogleNet features will affect the performance. As for LeNet-5 features, the patch size and the number of patches used for training of GoogleNet features are tested in the following experiments. In order to reduce overfitting, 1000 patches per image (85,000 in total) are used to train the GoogleNet. This value has been chosen based on the available hardware (see Chapter 6, Section 6.3.1). Table 6.15 reports the evaluation measures for the GoogleNet deep learning features learnt using 85,000 patches of different sizes, and Figure 6.14 shows a sample of the corresponding segmentation results.

Table 6.15 Evaluation measures for GoogleNet deep features with different sizes of input patch

Size of patches	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
49x49	0.62 (0.20)	3 (2)	0.59 (0.25)	3 (1)	0.65 (0.16)	3 (2)	0.65 (0.21)	3 (1)	184.95 (128.21)	3 (3)	254.45 (160.18)	3 (1)	18 (30)
97x97	0.66 (0.20)	2 (2)	0.62 (0.26)	2 (3)	0.69 (0.16)	2 (2)	0.69 (0.21)	1 (1)	169.98 (109.20)	2 (2)	239.69 (173.37)	1 (2)	10 (22)
225x225	0.69 (0.18)	1 (1)	0.66 (0.25)	1 (1)	0.74 (0.15)	1 (1)	0.68 (0.22)	2 (2)	121.10 (84.84)	1 (1)	240.47 (182.16)	2 (3)	8 (17)

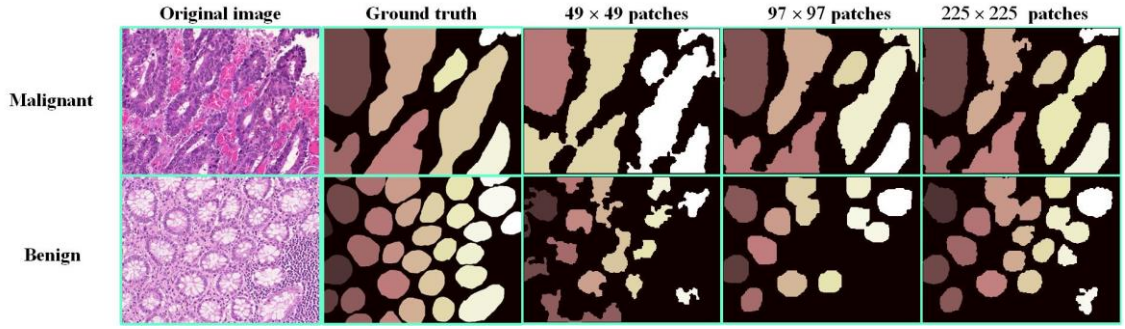


Figure 6.14 A sample of segmentation results with GoogleNet deep features using different sizes of input patch

Based on visual inspection and quantitative results, using 225-by-225 patches detected more gland objects than the other two values. From the quality measure, it also shows the best results are using 225-by-225 patches among these three results. From this experiment, 225-by-225 patches are the one close to the optimal size for GoogleNet feature based on the hardware.

After comparing the significance of the size of input patches, the following experiment tests the significance of the number of input patches for gland segmentation. Again, 1000 and 2000 patches per image are selected for tests. Table 6.16 shows the results for GoogleNet deep features with different numbers of training input patches, and a sample of corresponding segmentation results is shown in Figure 6.15.

Table 6.16 Evaluation measures for the GoogleNet features with different numbers of training patches

Number of input patches	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	rank	
85000	0.69 (0.18)	2 (1)	0.66 (0.25)	2 (1)	0.74 (0.14)	2 (1)	0.68 (0.22)	1 (1)	121.10 (84.84)	2 (2)	240.47 (182.16)	2 (2)	11 (19)
170000	0.70 (0.18)	1 (1)	0.69 (0.25)	2 (1)	0.75 (0.14)	1 (1)	0.68 (0.22)	1 (1)	105.14 (76.76)	1 (1)	224.00 (177.18)	1 (1)	7 (13)

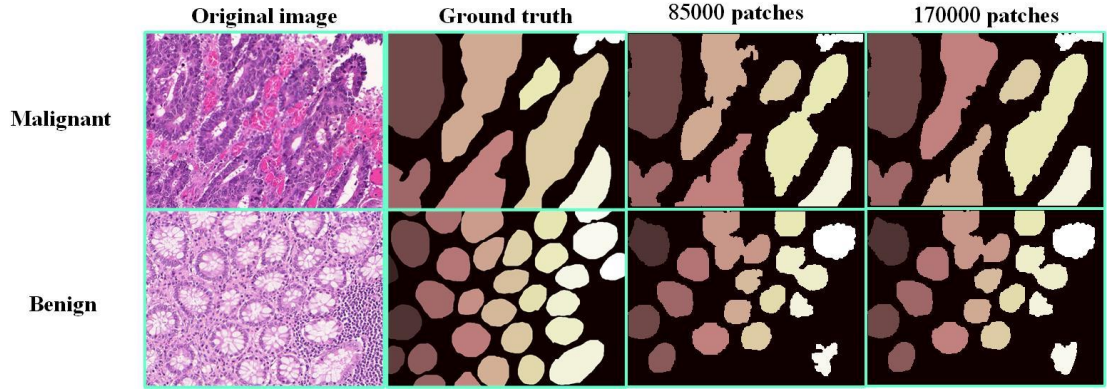


Figure 6.15 A sample of segmentation results with the GoogleNet deep features using different numbers of training patches

From the results shown in Figure 6.15, it can be concluded that using large number of training patches improves the segmentation results. This observation is confirmed by the quantitative evaluation shown in Table 6.16. This experiment has shown that the more input patches used for training the GoogleNet helps the network learn more discriminative patterns. However, in practice the number of patches is limited by the available hardware (see Chapter 6, Section 6.3.1).

6.3.3 Summary of segmentation without pre-classification

The above sections discuss the segmentation results with three most effective features used in segmentation without pre-classification. The results with other features (the details were discussed in Chapter 3) are shown in Appendix D. This section summarises of the best results of the different features in segmentation without pre-classification.

The reasons for choosing the values of the design parameters for each feature were discussed in the corresponding sections. For example, the reasons for choosing specific values of the design parameters for the ring histograms are discussed in Chapter 6 Section 6.3.1. All design parameters for that feature are discussed in Chapter 3.

Table 6.17 summarises the best segmentation results with different features using segmentation without pre-classification. Chapter 2 Section 2.2 provided details of the database used in this work. *RIULBP* in Table 6.17 denotes the rotation-invariant uniform LBP. *CHOG* is the circular Fourier HOG feature. *RIULBP&ring histogram* indicates the combination of rotation-invariant uniform LBP and ring histogram. LeNet5 & Ring histogram represents the hybrid features which combines the LeNet5 feature and ring

histogram.

Table 6.17 The overall performance of different features in segmentation without pre-classification

Different features	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	rank	
GoogleNet	0.7 (0.19)	1 (2)	0.67 (0.28)	1 (7)	0.74 (0.14)	1 (1)	0.66 (0.21)	1 (6)	105.14 (76.77)	1 (1)	236.29 (170.20)	1 (6)	6 (29)
LeNet5	0.6 (0.25)	2 (6)	0.63 (0.22)	3 (2)	0.62 (0.20)	2 (7)	0.6 (0.20)	4 (5)	212.59 (163.29)	6 (7)	256.21 (149.51)	2 (4)	19 (50)
Ring histogram	0.51 (0.26)	5 (7)	0.57 (0.23)	4 (3)	0.58 (0.18)	4 (4)	0.63 (0.17)	2 (1)	185.69 (156.99)	2 (6)	263.31 (125.57)	3 (1)	20 (42)
Ring histogram & RIULBP	0.55 (0.22)	3 (4)	0.67 (0.25)	2 (5)	0.59 (0.15)	3 (3)	0.59 (0.24)	5 (7)	195.78 (131.99)	4 (3)	301.8 (181.55)	6 (7)	23 (52)
RIULBP	0.4 (0.19)	6 (2)	0.54 (0.25)	5 (5)	0.5 (0.14)	6 (1)	0.63 (0.17)	2 (1)	198.05 (101.01)	5 (2)	264.79 (126.36)	4 (2)	28 (41)
LeNet5 & Ring histogram	0.52 (0.24)	4 (5)	0.52 (0.21)	6 (1)	0.57 (0.19)	5 (5)	0.51 (0.18)	7 (3)	195.53 (149.09)	3 (5)	319.63 (147.83)	7 (3)	32 (54)
CHOG	0.34 (0.17)	7 (1)	0.44 (0.24)	7 (4)	0.33 (0.19)	7 (5)	0.53 (0.18)	6 (3)	263.97 (134.98)	7 (4)	301.35 (151.87)	5 (5)	39 (61)

The best segmentation results, highlighted in blue, have been achieved using GoogleNet features. The results are sorted, from top to bottom, according the overall ranking (using rank sum) of the segmentation results. In general, deep features (GoogleNet and LeNet5) perform better than the hand-crafted features (ring histogram, rotation-invariant uniform LBP and circular Fourier HOG). For the same pixel-level classifier. As demonstrated in this section the performance of segmentation without pre-classification method using different features depends significantly on the selection of the correct values of the corresponding design parameters. As the deep features are optimised for specific segmentation problem (i.e. training data), the fact that they provide better segmentation results than the hand-crafted features has been somewhat expected.

The results reported in Section E.6 (in Appendix E) demonstrate that the intensity-based features (ring histogram and rotation-invariant uniform LBP) perform better than the gradient-based features (circular Fourier HOG) on the investigated gland segmentation problem.

Images with the benign and malignant tissue look very dissimilar; therefore, it is interesting to investigate the difference between performances of tested features on corresponding data subsets. Tables 6.18 and 6.19 show segmentation results with tested features for benign and malignant only cases respectively. One of the reasons to investigate such configurations is that it can give an insight into, and indeed motivation for, developing hybrid segmentation methods, i.e. methods using image pre-classification.

Table 6.18 Performance of different features for the benign category only, using segmentation without pre-classification

Different features	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	rank	
GoogleNet	0.71 (0.16)	1 (1)	0.84 (0.14)	2 (2)	0.75 (0.14)	1 (2)	0.88 (0.11)	1 (2)	106.34 (83.37)	1 (2)	77.04 (79.65)	1 (2)	7 (18)
Ring histogram & RIULBP	0.55 (0.20)	3 (4)	0.87 (0.04)	1 (1)	0.6 (0.14)	3 (2)	0.87 (0.16)	2 (4)	170.72 (121.48)	4 (3)	107.3 (71.25)	2 (1)	16 (30)
LeNet5	0.58±0.22	2 (5)	0.77 (0.25)	3 (6)	0.61 (0.20)	2 (7)	0.71 (0.21)	6 (7)	207.63 (162.52)	6 (7)	154.15 (109.49)	5 (5)	24 (61)
LeNet5 & Ring histogram	0.51 (0.23)	4 (6)	0.73 (0.14)	4 (2)	0.56 (0.17)	4 (4)	0.72 (0.17)	5 (5)	170.34 (133.73)	3 (4)	170.15 (120.28)	6 (6)	26 (53)
Ring histogram	0.45±0.25	5 (7)	0.72 (0.23)	5 (5)	0.55 (0.18)	5 (5)	0.82 (0.1)	3 (1)	173.68 (148.03)	5 (5)	131.23 (85.63)	3 (3)	26 (55)
RIULBP	0.41 (0.19)	6 (3)	0.64 (0.33)	6 (7)	0.5 (0.09)	6 (1)	0.8 (0.12)	4 (3)	160.8 (71.25)	2 (1)	132.5 (86.40)	4 (4)	28 (47)
CHOG	0.37±0.17	7 (2)	0.53 (0.17)	7 (4)	0.41 (0.18)	7 (5)	0.66 (0.20)	7 (6)	314.95 (158.32)	7 (6)	212.63 (136.02)	7 (7)	42 (72)

Table 6.19 Performance of different features for the malignant category only, using segmentation without pre-classification

Different features	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	rank	
GoogleNet	0.69 (0.22)	1 (3)	0.60 (0.29)	1 (7)	0.76 (0.14)	1 (1)	0.60 (0.19)	1 (5)	103.67 (69.36)	1 (2)	276.10 (164.23)	1 (6)	6 (30)
Ring histogram	0.58 (0.25)	3 (5)	0.52 (0.21)	4 (2)	0.62 (0.19)	3 (5)	0.59 (0.15)	2 (1)	200.38 (168.98)	2 (7)	296.33 (112.65)	3 (1)	17 (38)
LeNet5	0.63 (0.28)	2 (7)	0.59 (0.21)	3 (2)	0.65 (0.21)	2 (6)	0.57 (0.20)	4 (6)	218.65 (167.12)	4 (6)	281.73 (149.82)	2 (4)	17 (48)
Ring histogram & RIULBP	0.54 (0.23)	5 (4)	0.60 (0.24)	1 (5)	0.58 (0.15)	5 (2)	0.52 (0.21)	5 (7)	226.42 (140.00)	5 (4)	350.42 (167.70)	7 (7)	29 (57)
RIULBP	0.38 (0.20)	6 (2)	0.51 (0.22)	5 (4)	0.4 (0.18)	6 (4)	0.58 (0.15)	3 (1)	243.57 (113.97)	7 (3)	297.86 (113.56)	4 (2)	31 (47)
LeNet5 & Ring histogram	0.55 (0.25)	4 (5)	0.44 (0.17)	6 (1)	0.59 (0.21)	4 (6)	0.46 (0.15)	7 (1)	226.31 (163.22)	6 (5)	357.01 (131.68)	6 (3)	33 (54)
CHOG	0.21 (0.11)	7 (1)	0.43 (0.26)	7 (6)	0.25 (0.16)	7 (3)	0.50 (0.17)	6 (4)	201.65 (56.28)	3 (1)	323.53 (151.29)	5 (5)	35 (55)

From these two tables (6.18 and 6.19), it can be seen that on average the segmentation performance on the benign category is better than on the malignant category. This is because the morphological structure of malignant tissue is more complex than that of benign tissue (see Chapter 2 Section 2.3). Therefore, the proposed methods are more effective in dealing with the morphological structure of benign tissue. It can be also noticed that performance on each category alone is better than the overall performance, when all images are processed in the same way. This indicates that the proposed hybrid methods (which combines two-level classification) should perform better than the method described in this section, i.e. segmentation without pre-classification. Interestingly, the second best performing feature include the ring-histograms, when benign and malignant glands are segmented using methods trained respectively on the benign or malignant training data alone. This shows that ring-histograms extract discriminative enough characteristics of the glands.

6.4 Results for segmentation with pre-classification method

The details of segmentation with pre-classification methods were described in Chapter 5, Section 5.4, where two segmentation methods with pre-classification were introduced. Both these methods have two-stage classification approach, using image and pixel level classifications. This section reports on the results for image-level and pixel-level classifications.

6.4.1 Results for image-level classification

Chapter 5 introduced two segmentation methods, with and without pre-classification. The segmentation with pre-classification is a two-level classification problem. Image-level classification is to assign an image to one of the classes representing benign or malignant tissue.

For the image-level classification, a fusion method using the HMAX model and random forest techniques as well as deep learning techniques are tested. Classification using the fusion method could only achieve 70% classification accuracy no matter how the parameters in the HMAX model were selected. The 70% accuracy for an image classification problem is not satisfactory.

Three deep learning algorithms have been tested for image-level classification in the segmentation with pre-classification: AlexNet, GoogleNet and ResNet50. Although many versions of ResNet architecture could be used for the image-level classification, ResNet-50 was chosen for practical reasons as a good compromise between expected performance and the available resources (GTX 1080 Nvidia Graphic card was used in the experiments).

As explained in Chapter 2, there are 85 histology training images in the gland database. They include 37 images with benign tissue and 48 images with malignant tissue. First, 80% of each category was used for training and the rest of the corresponding category for validation. To avoid the networks overfitting, local image deformation and colour jitter techniques were used for data augmentation, what increased the number of images in the training dataset to a total of 6,392 (3,572 images with malignant tissue generated from 38 base images, and 2,820 images with benign tissue derived from 30

base images). The remaining 17 images used for validation consisted of 7 images with benign tissue and 10 images with malignant tissue. The details of colour jitter and local image deformation are explained in Chapter 5 Section 5.4.1. The original 80 test images were only used for testing. The Adam optimising method was chosen for each model, and the initial learning rate for all the networks was set to 0.0001. The maximum number of epochs was set to 200, if the classification accuracy for validation data achieved 100%, the network would stop training. The model selected was based on the best segmentation accuracy obtained on validation data. The screenshots of classification results for the networks are shown in Appendix H.

Table 6.20 shows the image-level classification results for the deep learning models (AlexNet, GoogleNet, ResNet-50) on the test data. **TP** refers to detecting image showing benign tissue; **TN** refers to detecting image showing malignant tissue; **FN** represents images with benign category predicted as images showing malignant tissue, **FP** represents images depicting malignant tissue predicted as images showing benign tissue.

Table 6.20 Image-level classification results of deep learning models (AlexNet, GoogleNet, ResNet-50) on testing data

Model Name	Number of **TP**	Number of **FP**	Number of **FN**	Number of **TN**
GoogleNet	34	3	8	35
AlexNet	23	14	13	30
ResNet-50	37	0	0	43

Based on the results shown in Table 6.20, the best image-level classification is obtained using the ResNet-50 architecture. The reason for the ResNet-50 being the best might be that there are more layers in that network than in the other two networks; residual connections between layers is another powerful tool in ResNet-50 to improve the network training process and therefore the classification results. In practice, however, the best performance is not always provided by the network with more layers. The performance for a specific problem depends on many criteria, such as the data augmentation used for increasing the amount of training data. ResNet-50 achieves 100% classification accuracy on these test data; the performance of segmentation with pre-classification depend on the results of pixel-level classification.

After image-level classification, the histology gland images are divided into benign and malignant categories. Subsequently, the pixel-level classification used for

segmentation implements a sliding window to extract the local patterns from images with either benign or malignant tissue to make predictions at a pixel level. In the next sections, the segmentation results of benign and malignant categories are discussed.

6.4.2 Summary of pixel-level classification

This section summaries the segmentation results with pre-classification method at feature extraction level (see Chapter 5, Section 5.4.2.3), and the results for either benign or malignant categories are shown in Appendix E. The results for two target classes and benign cases are shown in Appendix E, Section E.1, and the results for three target classes and benign cases are shown in Appendix E, Section E.2. The results for two target classes and malignant cases are shown in Appendix E, Section E.3, and finally the results for three target and malignant cases are shown in Appendix E, Section E.4.

The details for generating the three target classes' ground truth for images with benign and malignant tissue are explained in Chapter 5, Section 5.4.2.2. The parameters used for each feature are fixed, based on previous experiments, with the same hardware and the software environment (see Chapter 6, Section 6.3.1) as the other experiments. Again, the pixel-level classifier is the random forest model with 500 trees and the final segmentation results are generated by a set of morphological post-processing operations. However, unlike in the previous experiments, the features for pixel level-classification are only extracted from images representing a single category (i.e. benign or malignant). The experiments are designed in order to find the best way to describe the morphological structure in images with benign tissue and those with malignant tissue.

Table 6.21 shows the results obtained for different features using different number of target classes on the benign image category. Taking the LeNet5 feature in benign cases as an example, the performance of the three target classes is better than those of two target classes. This is because the morphological structure of benign gland tissue contains three main target structures (inside glands (cytoplasm), gland boundary

(epithelial cells) and gland outside (stroma)). This can be confirmed by visual inspection of Figure 2.2 shown in Chapter 2.

Table 6.21 Overall ranking of segmentation results for different features on images with benign tissue

Different Features	Number of Target classes	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
		Test A		Test B		Test A		Test B		Test A		Test B		
		Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	
LeNet5	3	0.73 (0.10)	2 (1)	0.73 (0.08)	3 (4)	0.74 (0.10)	1 (3)	0.83 (0.10)	2 (7)	102.95 (87.95)	2 (10)	144.68 (50.19)	7 (8)	17 (50)
GoogleNet	2	0.64 (0.20)	4 (11)	0.82 (0.11)	1 (6)	0.68 (0.15)	3 (12)	0.87 (0.08)	1 (3)	129.04 (73.06)	8 (5)	94.39 (69.07)	1 (9)	18 (65)
Ring histogram	3	0.74 (0.11)	1 (2)	0.59 (0.18)	7 (11)	0.74 (0.08)	1 (1)	0.72 (0.09)	7 (5)	89.42 (72.23)	1 (4)	138.38 (47.75)	6 (6)	23 (52)
LeNet5	2	0.52 (0.15)	5 (6)	0.69 (0.11)	5 (6)	0.64 (0.12)	7 (7)	0.79 (0.08)	4 (3)	147.54 (85.95)	10 (9)	120.26 (49.19)	2 (7)	33 (71)
Ring histogram	2	0.56 (0.20)	8 (11)	0.78 (0.02)	2 (2)	0.61 (0.15)	9 (12)	0.81 (0.12)	3 (8)	161.50 (122.87)	11(12)	136.89 (91.81)	5(11)	36 (94)
GoogleNet	3	0.55 (0.17)	9 (8)	0.62 (0.08)	6 (4)	0.6 (0.13)	10 (9)	0.73 (0.06)	6 (1)	113.49 (70.01)	6 (2)	133.14 (27.27)	3 (2)	40 (65)
RIULBP	3	0.6 (0.19)	7 (10)	0.37 (0.17)	10 (10)	0.65 (0.12)	6 (7)	0.6 (0.14)	8 (10)	111.38 (76.26)	4 (7)	177.73 (33.52)	8 (3)	43 (90)
Ring histogram & RIULBP	2	0.67 (0.14)	3 (3)	0.51 (0.18)	8 (11)	0.68 (0.09)	3 (2)	0.57 (0.17)	10 (12)	123.50 (82.80)	7 (8)	328.44 (152.57)	12 (14)	43 (93)
RIULBP	2	0.43 (0.17)	11 (8)	0.72 (0.29)	4 (14)	0.56 (0.10)	11 (3)	0.78 (0.09)	5(5)	140.05 (64.08)	9 (2)	133.89 (33.86)	4 (4)	44 (78)
LeNet5 & Ring histogram	2	0.62 (0.16)	5 (7)	0.27 (0.15)	14 (9)	0.66 (0.11)	5 (5)	0.48 (0.06)	12 (1)	113.45 (73.81)	5 (6)	292.56 (109.38)	11 (12)	52 (92)
Ring histogram & RIULBP	3	0.55 (0.2)	9 (11)	0.29 (0.01)	13 (1)	0.62 (0.11)	8 (5)	0.52 (0.12)	11 (8)	108.73 (61.33)	3 (1)	190.54 (19.96)	9 (1)	53 (80)
LeNet5 & Ring histogram	3	0.38 (0.22)	12(14)	0.43(0.22)	9 (13)	0.42(0.20)	12(14)	0.58 (0.29)	9 (14)	240.7 (158.09)	12(14)	222.29(113.81)	10(13)	64(146)
CHOG	3	0.3 (0.15)	14 (4)	0.33(0.06)	11 (3)	0.35 (0.13)	13 (9)	0.35 (0.16)	13 (11)	289.59 (137.06)	13(13)	376.17 (41.42)	13(5)	77(122)
CHOG	2	0.32 (0.15)	13 (4)	0.30(0.14)	12 (8)	0.35 (0.13)	13(9)	0.34 (0.19)	14 (13)	363.86 (114.84)	14(11)	395.34 (71.50)	14(10)	80(135)

Table 6.22 shows the results for different features with different number of target classes for images with malignant cases. Contrary to the benign cases, for the malignant tissue it is better to use the two target classes to describe the morphological structure. The reason for this is that the morphological structure of glands in malignant cases lacks the clear 'inside' pattern. This could be confirmed when comparing the benign and malignant tissue in images in Chapter 2.

Table 6.22 Overall ranking of segmentation results for different features on images with malignant tissue

Different Features	Number of Target classes	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
		Test A		Test B		Test A		Test B		Test A		Test B		
		Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	
Ring histogram	2	0.65 (0.27)	1 (12)	0.57 (0.25)	2 (7)	0.7 (0.19)	1 (7)	0.54 (0.2)	1 (9)	163.84 (155.22)	2 (9)	308.86 (147.3)	4 (6)	11 (61)
GoogleNet	2	0.62 (0.28)	2 (13)	0.5 (0.19)	5 (2)	0.67 (0.19)	2 (7)	0.54 (0.18)	1 (4)	162.66 (102.96)	1 (3)	299.8 (164.07)	2 (10)	13 (52)
LeNet5	2	0.61 (0.29)	3 (14)	0.52 (0.21)	3 (4)	0.66 (0.21)	3 (13)	0.53 (0.15)	3 (2)	195.23 (176.8)	3 (14)	326.55 (129.59)	9 (3)	24 (74)
LeNet5	3	0.57 (0.26)	4 (11)	0.58 (0.19)	1 (2)	0.58 (0.19)	6 (7)	0.52 (0.2)	4 (9)	215.69 (166.22)	5 (11)	319.33 (133.77)	6 (4)	26 (70)
Ring histogram	3	0.54 (0.23)	5 (8)	0.44(0.26)	11(8)	0.59 (0.17)	5(5)	0.45 (0.23)	10 (13)	204.39 (147.75)	4 (7)	263.1 (171,28)	1 (12)	36 (89)
RIULBP	2	0.42 (0.18)	7 (2)	0.34 (0.22)	13 (6)	0.52 (0.14)	7 (3)	0.44 (0.21)	11(11)	226.77 (115.71)	6 (4)	309.37 (171.41)	5(14)	49 (89)
LeNet5 & Ring histogram	2	0.52 (0.23)	6 (8)	0.43 (0.27)	12 (9)	0.6 (0.21)	4 (13)	0.46 (0.16)	9 (3)	249.85 (174.04)	7 (13)	371.86 (155.86)	12 (7)	50(103)
CHOG	3	0.39 (0.23)	8 (8)	0.49 (0.33)	6 (12)	0.39 (0.19)	10 (7)	0.47 (0.18)	8 (4)	393.65 (164.77)	12(10)	371.86 (163.17)	13 (9)	50(106)
CHOG	2	0.39 (0.21)	8 (6)	0.49 (0.32)	6 (10)	0.44 (0.18)	9 (6)	0.48 (0.18)	6 (4)	322.25 (150.5)	10 (8)	377.7 (170.21)	14 (13)	53(100)
RIULBP	3	0.39 (0.18)	10 (2)	0.46 (0.32)	9 (10)	0.45 (0.19)	8 (7)	0.42 (0.19)	12 (7)	263.21 (146.94)	9 (6)	321.96 (123.49)	8 (2)	56 (90)
Ring histogram & RIULBP	2	0.31 (0.18)	12 (2)	0.46 (0.21)	9 (4)	0.27 (0.1)	14 (2)	0.48 (0.22)	6 (12)	522.69 (85.57)	14 (2)	305.16 (165.2)	3 (11)	58 (91)
Ring histogram & RIULBP	3	0.38 (0.2)	11 (7)	0.48 (0.33)	8 (12)	0.33 (0.19)	12 (7)	0.49 (0.19)	5 (7)	362.51 (167.02)	11(12)	350.83(158.05)	11 (8)	58(111)
LeNet5 & Ring histogram	3	0.31 (0.18)	12 (2)	0.51 (0.41)	4 (14)	0.26 (0.09)	13 (1)	0.38 (0.23)	13 (13)	510.85 (81.63)	13 (1)	321.46 (107.66)	7 (1)	62 (94)
GoogleNet	3	0.28 (0.12)	14 (1)	0.33 (0.12)	14 (1)	0.38 (0.14)	11 (3)	0.3 (0.14)	14 (1)	257.89 (141.72)	8 (5)	344.36 (141.47)	10 (5)	71 (87)

Based on the results shown in this section, the conclusion from these experiments is that the best way to describe benign tissue is to use the three target classes, but for malignant cases it is two target classes.

6.5 Comparison of the two segmentation methods

Two main approaches have been proposed in this work for gland segmentation problem: segmentation with and without pre-classification. The proposed segmentation with the pre-classification has two variants: pre-classification, at the feature extraction level and at the pixel-level classifier (see Chapter 5, section 5.4.2.3). In this section the performance of the segmentation without pre-classification (so called Method 1) and the segmentation with pre-classification at the feature extraction level (so called Method 2) are compared. As reported in Chapter 6, Section 6.3, the best segmentation results for segmentation without pre-classification have been obtained using the GoogleNet features. The best results in segmentation with pre-classification (at the feature extraction level) are obtained by combining the best results from benign and malignant categories, as reported in Chapter 6, Section 6.4. Table 6.23 shows the ranking of these two approaches using three evaluation measures frequently used for gland segmentation evaluation.

Table 6.23 Performance of the best gland segmentation methods with and without pre-classification

Using pre-classification	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	Score	rank	
Yes	0.72 (0.18)	1 (1)	0.61 (0.20)	2 (1)	0.73 (0.14)	2 (1)	0.59 (0.23)	2 (2)	126.22 (124.15)	2 (2)	275.61 (156.21)	2 (1)	11 (19)
No	0.7 (0.19)	2 (2)	0.67 (0.28)	1 (2)	0.74 (0.14)	1 (1)	0.66 (0.21)	1 (1)	105.14 (76.77)	1 (1)	236.26 (170.20)	1 (2)	7 (17)

Using F1 score, object-level Dice index and object-level Hausdorff distance as evaluation measures, the best results of segmentation with pre-classification is not as good as the best results of segmentation without pre-classification. The segmentation with pre-classification has not improved the segmentation performance when compared to segmentation without pre-classification. This somewhat surprising result could be explained by the fact that the deep feature extraction methods are using smaller number of patches for training and therefore could not find strong discriminative features.

As discussed in Section 6.1.3, the object-level Hausdorff distance is not suitable for evaluation of the segmentation performance in gland segmentation. The object-level

Boundary Jaccard index is considered better suited for that purpose as it is not sensitive to outliers and is bounded with values between 0 and 1. Table 6.24 shows the ranking of the same two proposed segmentation methods using the object-level Boundary Jaccard index.

Table 6.24 Performance of the best gland segmentation methods with and without pre-classification based on the object-level Boundary Jaccard index

Segmentation using pre-classification	**Boundary Jaccard index**				Rank sum
	Test A		Test B		
	Value	Rank	Value	Rank	
No	0.769 (0.140)	1 (1)	0.663 (0.138)	1 (1)	2 (4)
Yes	0.744 (0.149)	2 (2)	0.594 (0.231)	2 (2)	4 (8)

From the results shown in Table 6.26 the result of segmentation without pre-classification is better than the result of segmentation with pre-classification using object-level Boundary Jaccard index as the evaluation measure. It is suggested that it is better to use object-level boundary Jaccard index as the shape similarity evaluation measure. The reason is that frequently used object-level Hausdorff distance is sensitive to outliers and is not bounded, what makes it difficult to combine it with existing other metrics such as Dice index.

6.6 Comparison of three segmentation methods

As reported in the previous section (Chapter 6, Section 6.5) the segmentation with pre-classification did not improve the segmentation results as expected. It is hypothesized that one of the reasons might be that the number of input samples used to learn the deep features when using pre-classification is smaller than when using segmentation without pre-classification method, and therefore the learned deep features are suboptimal. This was one of the reasons to propose a different version of the segmentation with pre-classification, i.e. segmentation with the pre-classification at the pixel level classifier (the so called “Method 3” - see Chapter 5, Section 5.4.2.3). In this case, all the training images, with both benign and malignant tissue are used for learning deep features. In the subsequent step after images are directed into separate respective decisions paths for the images showing benign and malignant tissue, the dedicated pixel-level classifiers are trained independently but with the same set of

training features.

The GoogleNet architecture (shown in Figure 4.16) has been selected and used for learning local features. One thousand 224-by-224 patches are sampled from each training histology image in the gland data for training. The reasons for choosing these parameters are discussed in Chapter 6 Section 6.3.2. The results reported in Chapter 6 Section 6.3.3 describe performance of the segmentation without pre-classification (Method 1), whereas the results reported in Chapter 6 Section 6.4 represent performance of the segmentation with pre-classification at the feature extraction level (Method 2). Segmentation with pre-classification at the pixel-level classifier (Method 3) uses two random forest models trained separately for benign and malignant cases but with the same set of features.

For the results reported in this section, the subsets Part A and Part B, previously used for evaluation, have been combined and the results are reported for the whole test dataset. The reason for combining these two sets and reporting the results for the whole dataset is to focus on the overall performance of the method, rather than on somewhat arbitrary selected two subsets (Part A and Part B).

Table 6.25 The overall segmentation results for the three methods shown in Figure 5.13

Different Method	F1 Score		Object-level Dice index		Object-level Hausdorff distance		Rank sum
	Value	Rank	Value	Rank	Value	Rank	
Method 3	0.67 (0.21)	1 (1)	0.73 (0.17)	1 (1)	136.48 (120.86)	1 (1)	3 (9)
Method 1	0.67 (0.22)	1 (2)	0.71 (0.18)	2 (2)	150.38 (123.54)	2 (2)	5 (11)
Method 2	0.63 (0.22)	3 (2)	0.66 (0.18)	3 (2)	173.75 (123.54)	3 (2)	9 (15)

Tables 6.26 and 6.27 show the results for malignant and benign categories using three segmentation methods introduced in Section 5.5.

Table 6.26 Malignant category segmentation results for three methods shown in Figure 5.13

Different Method	F1 Score		Object-level Dice index		Object-level Hausdorff distance		Rank sum
	Value	Rank	Value	Rank	Value	Rank	
Method 3	0.61 (0.23)	2 (1)	0.70 (0.18)	1 (1)	164.69 (141.27)	1 (1)	3 (7)
Method 1	0.64 (0.23)	1 (1)	0.67 (0.19)	2 (2)	210.65 (147.62)	2 (3)	5 (11)
Method 2	0.60 (0.24)	3 (3)	0.62 (0.19)	3 (2)	215.64 (142.64)	3 (2)	9 (16)

Table 6.27 Benign category segmentation results for three methods shown in Figure 5.13

Different Method	F1 Score		Object-level Dice index		Object-level Hausdorff distance		Rank sum
	Value	Rank	Value	Rank	Value	Rank	
Method 3	0.73 (0.16)	1 (1)	0.76 (0.14)	1 (1)	103.70 (81.94)	1 (2)	3 (7)
Method 1	0.71 (0.16)	2 (1)	0.74 (0.15)	2 (2)	117.68 (90.92)	3 (3)	7 (13)
Method 2	0.66 (0.20)	3 (3)	0.70 (0.15)	3 (2)	125.29 (72.54)	2 (1)	8 (14)

Segmentation with pre-classification at feature extraction level (Method 2 in Table 6.25) is the worst performing method. Segmentation with pre-classification at feature extraction level (Method 3 in Table 6.25) achieved the best segmentation results. The reason for performance of Method 1 being better than that of Method 2 is that the Method 2 uses two GoogleNet networks trained on smaller data subsets, leading to their poorer feature extraction performance. The reason why Method 3 performs better than Method 1 is that Method 3 uses two random forests (classifiers) to discriminate between morphological structures separately for images with benign and malignant tissue, whereas Method 1 only uses a single classifier (random forest) which may find harder to draw decision boundary between features representing gland and surrounding tissue where glands have very different morphology.

6.7 Summary

This chapter is focused on experimental validation of various methods proposed in the reported research. First, a number of metrics are proposed for region and shape based evaluation of the segmentation results. The properties of these metrics are discussed and the argument is but forward, for using proposed object-level boundary Jaccard index measure instead of frequently used in literature (Sirinukunwattana et al., 2017) object-level Hausdorff distance. Subsequently, the ranking strategy adopted for comparing different segmentation schemes is explained. Performance of segmentation methods is strongly depended on values of their design parameters. Therefore, design parameters for the investigated feature spaces, including hand-crafted and deep features, are experimentally investigated with the relevant values selected based on their best experimental performance.

In Chapter 5, a number of image segmentation approaches have been proposed. In

this chapter the performance of these methods have been experimentally validated using consisted validation framework. Firstly, segmentation without pre-classification has been tested when used with different set of feature spaces. It has been shown that this approach achieved best results when used with the GooleNet deep features. It has been also demonstrated that when the segmentation is performed on the subset of the data, representing exclusively benign or malignant tissue only, the segmentation performance improves and the hand-crafted features, namely ring histograms, can provide competitive results. This has confirmed hypothesis proposed in chapter 5, that a hybrid method combining image-level and pixel-level classification could improve overall performance of gland segmentation. Subsequently a number of the image-level classification methods, necessary for implementation of the hybrid methods, have been tested. It has been demonstrated that with the proposed local deformation and colour jitter training data augmentation, the ResNet-50 deep model provides excellent image level classification, with 100% accuracy on the available test data.

Two hybrid segmentation methods with the image pre-classification have been tested. Segmentation with pre-classification at the feature extraction level is to separate images into benign or malignant categories first and then train the corresponding pixel-level classifiers, whereas segmentation with pre-classification at the classifier level is to extracted the features first and then separate the features into benign and malignant cases and train the corresponding random forest models. From the reported experimental results, it can be seen that the hybrid segmentation method with the pre-classification at the pixel-level classifier, using deep GoogleNet features, provides the best segmentation results overall, but also when only images with benign or malignant tissue are segmented. This is because the pre-classification helps the classifiers in each classification path adopted to specific characteristics of the benign or malignant gland tissue, and in the same time, the deep features are trained with all the available data enabling the learning method to find the most discriminative deep image features.

Chapter 7

Summary, contributions and future work

This chapter summarises the research reported in the thesis, details the novel contributions and describes future research plans. The publication list is provided in Appendix G.

7.1 Summary

A brief review of image segmentation methods is included in Chapter 1. The described methods are generic image segmentation ordered by increasing complexity, starting with simple segmentation approaches based on the homogeneous regions, through semantic segmentation, to the most complex approach (instance segmentation).

The definition of gland instance segmentation and the comprehensive review of gland segmentation methods are provided in Chapter 2. Subsequently, the principles of the segmentation methodology adopted in this work, utilising pixel-level classification and the corresponding feature spaces (including both deep and hand-crafted features) investigated in this work, are provided in Chapters 3 and 4.

The two main approaches investigated in this work, i.e. segmentation with and without pre-classification, are detailed in Chapter 5. The reasons for developing the proposed hybrid methods, with pre-classification, are explained in Section 5.3, whereas the hybrid methods themselves are introduced in Section 5.4.

For these proposed hybrid methods, Chapter 6 describes extensive comparative experiments, with both deep and hand-crafted features. Based on the experimental results, it can be concluded that hybrid segmentation using learned deep features performs better than when using hand-crated features. The strong discriminative properties of deep features are also confirmed with the evidence provided in Appendix D. Although hand-crafted features don't provide as strong discrimination between different tissue types (i.e. benign and malignant), they still can be effectively used for

segmentation, especially for benign tissue glands.

The performance of hybrid methods, combining two-level classifications (i.e. image and pixel level) highly depends on the performance of pixel-level classification, as the proposed image level classification is very effective achieving 100% accuracy on the available test data. The experimental evidence demonstrates that segmentation with pre-classification is more effective in forcing the random forest classifier to separate between different tissue categories.

Four evaluation metrics are used to assess quality of the gland segmentation, with two of them used for assessment of shape similarity between predefined ground truth and segmentation results. The Boundary Jaccard index measure has been adapted for assessment of gland instance segmentation. It is more suitable for a shape similarity measure than the previously used object-level Hausdorff distance, as it is not sensitive to the outliers and is bounded. The relevant evidence is reported in Chapter 6, Section 6.1.3.

7.2 Contributions

The novel research contributions include image classification with proposed data augmentation for re-training deep learning models, a hybrid segmentation model combining image and pixel level classification, and the object-level Boundary Jaccard metric adopted for evaluation of instance segmentation methods.

Image classification

In Chapter 5, Section 5.4.1, an image classification algorithm is proposed as part of the segmentation method with image pre-classification. In Chapter 6, Section 6.4.1, it is shown that this method is able to effectively differentiate between benign and malignant gland images with very high accuracy (100% on the available test data), when used with the proposed data augmentation methods.

The data augmentation methods have been used in this work, include colour jitter and local deformations. The colour jitter is introduced to change the colour without changing the morphological structure of the gland objects, whereas local image deformations are used to modify shape of structures and textures without changing the

colour. With the proposed augmentations, the image-level classifier can learn more generic image features differentiating between benign and malignant glands, improving overall performance of the image classifier.

Segmentation methods

A number of segmentation methods are proposed in Chapter 5, Section 5.4.2.3. The random forest method is selected for the pixel-level classification. This is because it provides a good compromise between accuracy, scalability and flexibility of the design. Different configurations of random forests and their corresponding design parameters are evaluated using representative selection of data subsets from the UCI database in Chapter 3, Section 3.4. Subsequently the best preforming M-A-GI configuration was selected for the pixel-level classification in the proposed segmentation methods.

Two segmentation processing pipelines, with and without image pre-classification, are proposed. The processing with image pre-classification is further divided into pre-classification at the feature extraction level and the pixel-classification level. In Chapter 6, Section 6.3 to 6.6, these processing architectures are extensively tested with different gland categories, number of target classes, and different image feature sets. It is demonstrated that all these parameters are important when selecting the optimal segmentation method for a given problem. The experimental test results demonstrate that the segmentation with pre-classification, implemented at the pixel-classification level, with the Google deep image features provides the most competitive results overall. The reason for the segmentation with pre-classification at pixel-level classification (Method 3 in Chapter 5, Section 5.4.2.3) having the best performance is that the deep features are learned using all the available data, i.e. including images of both benign and malignant tissue. However, the pre-classification step helps the pixel-level classifiers to select the most distinctive deep features for segmentation of gland objects separately for the benign and malignant cases.

Evaluation metrics

In Chapter 6, Section 6.1, a number of region and contour based metrics are described for evaluation of the proposed gland instance segmentation methods. These include object-level F1 score, object-level Dice index, object-level Hausdorff distance and

the proposed object-level Boundary Jaccard index adopted for the instance segmentation evaluation. It has been demonstrated that all these metrics have complementary characteristics, leading to the conclusion that they should be combined to provide a more holistic assessment of the segmentation methods.

These different metrics are evaluated from the perspective of their significance for ranking different segmentation methods. It is argued that the object-level Boundary Jaccard metric adopted in the thesis for evaluation of instance segmentation methods is more suitable for segmentation ranking than the previously used object-level Hausdorff distance, as it is not sensitive to outliers and is bounded, therefore can be easily integrated with region-based metrics such as the object-level Dice index.

7.3 Future research

Image representation for images with malignant tissue

Although this research provides comprehensive comparative results for different image representations for both benign and malignant tissue, more features could be investigated providing a more complete picture for selection of the optimal image features. For example, the Gabor features, Haar features or SIFT features are possible options for further investigations.

The random forests techniques can facilitate semi-supervised learning. It would be interesting to adopt the techniques proposed in this research to work in a semi-supervised fashion and therefore enable handling larger, not fully annotated training datasets – reducing the burden for manual segmentation were much larger datasets to become eventually available. Furthermore, operation of the classifier could be better understood, by finding the most important features or by describing feature interactions.

Several sizes of input patches have been tested in order to find the one that effectively describes the local patterns. However, one could instead consider using patches of different sizes at the same time. For example, small input patches would represent the local image characteristics (corners, edges, etc.) with high spatial resolution, medium size patches would describe local morphological structure, whereas the largest patches would describe local image context. The classifier, in this case, would

not only use the local information but also take advantage of the more global contextual image information, which might improve the segmentation performance.

End-to-end feature extraction methods and random forest techniques

Both segmentation with and without pre-classification are patch-based approaches which use sliding windows to extract the local patterns from the histology gland images and use these patterns to train the classifier and make predictions. One of the disadvantages of these patch-based approaches is that the classifier could not learn and use the global contextual information present in the histology images.

In recent years, fully convolutional end-to-end architectures have become popular for semantic instance image segmentation. These end-to-end approaches can directly compute the segmentation results, and can utilise contextual information about the overall structure and the patterns in a whole image. It would be interesting to test the limitations of the end-to-end architectures and consider construction of hybrid methods which would combine fully convolutional deep learning architecture for feature vector estimation and random forest models as the pixel-level classifier.

The results reported in literature (Long et al., 2015; Xu et al., 2016) on using of the fully convolutional neural networks for semantic and instance segmentation are impressive (the performance of fully convolutional neural networks achieves top-rank). However, interpretation and understanding of these results is somewhat difficult. Although the overall performance of these networks is very good, it is not easy to associate this with any particular characteristics of the images or indeed specific parts of the network. It would be interesting to understand what specific image characteristic lead to successful gland segmentation.

Post-processing method

In this work, morphological and level-set techniques have been used for post-processing. Although currently segmentation using morphological processing provides overall better segmentation rankings, a more complex level-set method could be further investigated, i.e. including analysis of the level-set segmentation with topology constraints which proved useful for other segmentation problems.

Appendices

Appendix A

Random forest results on UCI dataset

The following figures A.1 to A.4 shows the results of different forest models on different datasets using different evaluation measures.

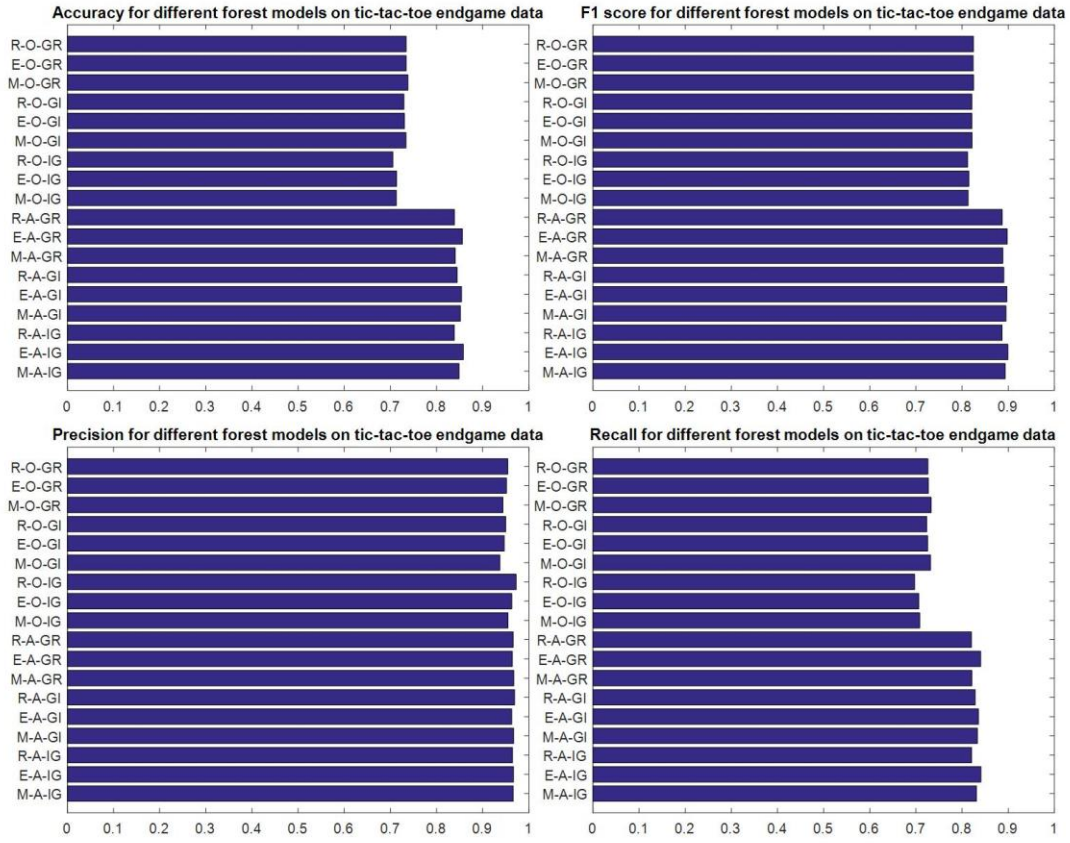


Figure A.1 Different evaluation measures of different forest models on **tic-tac-toe endgame**

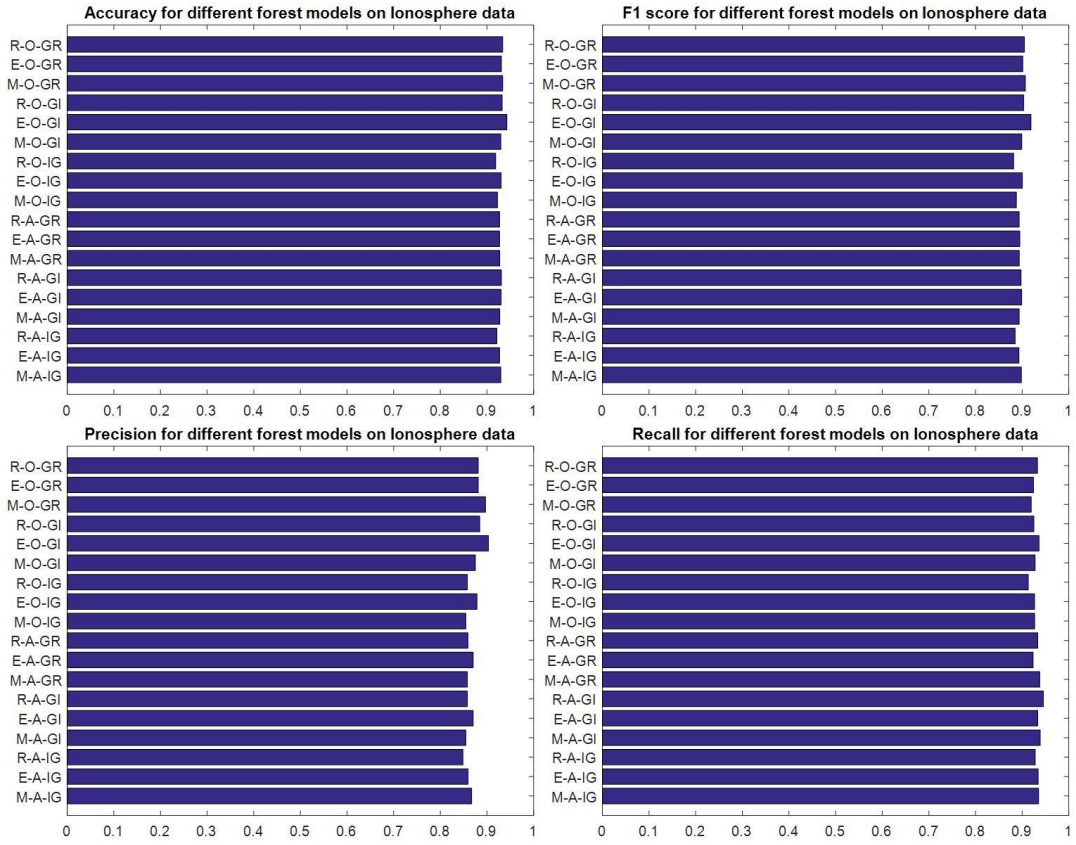


Figure A.2 Different evaluation measures of different forest models on **Ionosphere** data

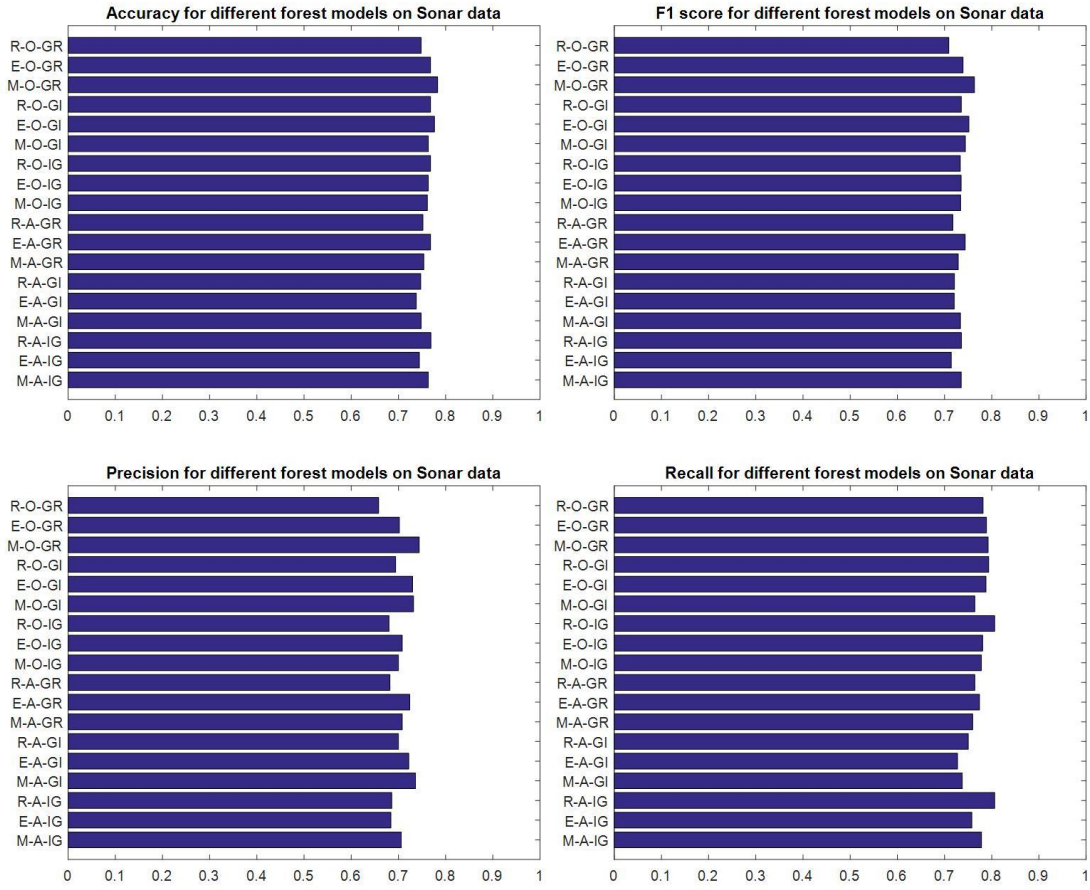


Figure A.3 Different evaluation measures of different forest models on **Sonar** data

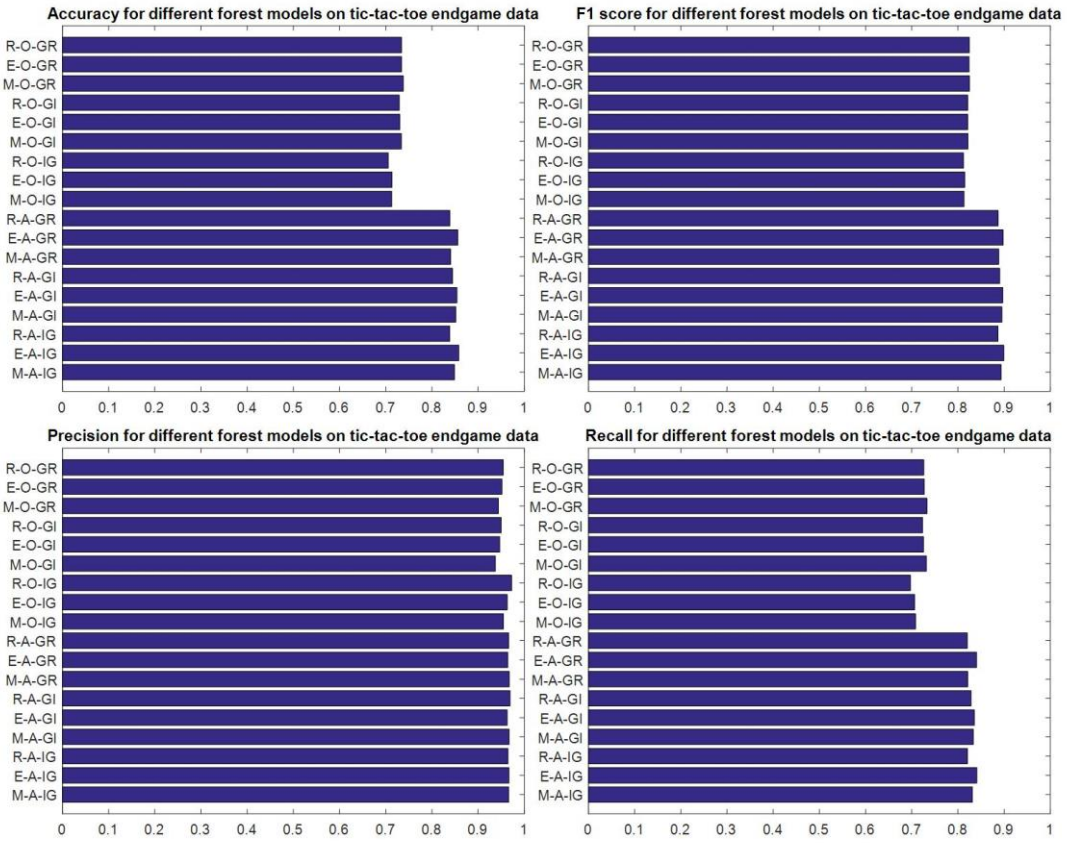


Figure A.4 Different evaluation measures of different forest models on **Tic-tac-toe** data

Appendix B

Visualised extracted features

In the following figures, the green points represent the gland feature and the red points represent the background feature.

Ring histogram in segmentation without pre-classification

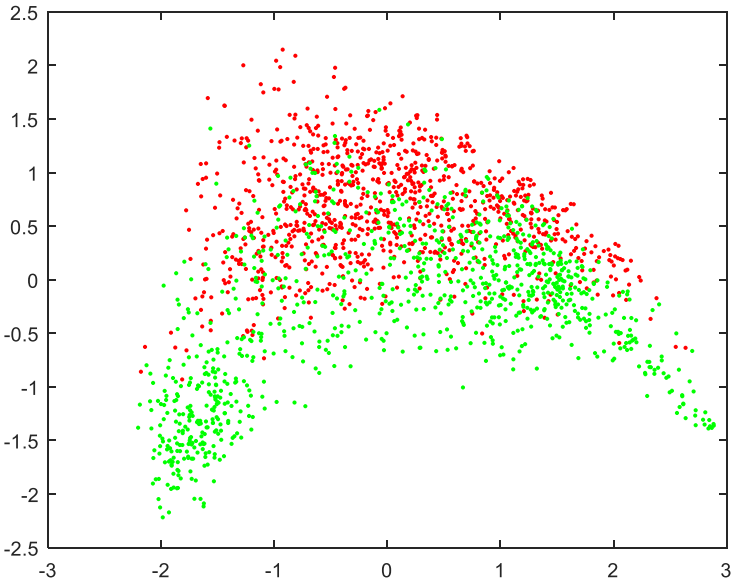


Figure B.1 Ring histogram after using PCA algorithm in 2D feature space.

HOG feature in segmentation without pre-classification

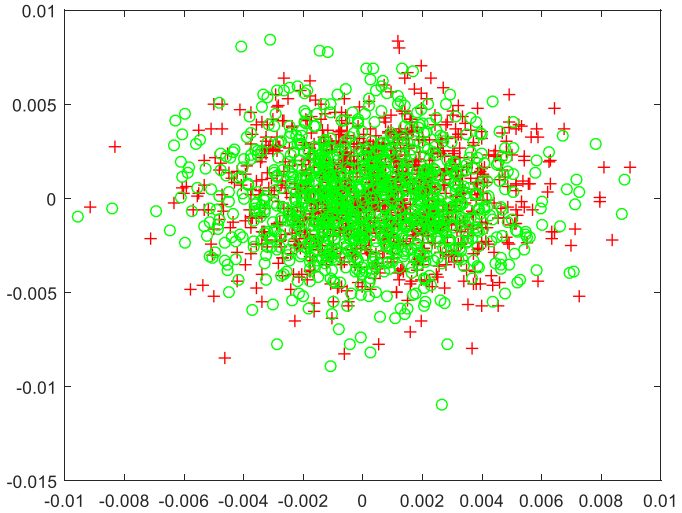


Figure B.2 The original HOG feature after using PCA algorithm in 2D feature space

Visualising circular Fourier HOG feature

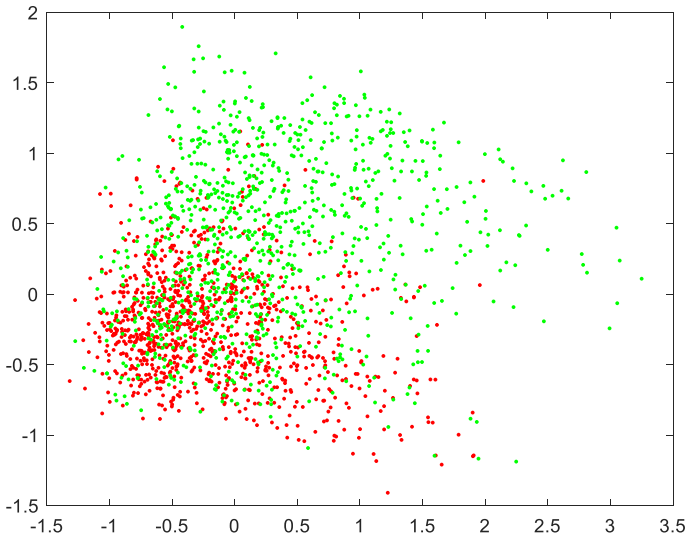


Figure B.3 The circular Fourier HOG feature after using PCA in 2D feature space

LBP feature in segmentation without pre-classification

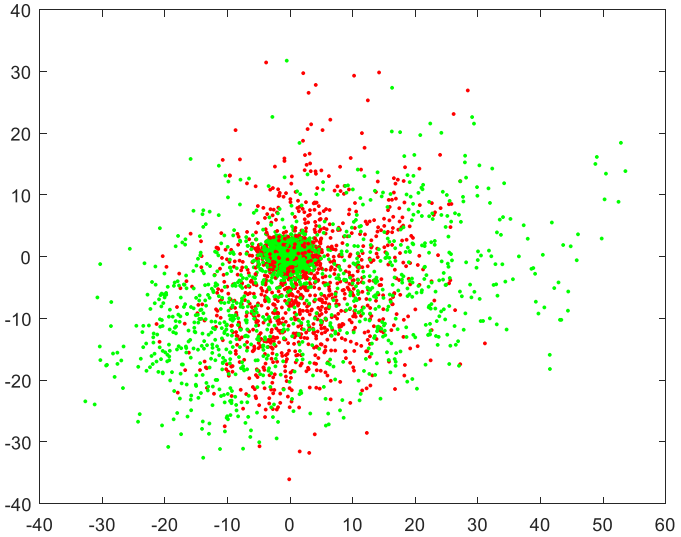


Figure B.4 Original LBP feature after PCA algorithm in 2D feature space

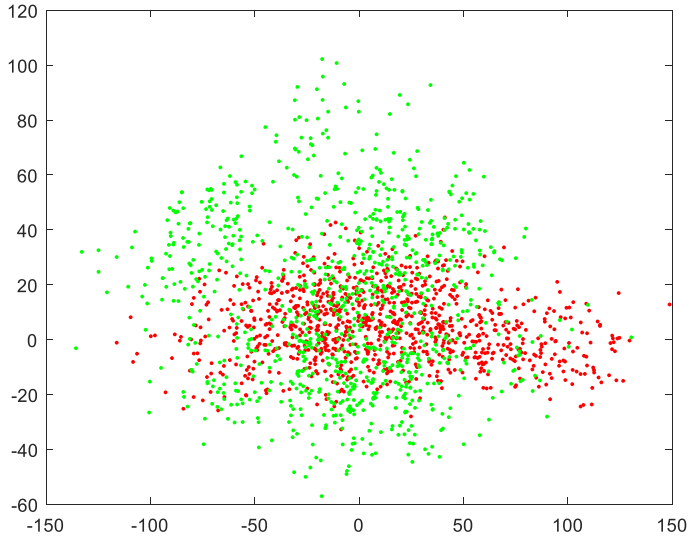


Figure B.5 Original rotation-invariant LBP feature after PCA algorithm in 2D feature space

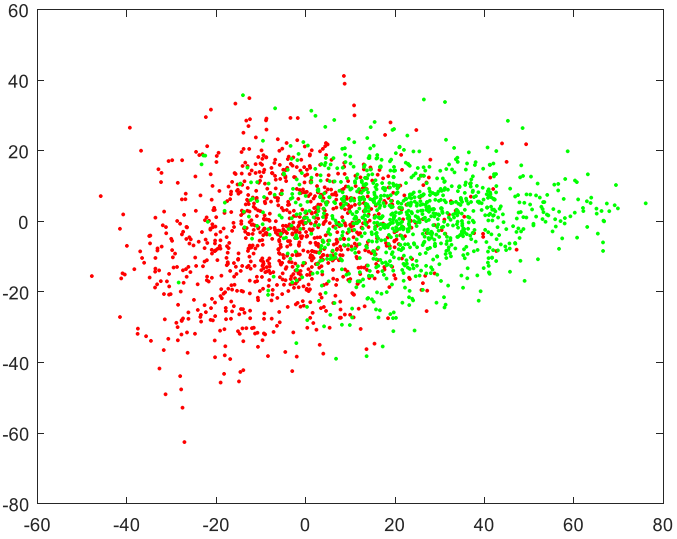


Figure B.6 Uniform LBP after using PCA algorithm in 2D feature space

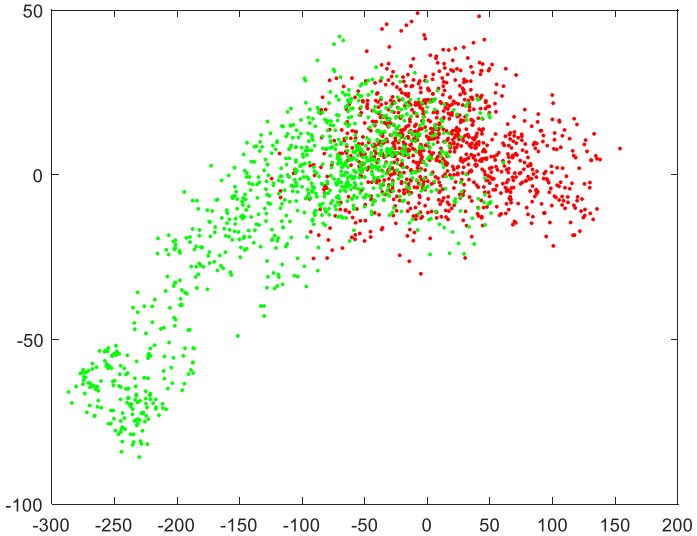


Figure B.7 Rotation-invariant uniform LBP after using PCA algorithm in 2D feature space

Deep features in segmentation without pre-classification

Figure B.8 shows the deep features from the LeNet-5 architecture.

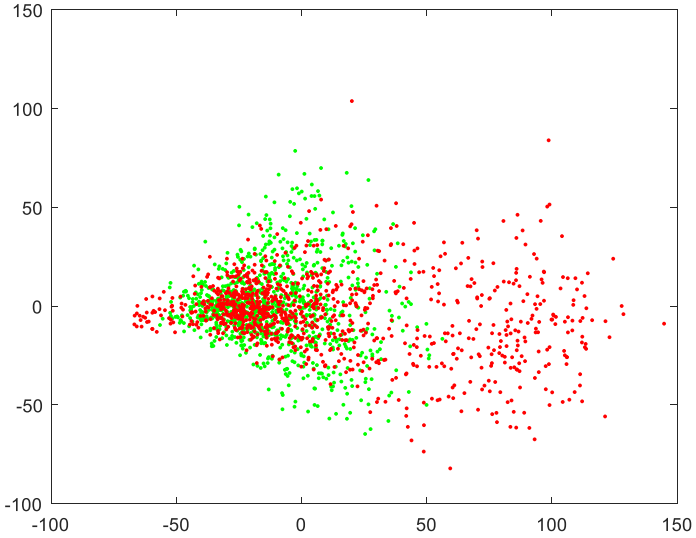


Figure B.8 Deep feature from LeNet-5 after using PCA algorithm in 2D feature space

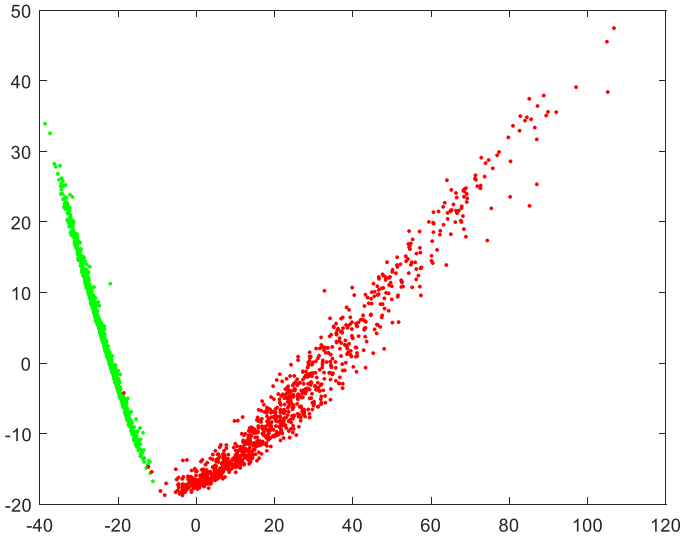


Figure B.9 Deep features from GoogleNet after using PCA algorithm in 2D feature space

Visualised features for benign category

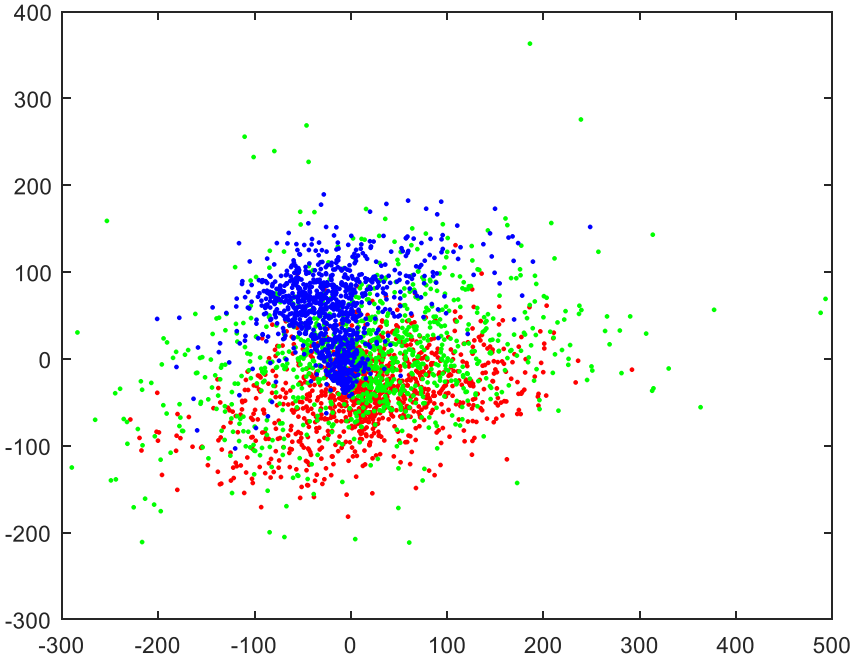


Figure B.10 Deep features from LeNet-5 with three classes after using PCA in 2D feature space

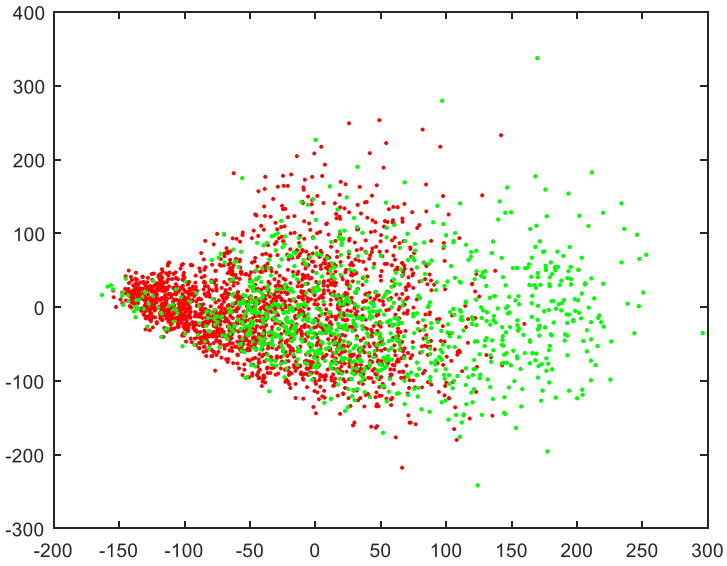


Figure B.11 Deep features from LeNet-5 with 2 classes after using PCA in 2D feature space

Visualised features for malignant category

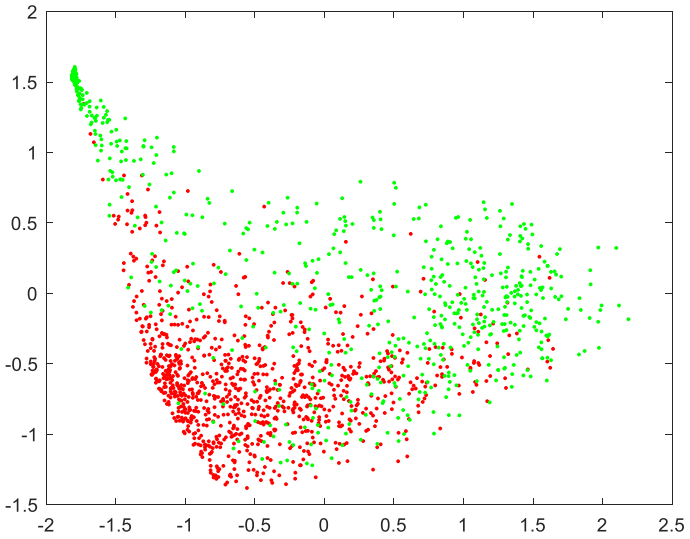


Figure B.12 Ring histogram with 2 classes after using PCA in 2D feature space in segmentation with pre-classification

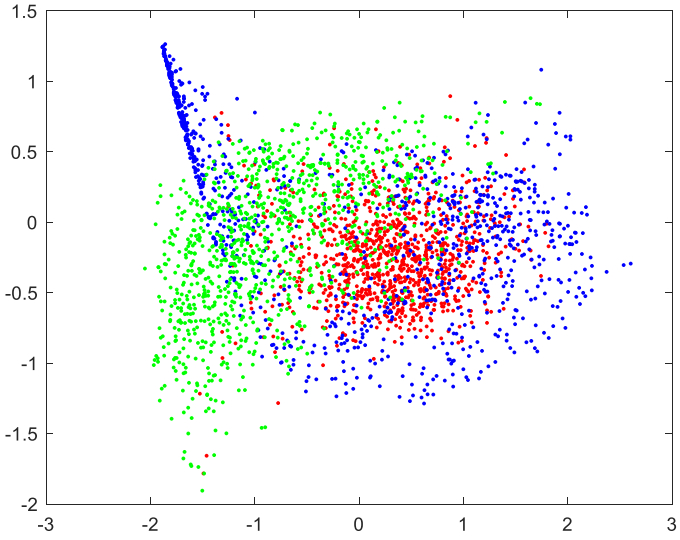


Figure B.13 Ring histogram with 3 classes after using PCA in 2D feature space in segmentation with pre-classification

Appendix C

Evaluation of features discriminative properties

C.1 Estimation of the features generated from the GLCM

Four different feature vectors are generated from the gland images:

- Model 1: The feature generated from the co-occurrence matrix in direction $\{0^{\circ}\}$
- Model 2: The feature generated from the co-occurrence matrix in direction $\{45^{\circ}\}$
- Model 3: The feature generated from the co-occurrence matrix in direction $\{90^{\circ}\}$
- Model 4: The feature generated from the co-occurrence matrix in direction $\{135^{\circ}\}$

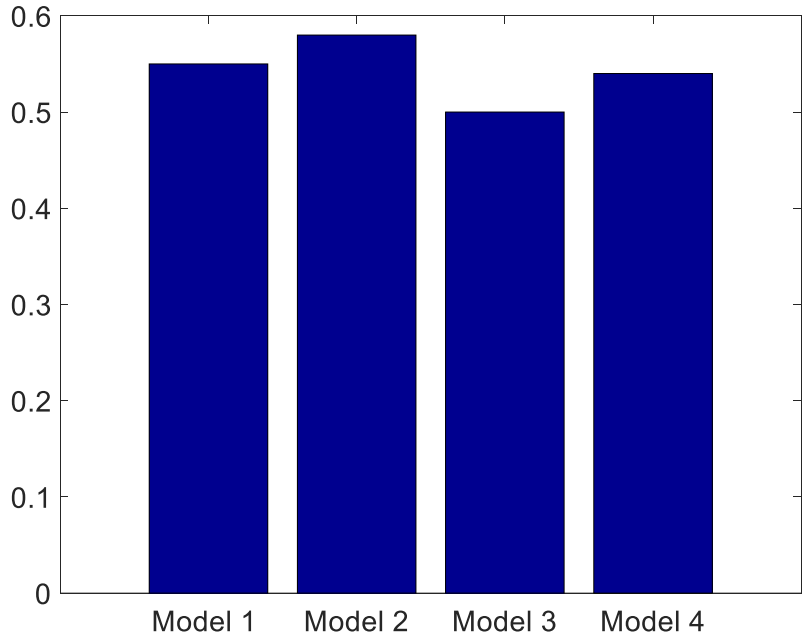


Figure C.1 The accuracy of different features generated from different co-occurrence matrices from the gland images

C.2 Estimation of discriminative properties of histogram features

Estimate discriminative properties of different sizes of patch using histogram feature in segmentation without pre-classification

- Model 1: the ring histogram feature generated from 15-by-15 patches
- Model 2: the ring histogram feature generated from 19-by-19 patches
- Model 3: the ring histogram feature generated from 23-by-23 patches
- Model 4: the ring histogram feature generated from 27-by-27 patches

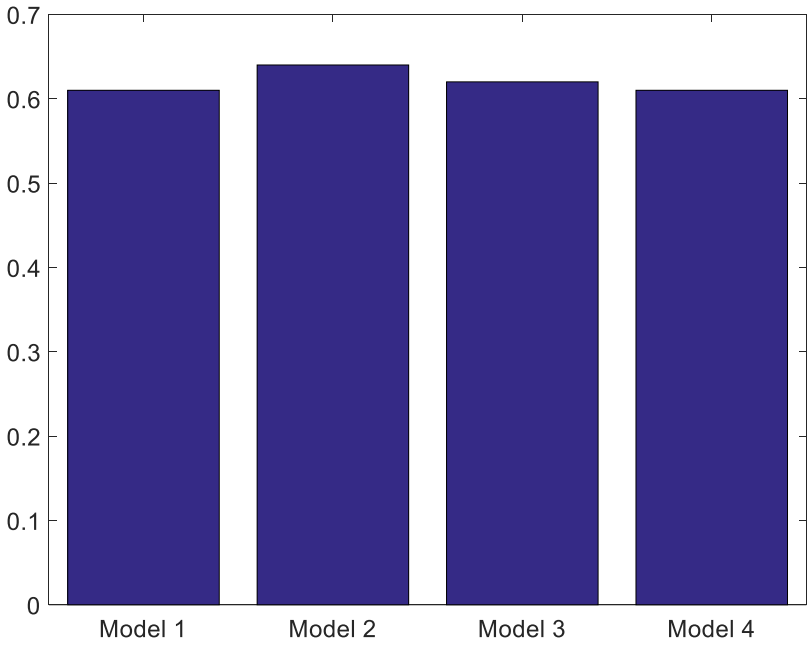


Figure C.2 Accuracy of different models (different sized patches) using K-means algorithm

Estimate discriminative properties for different number of rings in each patch of ring histogram in segmentation without pre-classification

- Model 1: 7 different rings were used to extract the patterns from the patches.
- Model 2: 8 different rings were used to extract the patterns from the patches.
- Model 3: 9 different rings were used to extract the patterns from the patches.

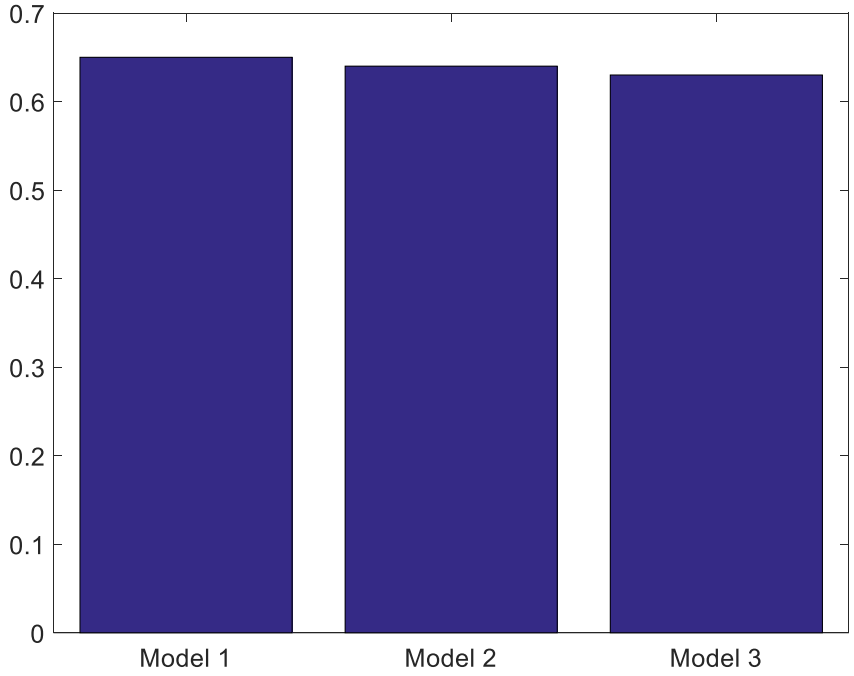


Figure C.3 Accuracy of different models (different number of rings in each patch) using K-means

Estimate discriminative properties for all ring histogram features in segmentation with a pre-classification approach

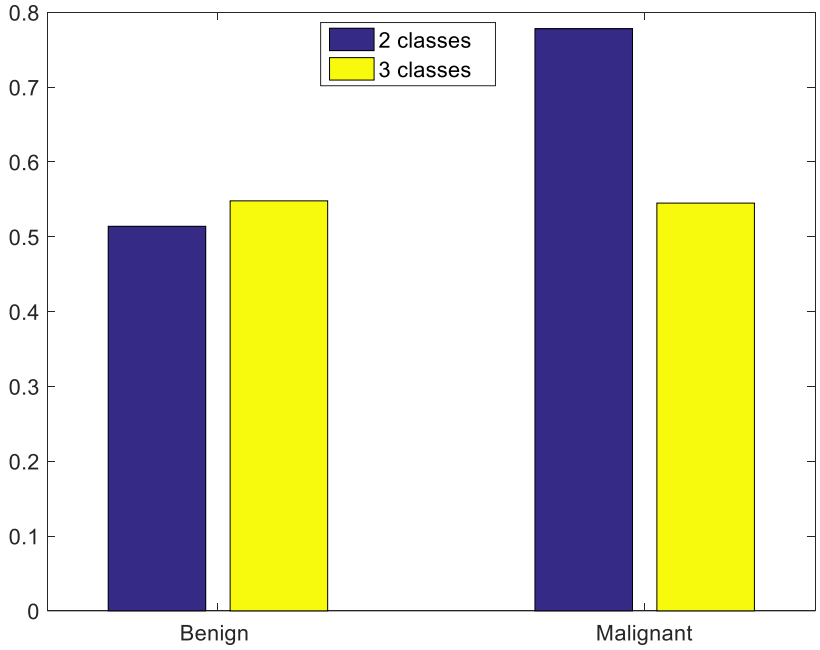


Figure C.4 Accuracy of ring histogram features in segmentation with pre-classification method

C.3 Estimation of the LBP features

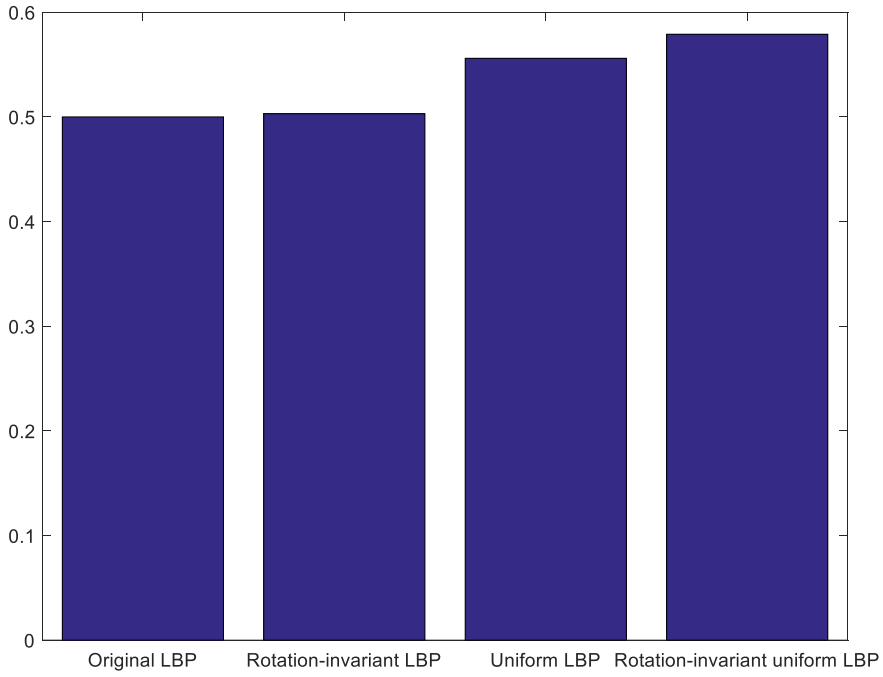


Figure C.5 The estimation results for different LBP features in segmentation without pre-classification

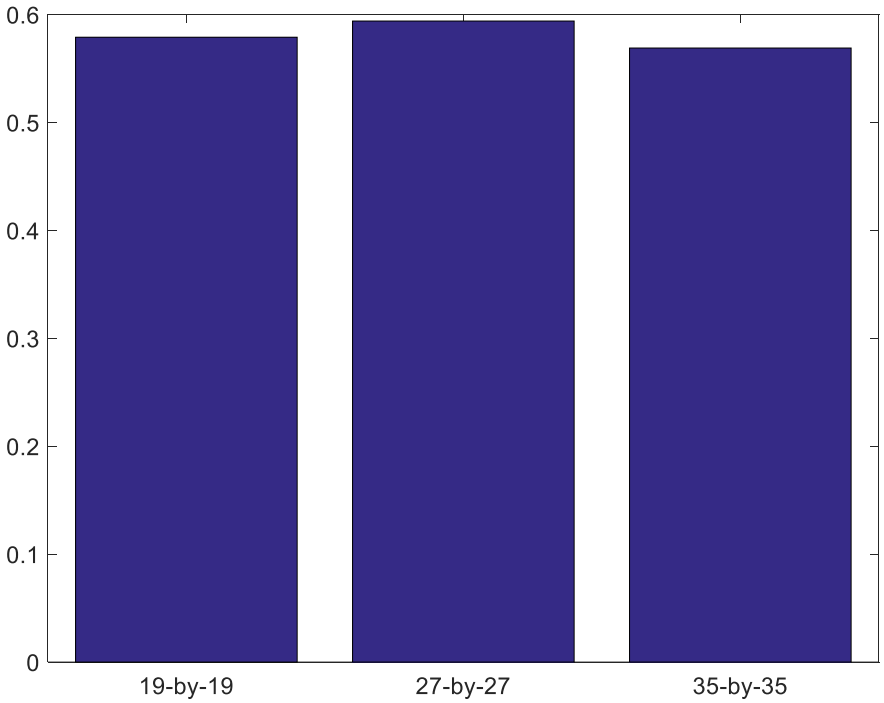


Figure C.6 Estimation results for different size of rotation-invariant uniform LBP feature

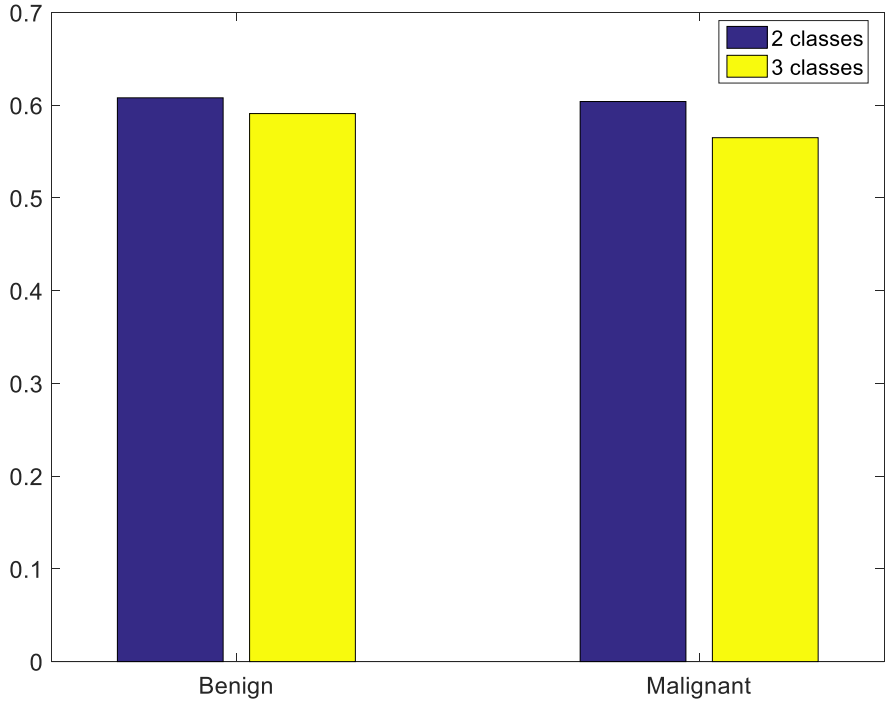


Figure C.7 Accuracy of rotation-invariant uniform LBP features in segmentation with pre-classification

C.4 Estimation of the HOG features

- Model 1: Extracting the original HOG feature vectors from the histology images
- Model 2: Extracting the circular Fourier HOG feature from histology images

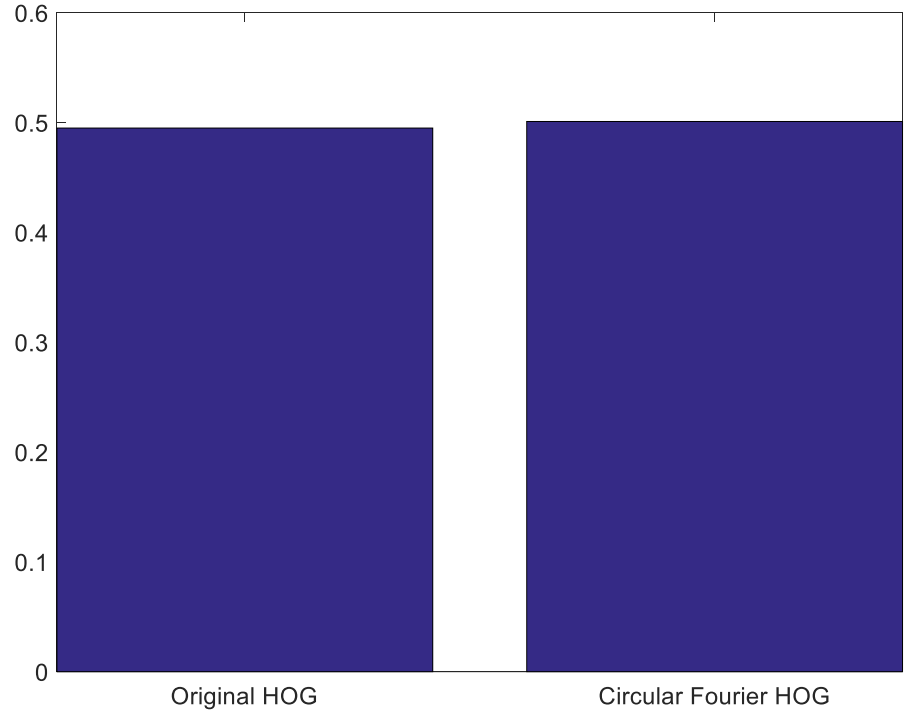


Figure C.8 Accuracy of two versions of HOG features in segmentation without pre-classification.

- Model 1: Extracting circular Fourier HOG feature using 19-by-19 patches
- Model 2: Extracting circular Fourier HOG feature using 27-by-27 patches

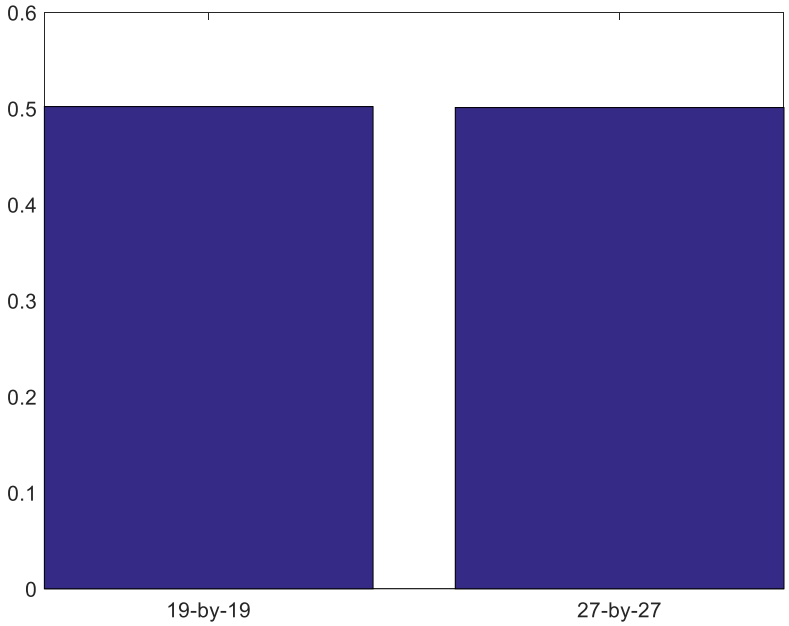


Figure C.9 Accuracy for different Fourier HOG features in segmentation with pre-classification

- Model 1: Extracting circular Fourier HOG using 2-by-2 circle in each selected patch
- Model 2: Extracting circular Fourier HOG using 5-by-5 circle in each selected patch

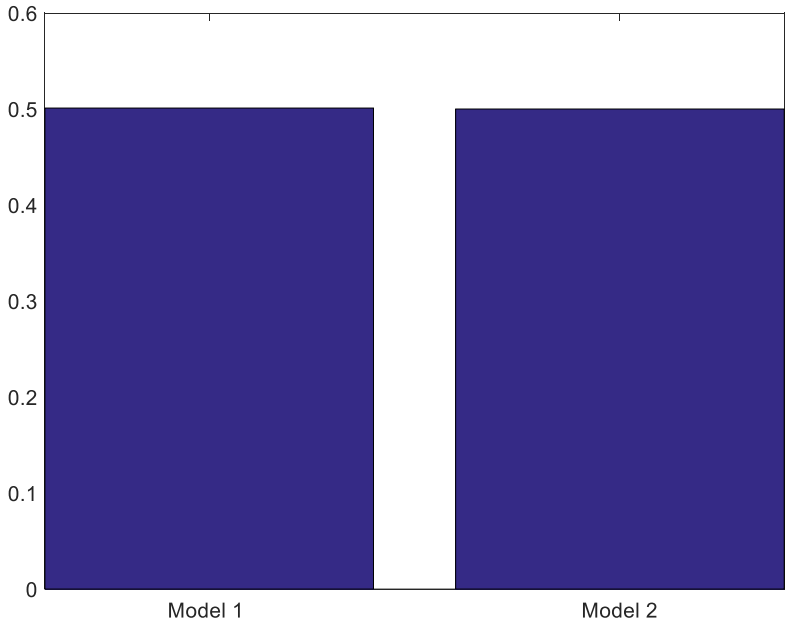


Figure C.10 The accuracy of circular Fourier HOG with different sizes of circles in each patch

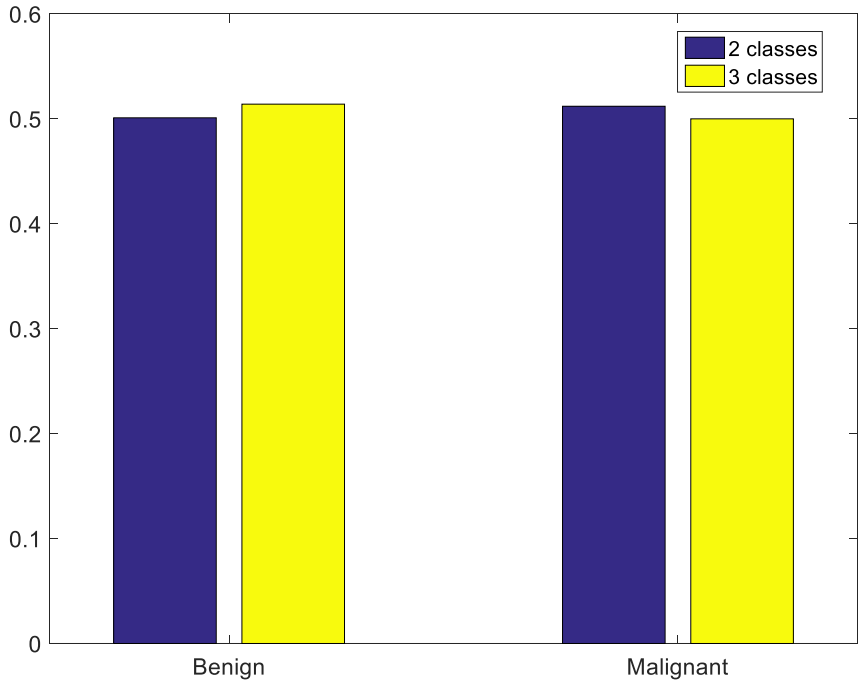


Figure C.11 Accuracy of different circular Fourier HOG in segmentation with pre-classification

C.5 Estimation of the deep features

- Model 1: Using 85,000 19-by-19 patches to generate deep learning features from LeNet-5 architecture.
- Model 2: Using 170,000 19-by-19 patches to generate deep learning features from LeNet-5 architecture.

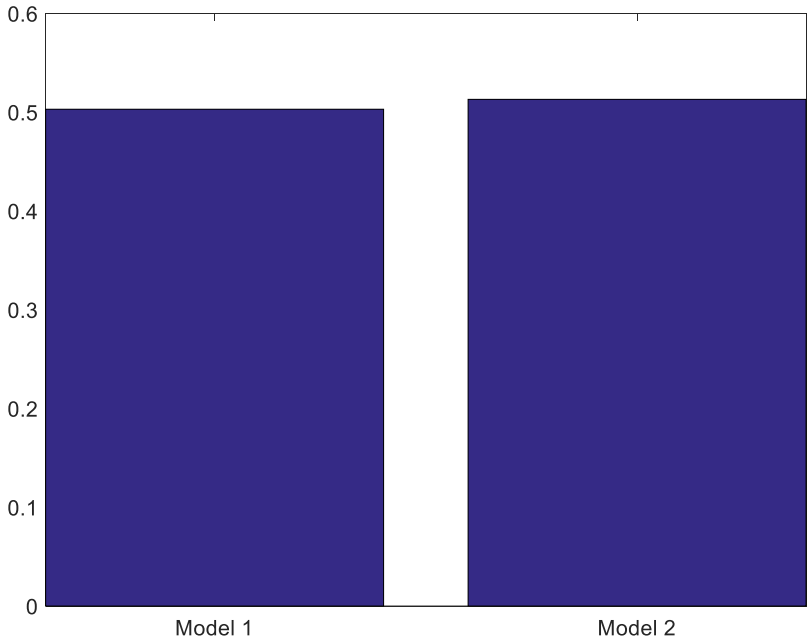


Figure C.12 Accuracy of deep feature from LeNet-5 architecture with different numbers of input patches in segmentation without pre-classification

- Model 1: Using 85,000 15-by-15 patches to generate deep learning features from LeNet-5 architecture.
- Model 2: Using 85,000 19-by-19 patches to generate deep learning features from LeNet-5 architecture.

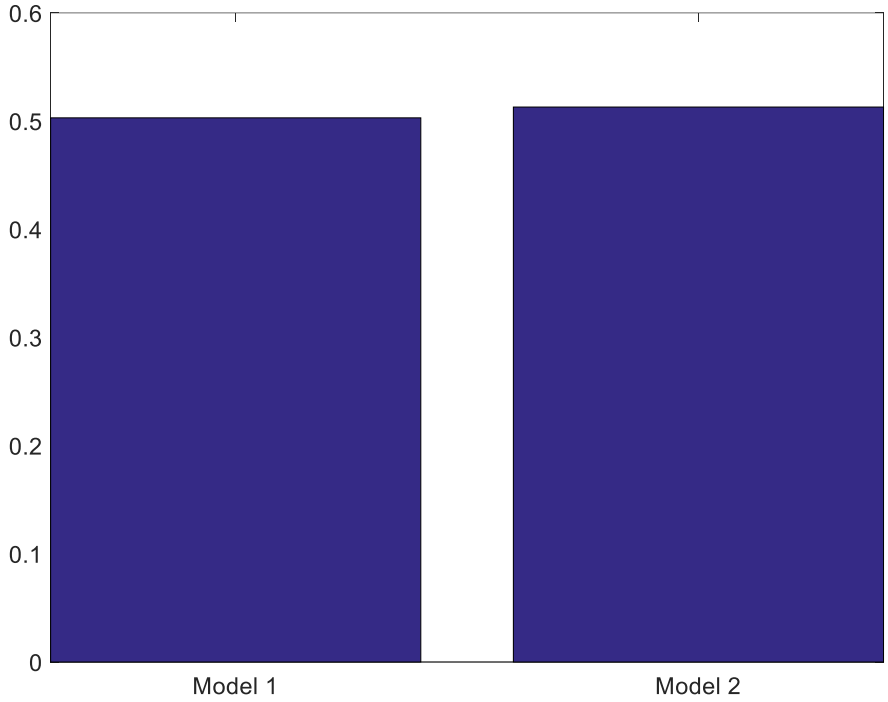


Figure C.13 Accuracy for different deep learning feature vectors from LeNet-5 architecture with different sizes of input patches in segmentation with pre-classification

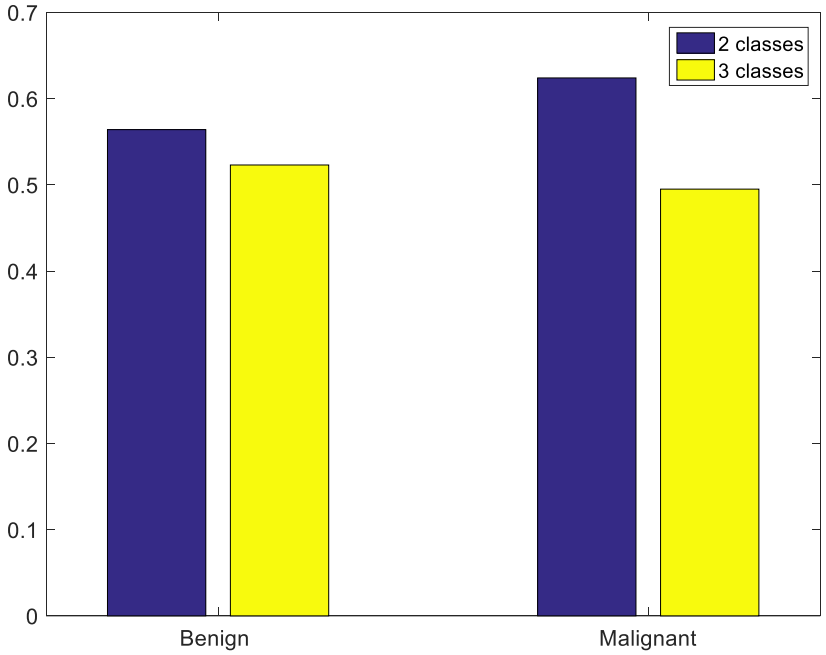


Figure C.14 Accuracy of different deep learning features with LeNet-5 architecture in segmentation with pre-classification.

- Model 1: Using 85,000 49-by-49 patches to generate deep learning features from GoogleNet architecture.
- Model 2: Using 85,000 97-by-97 patches to generate deep learning features from GoogleNet architecture.
- Model 3: Using 85,000 225-by-225 patches to generate deep learning features from GoogleNet architecture.

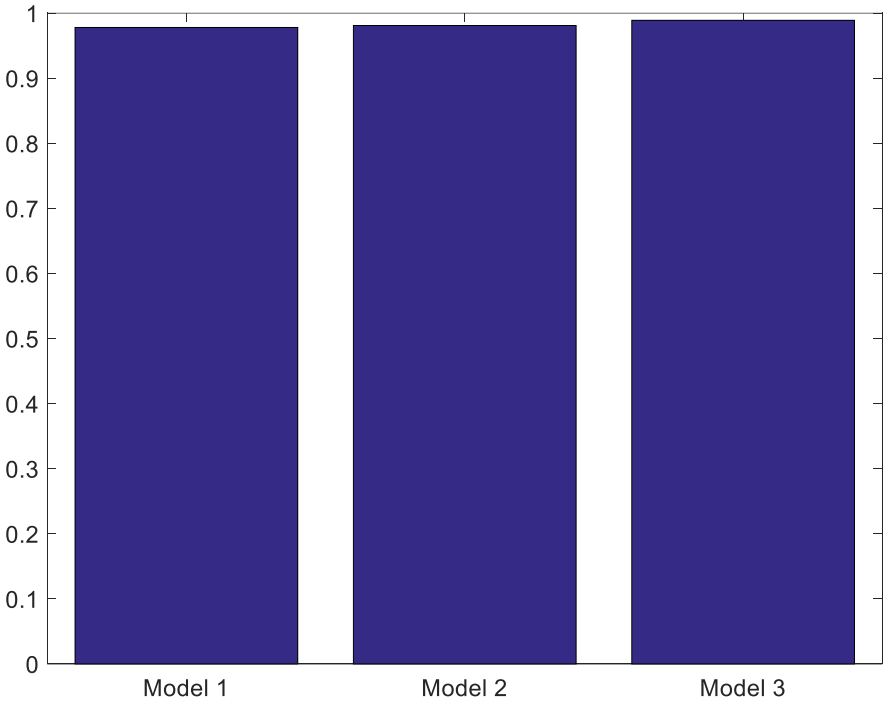


Figure C.15 Accuracy for different deep learning features from GoogleNet with different sizes of input patches in segmentation with pre-classification

Over 90% of data after K-means clustering could separate data in the same category as the original deep features.

- Model 1: Using 85,000 225-by-225 patches to generate deep learning features from GoogleNet architecture.
- Model 2: Using 170,000 225-by-225 patches to generate deep learning features from GoogleNet architecture.

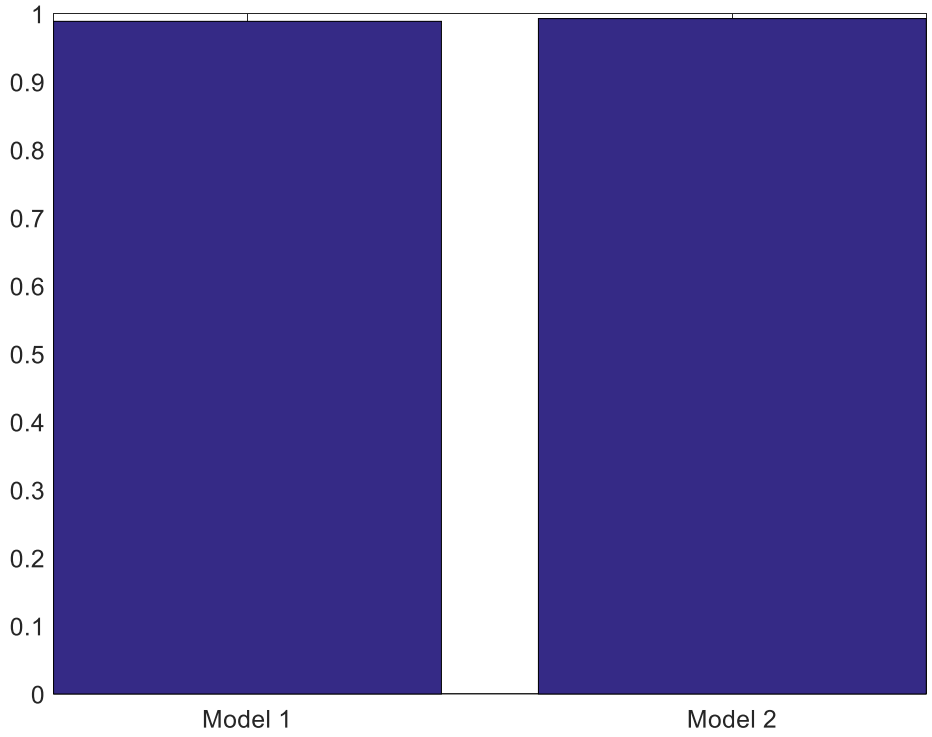


Figure C.16 Accuracy of deep features from GoogleNet with different input patches after using K-means clustering in segmentation without pre-classification

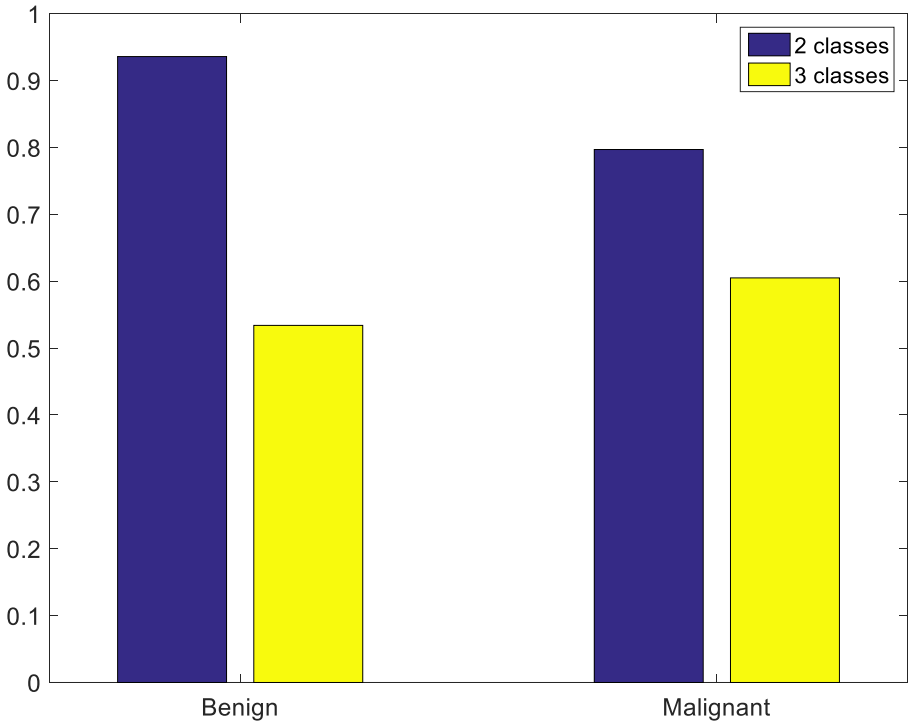


Figure C.17 Accuracy of deep features from GoogleNet with different input patches after using K-means clustering in segmentation without pre-classification

Appendix D

Segmentation results without pre-classification

D.1 Segmentation results for grey-level co-occurrence matrix

Figure D.1 is an example of the probability maps generated by random forest using the features from co-occurrence matrix.

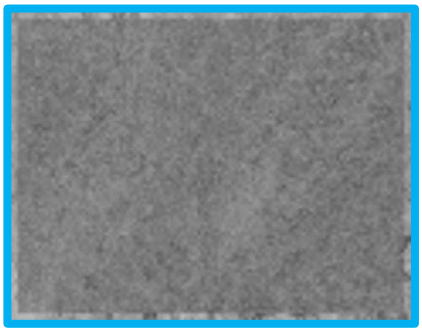


Figure D.1 Example of a probability map using features from co-occurrence matrix

The background and the gland patterns from the selected patches are similar, the classifier could not learn the differences between these two classes.

D.2 Segmentation results for LBP features

Figure D.2 shows examples of probability maps obtained from these features, which could not be used to describe the differences between gland and non-gland in the histology images.

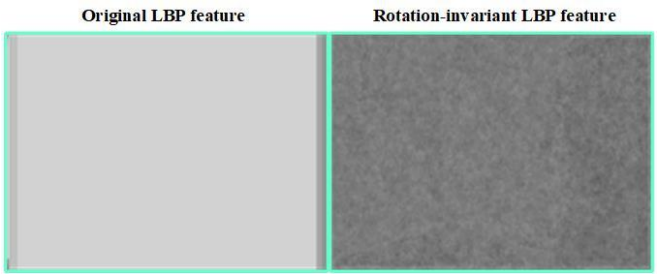


Figure D.2 Probability maps for the two failed LBP features

Figure D.3 shows segmentation results using the rotation-invariant uniform LBP and uniform LBP features.

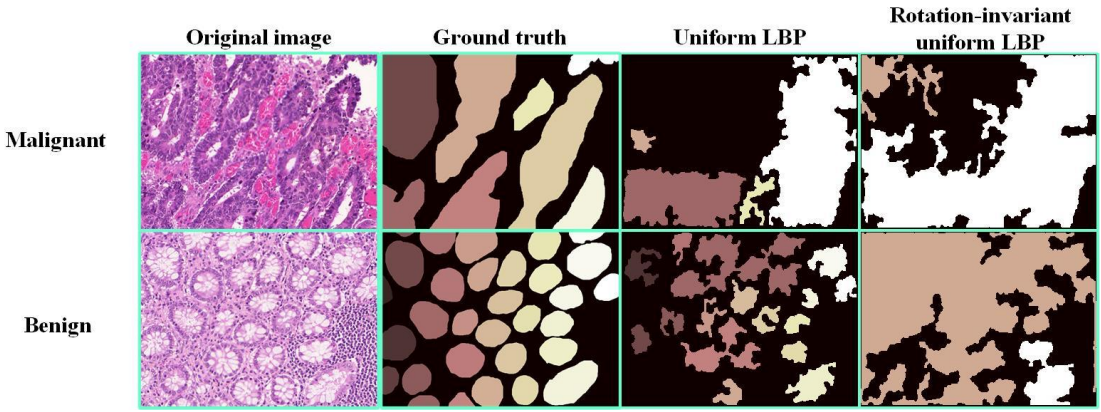


Figure D.3 Example of segmentation performance with the two uniform LBP features

Table D.1 Ranking of segmentation performance of different uniform LBP features

Different LBP	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
Rotation invariant uniform LBP	0.40 (0.19)	1 (1)	0.54 (0.25)	1 (2)	0.50 (0.14)	2 (1)	0.63 (0.17)	1 (1)	198.05 (101.01)	2(1)	264.79 (126.37)	1 (1)	8 (15)
Uniform LBP	0.40 (0.21)	1 (2)	0.10 (0.01)	2 (1)	0.51 (0.14)	1 (1)	0.23 (0.17)	2 (1)	186.18 (103.80)	1(2)	344.27(178.32)	2 (2)	9 (17)

The reason for rotation invariant uniform LBP being better is LBP with rotation-invariant properties could using fewer patterns to describe more characteristics in the images.

Table D.2 shows the quantitative results of different sized input patches for rotation-invariant uniform LBP in segmentation without pre-classification. The segmentation results are shown in Figure D.4.

Table D.2 Ranking of rotation-invariant uniform LBP feature using different sizes of input patch

Size of patches	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
19 x 19	0.38 (0.16)	2 (1)	0.61 (0.25)	1 (2)	0.46 (0.15)	3 (3)	0.55 (0.22)	3 (3)	257.32 (124.31)	3 (3)	270.08 (164.05)	2 (3)	14 (29)
27 x 27	0.40 (0.19)	1 (2)	0.54 (0.25)	2 (2)	0.50 (0.14)	1 (1)	0.63 (0.17)	1 (1)	198.05 (101.01)	2 (2)	264.79 (126.37)	1 (1)	8 (19)
35 x 35	0.35 (0.19)	3 (2)	0.54 (0.23)	2(1)	0.48 (0.14)	2 (1)	0.62 (0.2)	2 (2)	187.28 (99.83)	1 (1)	277.51 (145.82)	3 (2)	13 (22)

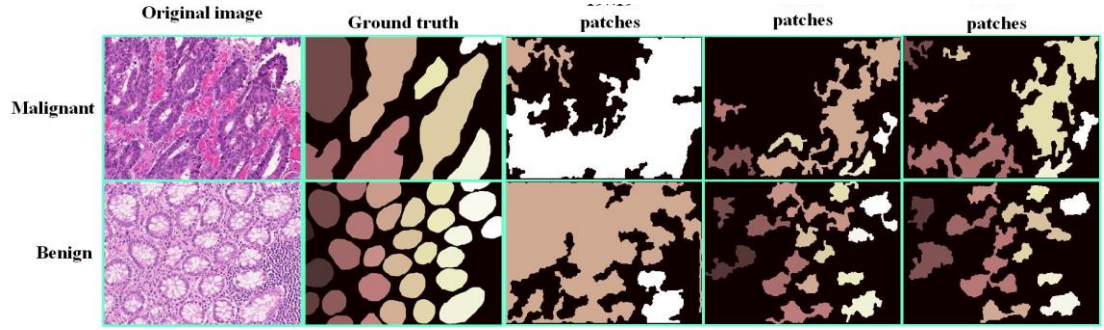


Figure D.4 Results of different input patches size using rotation invariant uniform LBP

D.3 Segmentation results for HOG features

Figure D.5 shows one map from random forest using the original HOG feature.

Figure D.5 Probability maps for HOG feature

The following experiment is designed to evaluation of significance of different size of input patches.

Table D.3 Ranking of Fourier HOG features using different sizes of circle

Size of patches	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
19 x 19	0.34 (0.12)	2 (1)	0.45 (0.24)	1 (1)	0.33 (0.19)	2 (1)	0.53 (0.18)	1 (2)	263.97 (134.98)	1 (2)	301.35 (151.87)	1 (2)	8 (17)
27 x 27	0.35 (0.19)	1 (2)	0.38 (0.16)	1 (2)	0.38 (0.16)	1 (2)	0.32 (0.14)	2(1)	323.66 (133.8)	2 (1)	366.4 (97.82)	2 (1)	9 (18)

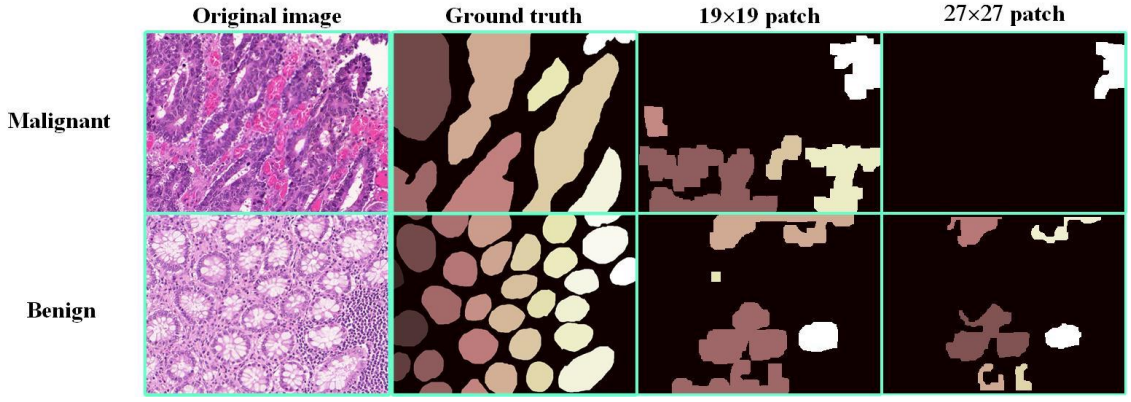


Figure D.6 Results of different input patches size using circular Fourier HOG

Appendix E

Segmentation results with pre-classification

E.1 Segmentation results for the two classes of benign category

Table E.1 shows the ranking of the segmentation performance of the different feature extraction techniques for benign category images, and again the best results are highlighted in blue.

Table E.1 Ranking of different feature of two classes of benign category in segmentation with pre-classification

Different features	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	rank	
GoogleNet	0.64 (0.20)	2 (6)	0.82 (0.11)	1 (2)	0.68 (0.15)	3 (6)	0.87 (0.08)	1 (2)	129.04 (73.06)	3 (2)	94.39 (69.07)	2 (3)	12 (33)
LeNet5	0.62 (0.15)	3 (2)	0.69 (0.11)	4 (2)	0.64 (0.12)	3 (4)	0.79 (0.08)	3 (2)	147.54 (85.95)	5 (5)	120.26 (49.19)	3 (2)	21 (38)
Ring histogram & RIULBP	0.67 (0.14)	1 (1)	0.51 (0.18)	5 (6)	0.68 (0.09)	1 (1)	0.57 (0.17)	5 (6)	123.50 (82.80)	2 (4)	328.44 (152.57)	7 (7)	21 (46)
LeNet5 & Ring histogram	0.62 (0.16)	3 (4)	0.27 (0.15)	7 (5)	0.66 (0.11)	2 (3)	0.48 (0.06)	6 (1)	113.45 (73.61)	1 (3)	292.56 (109.38)	6 (6)	25 (47)
RIULBP	0.43 (0.17)	6 (5)	0.72 (0.29)	3 (7)	0.56 (0.10)	6 (2)	0.78 (0.09)	4 (4)	140.05 (64.08)	4 (1)	133.89 (33.86)	4 (1)	27 (47)
Ring histogram	0.56 (0.20)	5 (6)	0.78 (0.02)	2 (1)	0.61 (0.15)	5 (6)	0.81 (0.12)	2 (5)	161.50 (122.87)	6 (7)	136.89 (91.81)	5 (5)	25 (55)
CHOG	0.32 (0.15)	7 (2)	0.30 (0.14)	6 (4)	0.35 (0.13)	7 (5)	0.34 (0.19)	7 (7)	363.86 (114.84)	7 (6)	395.34 (71.50)	7 (4)	41 (69)

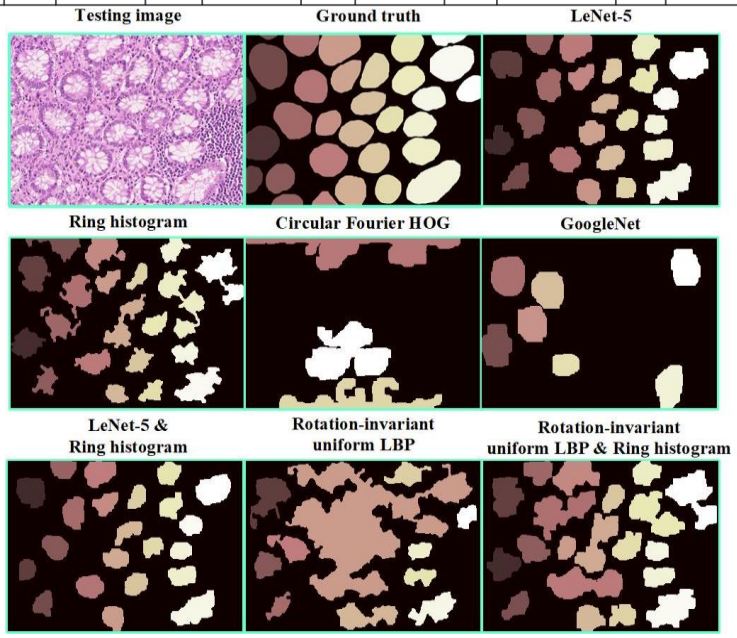


Figure E.1 Examples of results for the benign category with two classes of different features in segmentation with pre-classification

E.2 Segmentation results for the three classes of benign category

Table E.2 shows the ranking of the segmentation performance of different feature extraction methods for three classes of the benign category, with the best segmentation results again highlighted.

Table E.2 Ranking of different features of three categories of benign images in segmentation with pre-classification

Different features	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	rank	
LeNet5	0.73 (0.10)	2 (1)	0.73 (0.08)	1 (3)	0.74 (0.10)	1 (2)	0.83 (0.10)	1 (3)	102.95 (87.95)	2 (5)	144.68 (50.19)	3 (6)	10 (30)
Ring histogram	0.74 (0.11)	1 (2)	0.59 (0.18)	3 (6)	0.74 (0.08)	1 (1)	0.72 (0.09)	3 (2)	89.42 (72.23)	1 (3)	138.38 (47.75)	2 (5)	11 (30)
GoogleNet	0.55 (0.17)	4 (4)	0.62 (0.08)	2 (3)	0.6 (0.13)	5 (5)	0.73 (0.06)	2 (1)	113.49 (70.01)	5 (2)	133.14 (27.27)	1 (2)	19 (36)
Ring histogram & RIULBP	0.55 (0.20)	4 (6)	0.29 (0.01)	7 (1)	0.62 (0.11)	4 (3)	0.52 (0.12)	6 (4)	108.73 (61.33)	3 (1)	190.54 (19.96)	5 (1)	29 (45)
RIULBP	0.6 (0.19)	3 (5)	0.37 (0.17)	5(5)	0.65 (0.12)	3 (4)	0.6 (0.14)	4 (5)	111.38 (76.26)	4 (4)	177.73 (33.52)	4 (3)	23 (49)
CHOG	0.3 (0.15)	7 (3)	0.33±0.06	6 (2)	0.35 (0.13)	7 (5)	0.35 (0.16)	7 (6)	289.59 (137.06)	7 (6)	376.17 (41.42)	7 (4)	41 (67)
LeNet5 & Ring histogram	0.38 (0.22)	6 (7)	0.43±0.22	4 (7)	0.42 (0.20)	6 (7)	0.58 (0.29)	5 (7)	240.7±158.09	6 (7)	222.29 (113.81)	6 (7)	32 (75)

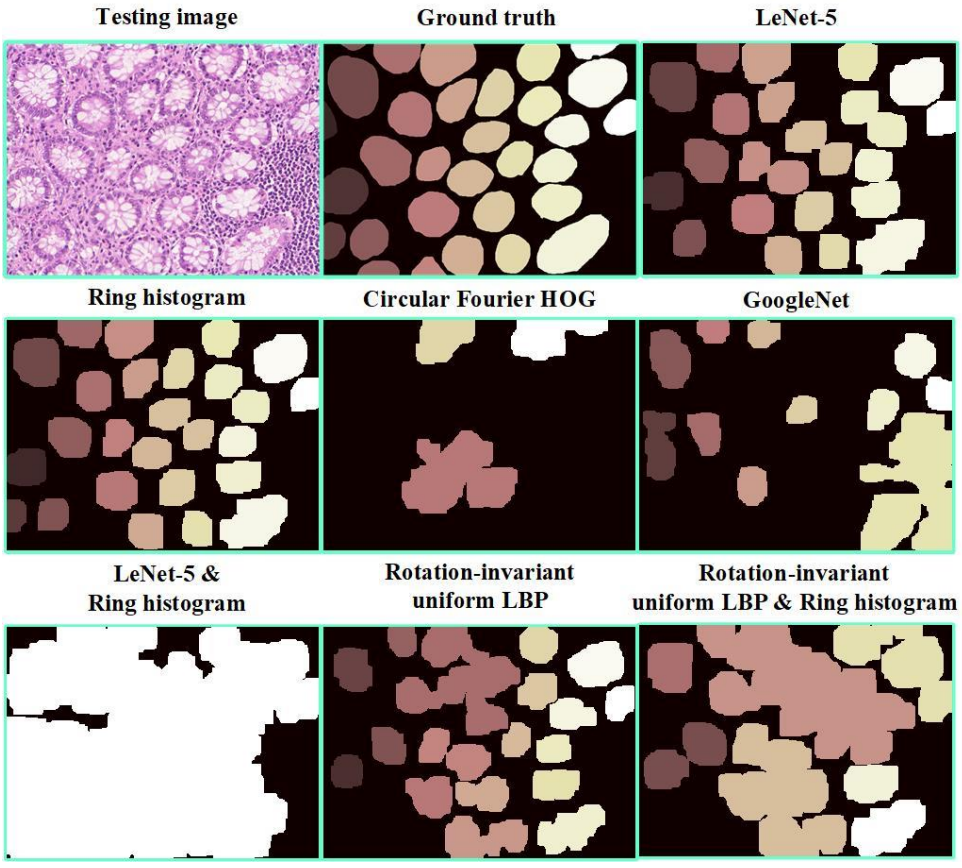


Figure E.2 Example results of benign category with three classes of different features in segmentation with pre-classification method.

E.3 Segmentation results for two classes of malignant category

Table E.3 shows the ranking of the segmentation results for different features of the two classes of malignant category gland images. The best performance is the ring histogram feature, highlighted in blue.

Table E.3 Ranking of segmentation results of different feature of two classes of malignant images

Different features	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	rank	
GoogleNet	0.62 (0.28)	2 (6)	0.5 (0.19)	3 (1)	0.67 (0.19)	2 (4)	0.54 (0.18)	1 (3)	162.86 (102.96)	1 (2)	299.8 (164.07)	1 (4)	10 (30)
Ring histogram	0.65 (0.27)	1 (5)	0.57 (0.25)	1 (5)	0.7 (0.19)	1 (4)	0.54 (0.20)	1 (5)	163.84 (155.22)	2 (4)	308.86 (147.30)	3 (2)	9 (34)
LeNet5	0.61 (0.29)	3 (7)	0.52 (0.21)	2 (2)	0.66 (0.21)	3 (6)	0.53 (0.15)	3 (1)	195.23 (176.80)	3 (7)	326.55 (129.59)	5 (1)	19 (43)
Ring histogram & RIULBP	0.31 (0.18)	7 (1)	0.46 (0.21)	5 (2)	0.27 (0.10)	7 (1)	0.48 (0.22)	4 (7)	522.69 (85.57)	7 (1)	305.16 (165.20)	2 (5)	32 (49)
RIULBP	0.42 (0.18)	5 (1)	0.34 (0.22)	7 (4)	0.52 (0.14)	5 (2)	0.44 (0.21)	7 (6)	226.77 (115.71)	4 (3)	309.37 (171.41)	4 (7)	32 (55)
LeNet5 & Ring histogram	0.52 (0.23)	4 (4)	0.43 (0.27)	6 (6)	0.6 (0.21)	4 (6)	0.46 (0.16)	6 (2)	249.85 (174.04)	5 (6)	371.56 (155.86)	6 (3)	31 (58)
CHOG	0.39 (0.21)	6 (3)	0.49 (0.32)	4 (7)	0.44 (0.18)	6 (3)	0.48 (0.18)	4 (3)	322.25 (150.50)	6 (5)	377.7 (170.21)	7 (6)	33 (60)

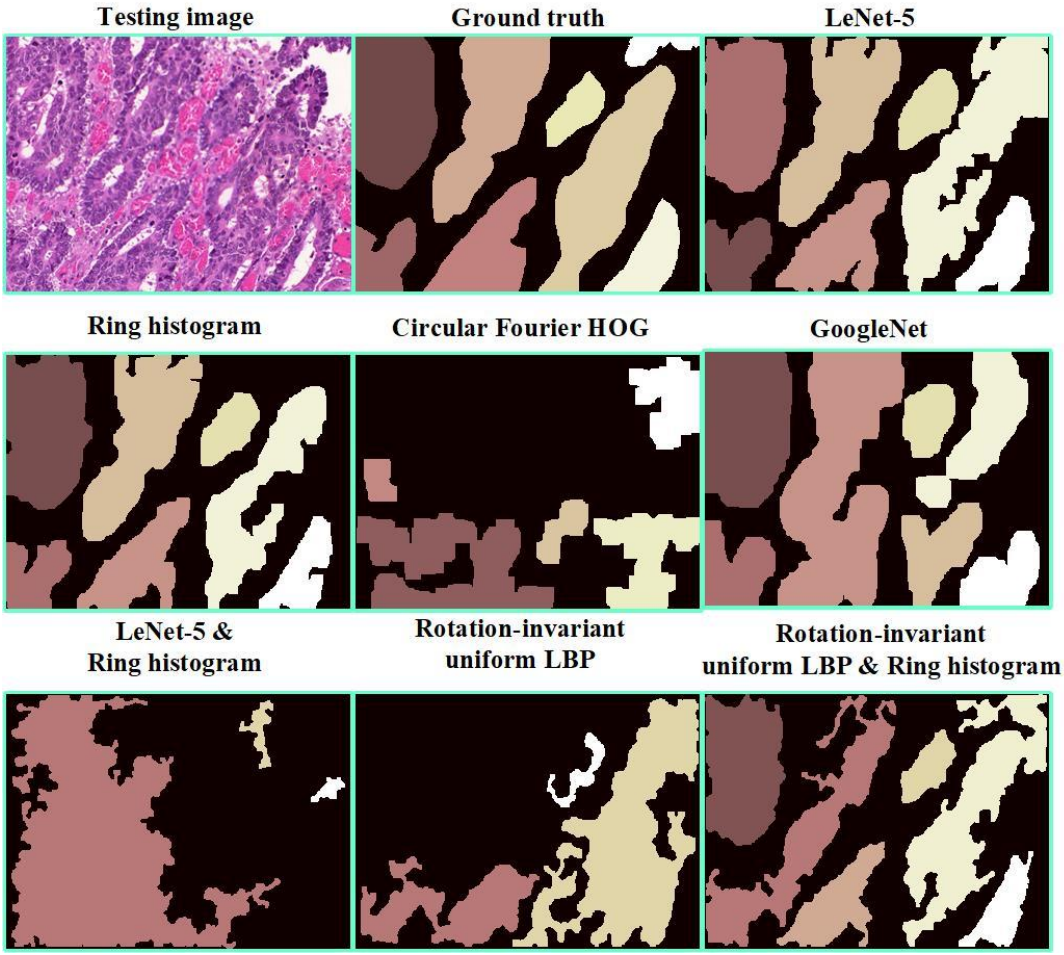


Figure E.3 Example results of malignant category with two classes of different features in segmentation with pre-classification method.

E.4 Segmentation results of three classes of malignant category

Table E.4 shows the ranking of the segmentation results of malignant category with three target classes.

Table E.4 Ranking of segmentation results of different features of three classes in the malignant category

Different features	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	rank	
LeNet5	0.57 (0.26)	1 (7)	0.58 (0.19)	1 (2)	0.58 (0.19)	2 (4)	0.52 (0.20)	1 (5)	215.69 (166.22)	2 (6)	319.33 (133.77)	2 (3)	9 (36)
Ring histogram	0.54 (0.23)	2 (5)	0.44 (0.26)	6 (3)	0.59 (0.17)	1 (3)	0.45 (0.23)	4 (6)	204.39 (147.75)	1 (4)	263.1 (171.23)	1 (7)	15 (43)
RIULBP	0.39 (0.18)	3 (2)	0.46 (0.32)	5 (4)	0.45 (0.19)	3 (4)	0.42 (0.19)	5 (3)	263.21 (146.93)	4 (3)	321.96 (123.49)	4 (2)	24 (42)
CHOG	0.39 (0.23)	3 (5)	0.49 (0.33)	3 (5)	0.39 (0.19)	4 (4)	0.47 (0.18)	3 (2)	393.65 (164.77)	6 (5)	371.86 (163.17)	7 (6)	26 (53)
Ring histogram & RIULBP	0.38 (0.20)	5 (4)	0.48 (0.33)	4 (5)	0.33 (0.19)	6 (4)	0.49 (0.19)	2 (3)	362.51 (167.02)	5 (7)	350.83 (158.05)	6 (5)	28 (56)
LeNet5 & Ring histogram	0.31 (0.18)	6 (2)	0.51 (0.41)	2 (7)	0.26 (0.09)	7 (1)	0.38 (0.23)	6 (6)	510.85 (81.63)	7 (1)	321.46 (107.66)	3 (1)	31 (49)
GoogleNet	0.28 (0.12)	7 (1)	0.33 (0.12)	7 (1)	0.38 (0.14)	5 (2)	0.3 (0.14)	7 (1)	257.89 (141.72)	3 (2)	344.36 (141.47)	5 (4)	34 (45)

Figure F.4 shows examples of results of different features of three classes of malignant category in segmentation with pre-classification.

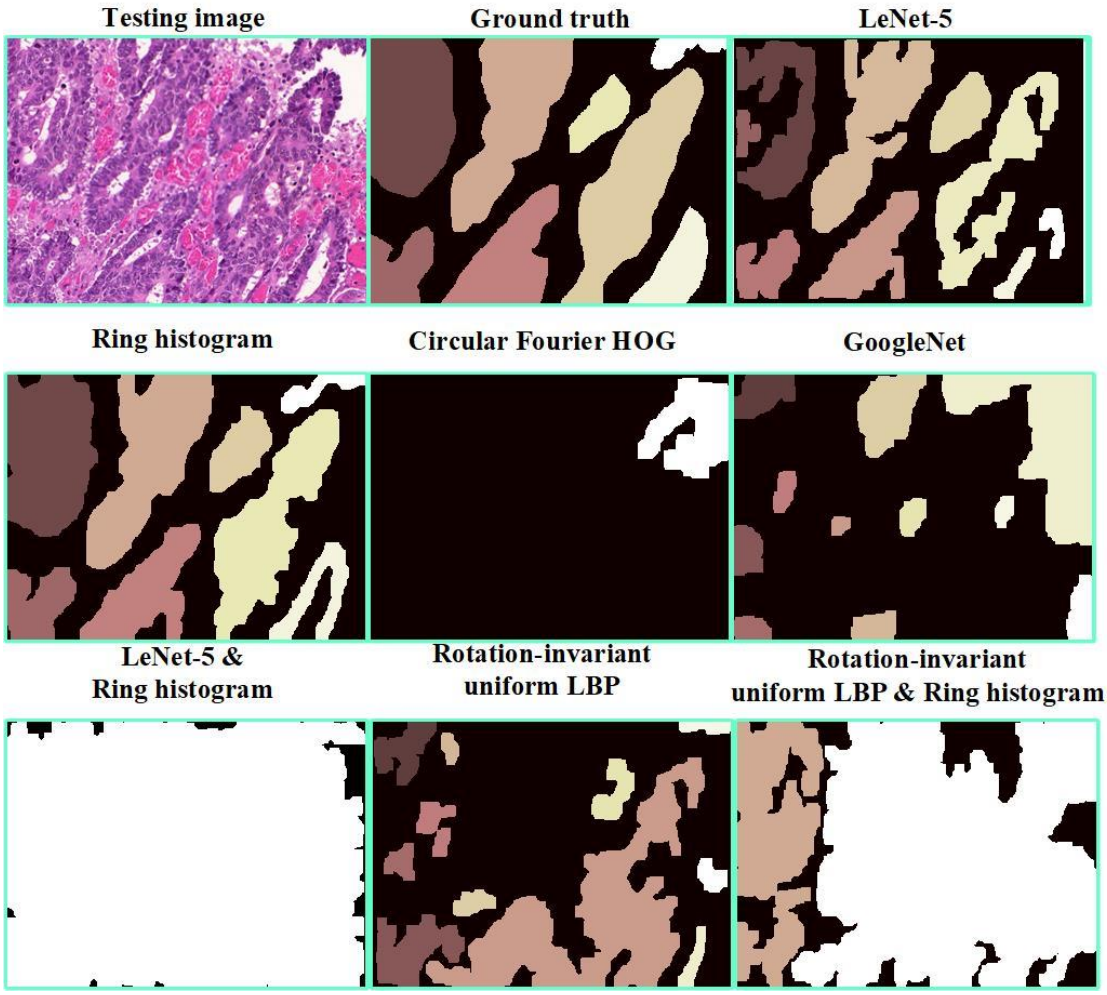


Figure E.4 Example of results of malignant category with two classes of different features in segmentation with pre-classification

E.5 Further improvement of the best performance of pixel-level classification in segmentation with the pre-classification method

F.5 shows examples of malignant images in the training part (on the left) compared with the corresponding images after applying local image deformation (on the right)

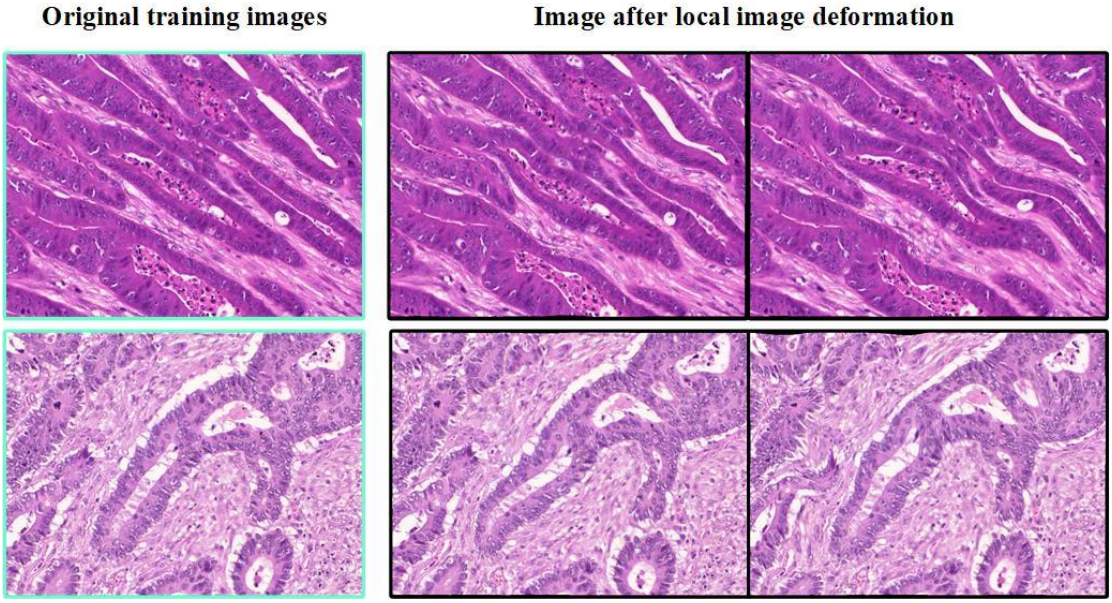


Figure E.5 Examples of training malignant images and the image after using local image deformation

Table E.5 Ranking of two classes of malignant category of the ring histogram feature with/without local image deformation

Using deformation	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
Yes	0.7 (0.26)	1(1)	0.52(0.18)	2(1)	0.71(0.19)	1(1)	0.53 (0.21)	2(2)	156.06 (154.06)	1 (1)	309.7 (161.66)	2 (2)	9 (17)
No	0.65(0.27)	2(2)	0.57 (0.25)	1(2)	0.7 (0.19)	2(1)	0.54 (0.2)	1(1)	163.84(155.22)	2(2)	308.86 (147.3)	1(1)	9 (18)

Even using the hand-crafted feature, the local image deformation could improve the segmentation performance.

Figure E.6 shows examples of benign images in training data (left) and the results after using local deformation (right).

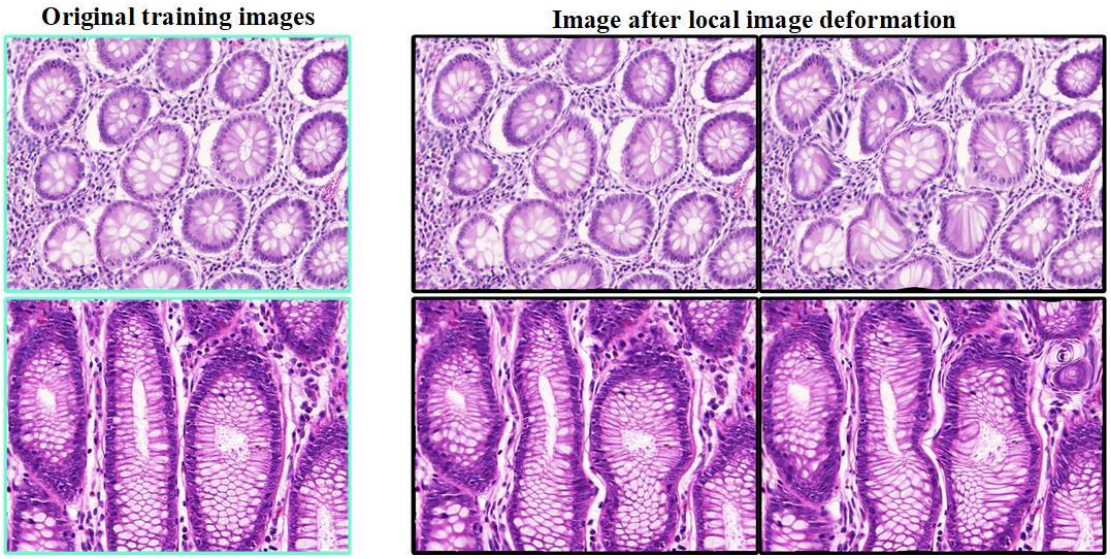


Figure E.6 Example of training benign images and the corresponding images after using local image deformation

The evaluation measure of three classes of benign category of ring histogram with/without local image deformation is shown in Table E.6.

Table E.6 Ranking of segmentation results of benign category three classes using histogram with/without local image deformation

Using deformation	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
Yes	0.64 (0.21)	2 (2)	0.43(0.21)	2(2)	0.67 (0.13)	2(2)	0.62(0.19)	2(1)	99.44(60.04)	2(1)	146.03 (48.02)	2(2)	12 (22)
No	0.74(0.11)	1(1)	0.59 (0.18)	1(1)	0.74(0.08)	1(1)	0.72(0.09)	1(1)	89.42 (72.23)	1(2)	138.38 (47.75)	1(1)	6 (13)

Using local image deformation, the patterns provided for training the classifier is more but the results is not improve. The reason might be that they are not providing the effective patterns to classifier the patterns in malignant tissue.

E.6 Comparison of intensity-based and gradient-based histograms

Tables E.7 and E.8 show the ranking of segmentation performance of benign category glands with two and three class labels respectively.

Table E.7 Segmentation results for two types of histogram with three classes of benign category images

Histogram types	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
Intensity-based	0.74(0.11)	1(1)	0.59(0.18)	1(2)	0.74(0.08)	1(1)	0.72(0.08)	1(1)	89.42(72.23)	1(1)	138.38 (47.75)	1(1)	6 (13)
Gradient-based	0.56 (0.21)	2(2)	0.25(0.22)	2(1)	0.61 (0.19)	2(2)	0.48(0.26)	2(2)	131.87(88.76)	2(2)	242.04(82.23)	2(2)	12 (23)

Table E.8 Segmentation results for two types of histogram with two classes of benign category images

Histogram types	F1 score				Object-level Dice index				Object-level Hausdorff distance				Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
Intensity-based	0.56 (0.2)	1(2)	0.78 (0.03)	1(1)	0.61 (0.15)	1 (1)	0.81 (0.12)	1(1)	161.5 (122.67)	1(1)	136.88 (91.81)	1(2)	6 (14)
Gradient-based	0.46 (0.19)	2(1)	0.63(0.13)	2(2)	0.52(0.15)	2(1)	0.72(0.15)	2(2)	204.89 (135.61)	2(2)	151.5 (73.18)	2(1)	12 (21)

Tables E.9 and E.10 show the segmentation results with two and three classes respectively for malignant images, with the best again highlighted.

Table E.9 Segmentation results for two types of histogram for two classes of malignant category images

Histogram types	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
Intensity-based	0.64 (0.27)	1(2)	0.57(0.25)	1(2)	0.70(0.19)	1(2)	0.54(0.2)	1(2)	163.84(155.22)	1(2)	308.86 (147.3)	1(2)	6 (18)
Gradient-based	0.49 (0.2)	2(1)	0.52 (0.19)	2(1)	0.53 (0.14)	2(1)	0.55(0.15)	2(1)	228.46(116.86)	2(1)	343.15 (124.34)	2(1)	12 (18)

Table E.10 Segmentation results for two types of histogram for two classes of malignant category images

Histogram types	F1 score				Object-level Dice index				Object-level Hausdorff distance				Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
Intensity-based	0.54 (0.23)	1(2)	0.44(0.26)	2(1)	0.59(0.17)	1(1)	0.45(0.23)	2(2)	204.39 (147.75)	1(2)	263.11(171.28)	1(1)	8 (17)
Gradient-based	0.3 (0.14)	2 (1)	0.49 (0.29)	1(2)	0.2(0.19)	2(2)	0.50(0.17)	1(1)	245.48(96.61)	2(1)	313.92(152.79)	2(2)	10 (19)

E.7 Comparison of results of with/without histogram correction

Table E.11 therefore indicates the results of histogram features with/without histogram correction in segmentation without pre-classification method.

Table E.11 Segmentation results of histogram feature with/without histogram correction in segmentation without pre-classification

Using colour correction	F1 score				Object-level Dice index				Object-level Hausdorff distance				Rank Sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	Score	rank	Score	Rank	Score	rank	Score	rank	Score	rank	Score	Rank	
Yes	0.51 (0.24)	1(1)	0.5(0.22)	2(1)	0.54(0.2)	2(2)	0.62 (0.17)	1(1)	267.88(181.94)	2(2)	266.43(125.82)	2(2)	10 (20)
No	0.51 (0.25)	1(2)	0.57 (0.23)	1(2)	0.58 (0.18)	1(1)	0.63(0.17)	2(1)	185.7 (156.99)	1(1)	263.31 (125.57)	1(1)	7 (14)

Appendix F

Ranking of different methods in gland segmentation

Table F.1 shows the ranking for all the methods that participated in the MICCAI 2015 Gland Segmentation Challenge. This table does include three methods introduced in this work indicated by the red bounding box. Two proposed method achieved the middle ranking.

Table F.1 The ranking of different methods in Gland Segmentation Challenge (Warwick.ac.uk, 2018b)

Method	Ranking						Sum
	F1 score A	F1 score B	Dice A	Dice B	Hausdorff A	Hausdorff B	
CUMedVision2	1	3	1	5	1	6	17
ExB1	4	4	4	2	6	1	21
ExB3	2	2	2	6	5	5	22
Freiburg2	5	6	5	3	3	3	24
CUMedVision1	6	1	8	1	8	4	28
ExB2	3	7	3	7	2	8	29
Freiburg1	8	9	6	4	4	2	32
CVIP	7	8	7	8	7	10	46
Method 2	11	5	11	9	12	13	61
Method 1	10	10	10	11	11	12	64
CVML	12	11	13	10	13	7	66
LIB	9	19	9	14	9	9	69
Vision4GlaS	13	12	12	16	10	11	74
LIST	15	13	16	12	16	14	86
Ching-Wei Wang 1	14	14	17	15	18	16	94
Biomage Informatics	18	17	19	12	20	14	100
Ching-Wei Wang 2	16	15	18	17	19	19	104
SUTECH	17	20	15	20	15	17	104
ISI Kolkatta	20	18	20	21	14	15	108
FIMM	21	21	14	19	17	21	113
Ching-Wei Wang 3	19	16	21	18	21	20	115

Image Analysis Lab Freiburg: Freiburg 2 = post-processing, Freiburg 1 = raw

CVIP Dundee: Feature level fusion

Ching-Wei Wang: Ching-Wei Wang 1 = no preprocess fill hole, Ching-Wei Wang 2 = no preprocess hole

Ching-Wei Wang 3 = preprocess fill hole

Method 1: Segmentation method without pre-classification

Method 2: Segmentation method with pre-classification at the pixel-level classifier

The two proposed methods introduced in this work used the random forest as the pixel-level classifier to predict the classes of each pixel in testing images. The two proposed methods not only achieved better results than any other method not using deep classifier, but also outperformed same methods using deep learning.

Appendix G

Publications

1. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U. and Böhm, A., 2017. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, *35*, pp.489-502.

2. Wang L., Zhou Yu, Matuszewski B.J. (2019) A New Hybrid Method for Gland Segmentation in Histology images. In: Vento M. et al. (eds) Computer Analysis of Images and Patterns. CAIP 2019. Communications in Computer and information Science, vol 1089, Springer, Cham.

Appendix H

Screenshot of image-level classification results

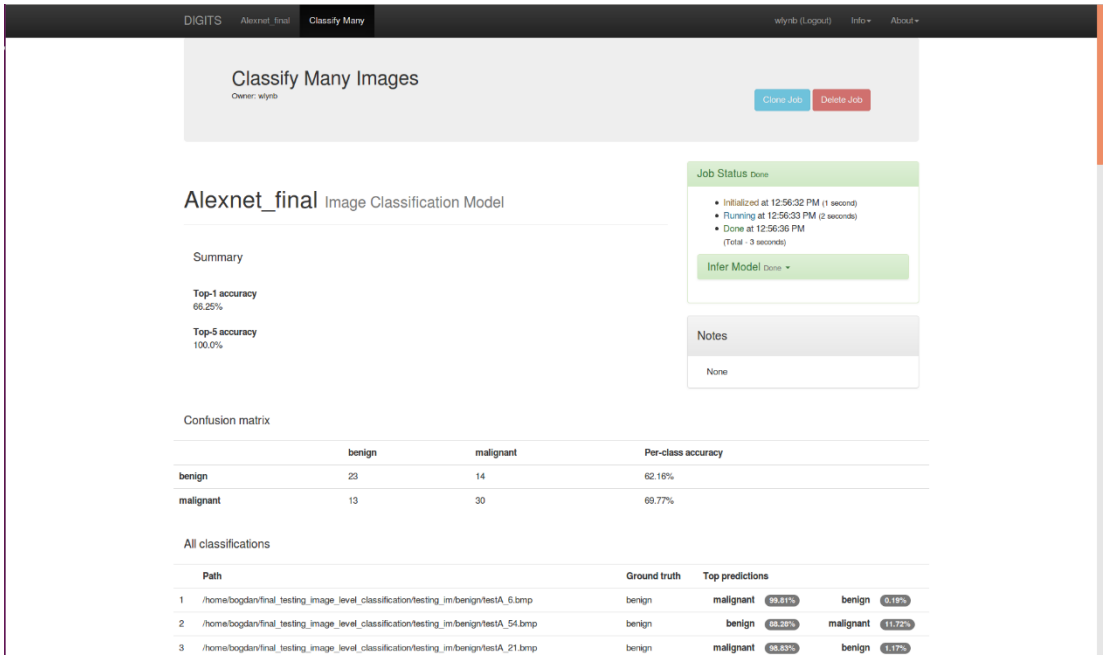


Figure H.1 Screenshot for Image-level classification results using AlexNet In DIGITS

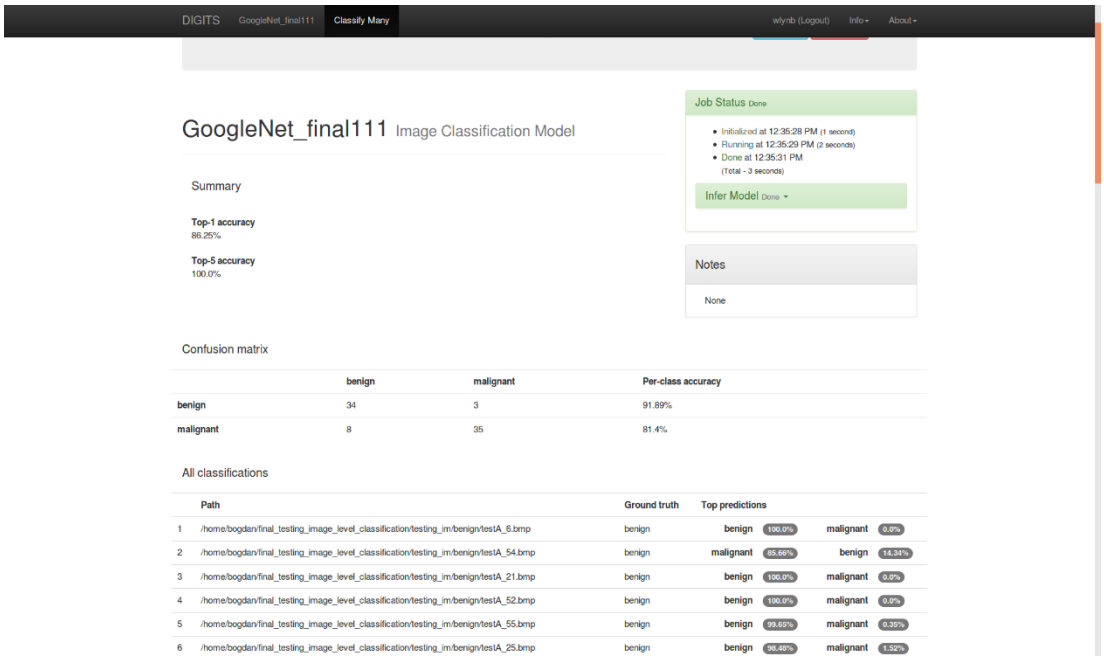


Figure H.2 Screenshot for image-level classification results using GoogleNe In DIGITS

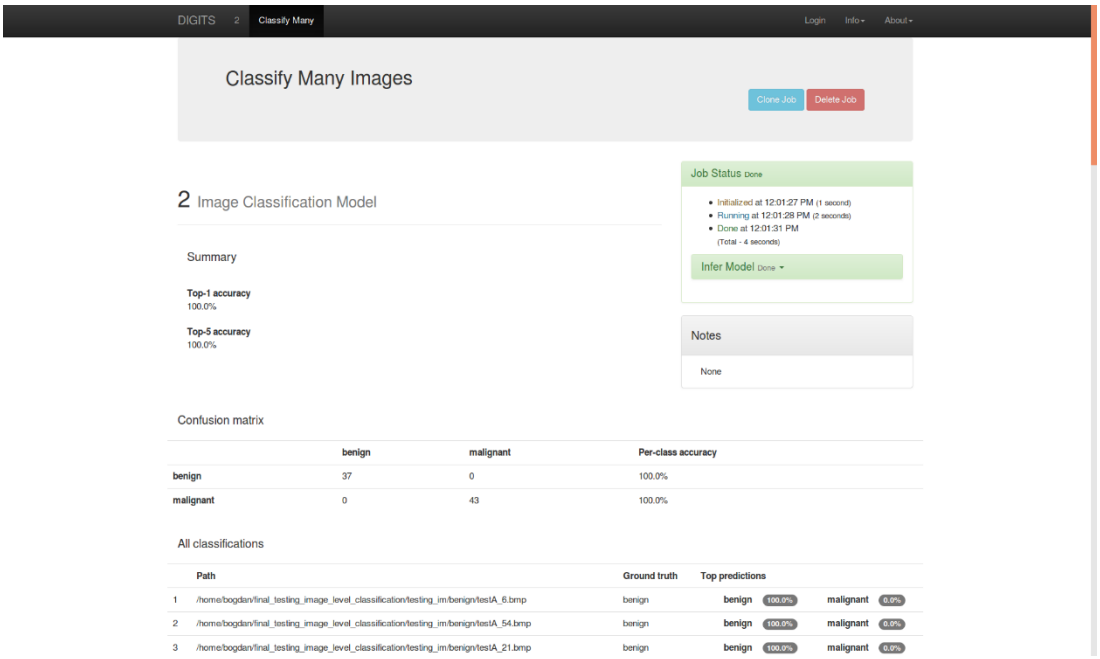


Figure H.3 Screenshot for image-level classification results using ResNet50 in DIGITS

References:

Adams, R. and Bischof, L., 1994. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, *16*(6), pp.641-647.

Akbar, B., Gopi, V.P. and Babu, V.S., 2015, February. Colon cancer detection based on structural and statistical pattern recognition. In *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on* (pp. 1735-1739). IEEE.

Altunbay, D., Cigir, C., Sokmensuer, C. and Gunduz-Demir, C., 2010. Colour graphs for automated cancer diagnosis and grading. *IEEE Transactions on Biomedical Engineering*, *57*(3), pp.665-674.

Ap, R., Khan, S.S., Anubhav, K. and Paul, A., 2017. Gland Segmentation in Histopathology Images Using Random Forest Guided Boundary Construction. *arXiv preprint arXiv:1705.04924*.

Arifin, A.Z. and Asano, A., 2006. Image segmentation by histogram thresholding using hierarchical cluster analysis. *Pattern Recognition Letters*, *27*(13), pp.1515-1521.

Ballard, D.H., 1981. Generalizing the Hough transform to detect arbitrary shapes. *Pattern recognition*, *13*(2), pp.111-122.

Barandiaran, I., 1998. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8).

Beauchemin, M., Thomson, K.P. and Edwards, G., 1998. On the Hausdorff distance used for the evaluation of segmentation results. *Canadian journal of remote sensing*, *24*(1), pp.3-8.

Belaid, L.J. and Mourou, W., 2011. Image segmentation: a watershed transformation algorithm. *Image Analysis & Stereology*, *28*(2), pp.93-102.

Bezdek, J.C., Ehrlich, R. and Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), pp.191-203.

Breiman, L., 1996. *ftp. stat. berkeley. edu/pub/users/breiman/OOBestimation. ps*.

Breiman, L., 2001. Random forests. *Machine Learning*, *45*(1), pp.5-32.

Breiman, L., 2017. *Classification and regression trees*. New York: Routledge.

Brejl, M. and Sonka, M., 2000. Automated initialization and automated design of border detection criteria in edge-based image segmentation. In *Image Analysis and Interpretation, 2000. Proceedings. 4th IEEE Southwest Symposium* (pp. 26-30). IEEE.

Brodley, C.E. and Utgoff, P.E., 1992. *Multivariate versus univariate decision trees*. Amherst, MA: University of Massachusetts, Department of Computer and Information Science.

Canny, J., *A Computational Approach To Edge Detection*, IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679–698, 1986.

Caselles, V., Kimmel, R. and Sapiro, G., 1997. Geodesic active contours. *International journal of computer vision*, *22*(1), pp.61-79.

Chan, T.F. and Vese, L.A., 2001. Active Contours Without Edges. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, *10*(2).

Chen, C., Liu, M., Liu, H., Zhang, B., Han, J. and Kehtarnavaz, N., 2017. Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition. *IEEE Access*, *5*, pp.22590-22604.

Chen, H., Qi, X., Yu, L. and Heng, P.A., 2016. DCAN: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 2487-2496).

Cheriet, M., Said, J.N. and Suen, C.Y., 1998. A recursive thresholding technique for image segmentation. *IEEE transactions on image processing*, *7*(6), pp.918-921.

Christ, P.F., Ettlinger, F., Grün, F., Elshaera, M.E.A., Lipkova, J., Schlecht, S., Ahmaddy, F., Tatavarty, S., Bickel, M., Bilic, P. and Rempfler, M., 2017. Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks. *arXiv preprint arXiv:1702.05970*.

Cohen, A., Rivlin, E., Shimshoni, I. and Sabo, E., 2015. Memory-based active contour algorithm using pixel-level classified images for colon crypt segmentation. *Computerized Medical Imaging and Graphics*, *43*, pp.150-164.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, *20*(3), pp.273-297.

Costantini, M., Sciallero, S., Giannini, A., Gatteschi, B., Rinaldi, P., Lanzanova, G., Bonelli, L., Casetti, T., Bertinelli, E., Giuliani, O. and Castiglione, G., 2003. Interobserver agreement in the histologic diagnosis of colorectal polyps: the experience of the multicenter adenoma colorectal study (SMAC). *Journal of clinical epidemiology*, *56*(3), pp.209-214.

Csurka, G., Larlus, D., Perronnin, F. and Meylan, F., 2013, September. What is a good evaluation measure for semantic segmentation?. In *BMVC* (Vol. 27, p. 2013).

Dalal, N. and Triggs, B., 2005, June. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 1, pp. 886-893). IEEE.

Davies, R.B., 2002. Exclusive OR (XOR) and hardware random number generators. *HYPERLINK" http://www. robertnz. net/pdf/xor2. pdf" http://www. robertnz. net/pdf/xor2. pdf*.

Dehzangi, A., Phon-Amnuaisuk, S. and Dehzangi, O., 2010. Using random forest for protein fold prediction problem: an empirical study. *Journal of Information Science and Engineering*, *26*(6), pp.1941-1956.

Diamond, J., Anderson, N.H., Bartels, P.H., Montironi, R. and Hamilton, P.W., 2004. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human pathology*, *35*(9), pp.1121-1131.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. *Ecology*, *26*(3), pp.297-302.

Doyle, S., Agner, S., Madabhushi, A., Feldman, M. and Tomaszewski, J., 2008, May. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on* (pp. 496-499). IEEE.

Du Buf, JM Hans, M. Kardan, and Michael Spann. "Texture feature performance for image segmentation." *Pattern recognition* 23, no. 3-4 (1990): 291-309.

Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [online] Available at: http://archive.ics.uci.edu/ml [Accessed 17 Oct 2017]

Efron, B., 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics* (pp. 569-593). Springer, New York, NY.

Eleyan, A. and Demirel, H., 2011. Co-occurrence matrix and its statistical features as a new approach for face recognition. *Turkish Journal of Electrical Engineering & Computer Sciences*, *19*(1), pp.97-107.

Everingham, M. and Winn, J., 2011. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*.

Farjam, R., Soltanian-Zadeh, H., Jafari-Khouzani, K. and Zoroofi, R.A., 2007. An image analysis approach for automatic malignancy determination of prostate pathological images. *Cytometry Part B: Clinical Cytometry: The Journal of the International Society for Analytical Cytology*, *72*(4), pp.227-240.

Fernández, C., Huerta, I. and Prati, A., 2015. A comparative evaluation of regression learning algorithms for facial age estimation. In *Face and facial expression recognition from real world videos* (pp. 133-144). Springer, Heidelberg.

Fernandez-Moral, E., Martins, R., Wolf, D. and Rives, P., 2018, June. A new metric for evaluating semantic segmentation: leveraging global and contour accuracy. In *2018 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1051-1056). IEEE.

Fleming, M., Ravula, S., Tatishchev, S.F. and Wang, H.L., 2012. Colorectal carcinoma: pathologic aspects. *Journal of gastrointestinal oncology*, *3*(3), p.153.

Freund, Y. and Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), pp.119-139.

Freund, Y., Iyer, R., Schapire, R.E. and Singer, Y., 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, *4*(Nov), pp.933-969.

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, *32*(200), pp.675-701.

Fu, H., Qiu, G., Shu, J. and Ilyas, M., 2014. A novel polar space random field model for the detection of glandular structures. *IEEE transactions on medical imaging*, *33*(3), pp.764-776.

Geurts, P., Ernst, D. and Wehenkel, L., 2006. Extremely randomized trees. *Machine learning*, *63*(1), pp.3-42.

Gini, C., 1912. Memorie di metodologia statistica. Variabilità e Concentrazione, vol. 1.

González, A., Villalonga, G., Xu, J., Vázquez, D., Amores, J. and López, A.M., 2015. Multiview random forest of local experts combining rgb and lidar data for pedestrian detection. In *Intelligent Vehicles Symposium (IV), 2015 IEEE* (pp. 356-361). IEEE.

Gonzalez, R. and Woods, R. (2002). Digital Image Processing. Prentice Hall, second edition

Grady, L. and Schwartz, E.L., 2006. Isoperimetric graph partitioning for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3), pp.469-475.

Grady, L., 2006. Random walks for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *28*(11), pp.1768-1783.

Graham, S., Chen, H., Dou, Q., Heng, P.A. and Rajpoot, N., 2018. MILD-Net: Minimal Information Loss Dilated Network for Gland Instance Segmentation in Colon Histology Images. *arXiv preprint arXiv:1806.01963*.

Gunduz-Demir, C., Kandemir, M., Tosun, A.B. and Sokmensuer, C., 2010. Automatic segmentation of colon glands using object-graphs. *Medical image analysis*, *14*(1), pp.1-12.

Gurcan, M.N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N. and Yener, B., 2009. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, *2*, p.147.

Hammouche, K., Diaf, M. and Siarry, P., 2008. A multilevel automatic thresholding method based on a genetic algorithm for a fast image segmentation. *Computer Vision and Image Understanding*, *109*(2), pp.163-175.

Haralick, R.M. and Shanmugam, K., 1973. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), pp.610-621.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Heath, D., Kasif, S., and Salzberg, S., 1993b. Learning oblique decision trees. In Proceedings of the 13th International Joint Conference on Articial Intelligence, pp. 1002-1007. Chambery, France. Morgan Kaufmann

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017, July. Densely connected convolutional networks. In *CVPR* (Vol. 1, No. 2, p. 3).

Hum, Y.C., Lai, K.W. and Mohamad Salim, M.I., 2014. Multiobjectives bihistogram equalization for image contrast enhancement. *Complexity*, *20*(2), pp.22-36.

Irshad, H., Jalali, S., Roux, L., Racoceanu, D., Hwee, L.J., Le Naour, G. and Capron, F., 2013. Automated mitosis detection using texture, SIFT features and HMAX biologically inspired approach. *Journal of pathology informatics*, *4*(Supplement).

Jagtap, S.B., 2013. Census data mining and data analysis using WEKA. *arXiv preprint arXiv:1310.4647*.

Jampour, M., Moin, M.S., Yu, L.F. and Bischof, H., 2018. Mapping forests: a comprehensive approach for nonlinear mapping problems. *Journal of Mathematical Imaging and Vision*, *60*(2), pp.232-245.

Kainz, P., Pfeiffer, M. and Urschler, M., 2015. Semantic segmentation of colon glands with deep convolutional neural networks and total variation segmentation. *arXiv preprint arXiv:1511.06919*.

Kass, M., Witkin, A. and Terzopoulos, D., 1988. Snakes: Active contour models. *International Journal of Computer Vision*, *1*(4), 321-331.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D. and Jansen, L., 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, *16*(1), p.e1002730.

Khan, R., Hanbury, A. and Stoettinger, J., 2010. Skin detection: a random forest approach. In *Image Processing (ICIP), 2010 17th IEEE International Conference on* (pp. 4613-4616). IEEE.

Khan, S., Islam, N., Jan, Z., Din, I.U. and Rodrigues, J.J.C., 2019. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, *125*, pp.1-6.

Kontschieder, P., Fiterau, M., Criminisi, A. and Rota Bulo, S., 2015. Deep neural decision forests. In *Proceedings of the IEEE international conference on computer vision* (pp. 1467-1475).

Kremic, E. and Subasi, A., 2016. Performance of random forest and SVM in face recognition. *International Arab Journal of Information Technology*, *13*(2), pp.287-293.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Latinne, P., Debeir, O. and Decaestecker, C., 2001, July. Limiting the number of trees in random forests. In *International workshop on multiple classifier systems* (pp. 178-187). Springer, Berlin, Heidelberg.

Lahdenoja, O., Poikonen, J. and Laiho, M., 2013. Towards understanding the formation of uniform local binary patterns. *ISRN Machine Vision*, *2013*.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), pp.2278-2324.

Li, W., Manivannan, S., Akbar, S., Zhang, J., Trucco, E. and McKenna, S.J., 2016, April. Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on* (pp. 1405-1408). IEEE.

Lloyd, S., 1982. Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), pp.129-137.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Maclin, R. and Opitz, D., 1997. An empirical evaluation of bagging and boosting. *AAAI/IAAI*, *1997*, pp.546-551.

Manivannan, S., Li, W., Zhang, J., Trucco, E. and McKenna, S.J., 2018. Structure Prediction for Gland Segmentation With Hand-Crafted and Deep Convolutional Features. *IEEE transactions on medical imaging*, *37*(1), pp.210-221.

McCann, M.T., Ozolek, J.A., Castro, C.A., Parvin, B. and Kovacevic, J., 2015. Automated histology analysis: Opportunities for signal processing. *IEEE Signal Processing Magazine*, *32*(1), pp.78-87.

Mehta, R. and Egiazarian, K.O., 2013, February. Rotated Local Binary Pattern (RLBP)-Rotation Invariant Texture Descriptor. In *ICPRAM* (pp. 497-502).

Menze, B.H., Kelm, B.M., Splitthoff, D.N., Koethe, U. and Hamprecht, F.A., 2011, September. On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 453-469). Springer, Berlin, Heidelberg.

Morar, A., Moldoveanu, F. and Gröller, E., 2012, August. Image segmentation based on active contours without edges. In *2012 IEEE 8th International Conference on Intelligent Computer Communication and Processing* (pp. 213-220). IEEE.

Murthy, S.K., Kasif, S. and Salzberg, S., 1994. A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, *2*, pp.1-32.

Nasir, (2015). GlaS@MICCAI'2015: Gland Segmentation Challenge Contest. [online] Available at: https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest/ [Accessed 17 Dec. 2017]

Ng, H.P., Ong, S.H., Foong, K.W.C., Goh, P.S. and Nowinski, W.L., 2006, March. Medical image segmentation using k-means clustering and improved watershed algorithm. In *Image Analysis and Interpretation, 2006 IEEE Southwest Symposium on* (pp. 61-65). IEEE.

Nguyen, K., Jain, A.K. and Allen, R.L., 2010, August. Automated gland segmentation and classification for Gleason grading of prostate tissue images. In *Pattern Recognition (ICPR), 2010 20th International Conference on* (pp. 1497-1500). IEEE.

Nicola Parry. 2017. *How Histology Slides are Prepared*. [online] Available at: https://bitesizebio.com/13398/how-histology-slides-are-prepared/. [Accessed 19 December 2018].

Ojala, T., c, M. and Mäenpää, T., 2000, June. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision* (pp. 404-420). Springer, Berlin, Heidelberg.

Ojala, T., Pietikäinen, M. and Harwood, D., 1994, October. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on* (Vol. 1, pp. 582-585). IEEE.

Ojala, T., Pietikäinen, M. and Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(7), pp.971-987.

Pal, N.R. and Pal, S.K., 1993. A review on image segmentation techniques. *Pattern recognition*, *26*(9), pp.1277-1294.

Paul, A. and Mukherjee, D.P., 2016, September. Gland segmentation from histology images using informative morphological scale space. In *Image Processing (ICIP), 2016 IEEE International Conference on* (pp. 4121-4125). IEEE.

Peizhuang, W., 1983. Pattern recognition with fuzzy objective function algorithms (James C. Bezdek). *SIAM Review*, *25*(3), p.442.

Pietikäinen, M., Hadid, A., Zhao, G. and Ahonen, T., 2011. Local binary patterns for still images. In *Computer vision using local binary patterns* (pp. 13-47). Springer, London.

Probst, P. and Boulesteix, A.L., 2017. To Tune or Not to Tune the Number of Trees in Random Forest. *Journal of Machine Learning Research*, *18*, pp.181-1.

Powell, S., Magnotta, V.A., Johnson, H., Jammalamadaka, V.K., Pierson, R. and Andreasen, N.C., 2008. Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. *Neuroimage*, *39*(1), pp.238-247.

Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Quinlan, J.R., 1986. Induction of decision trees. *Machine learning*, *1*(1), pp.81-106.

Quinlan, J.R., 1993. C4. 5: Programming for machine learning. *Morgan Kauffmann*, *38*, p.48.

Rathore, S., Iftikhar, M.A., Hussain, M. and Jalil, A., 2013, December. A novel approach for ensemble clustering of colon biopsy images. In *Frontiers of Information Technology (FIT), 2013 11th International Conference on* (pp. 25-30). IEEE.

Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., Sudhakar, P. and Vaidya, V., 2017, September. Learning and incorporating shape models for semantic segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 203-211). Springer, Cham.

Ren, X. and Malik, J., 2003, October. Learning a classification model for segmentation. In *null* (p. 10). IEEE.

Rodriguez, J.J., Kuncheva, L.I. and Alonso, C.J., 2006. Rotation forest: a new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), pp.1619-1630.

Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

Ross, R. and Kelleher, J., 2013, December. A comparative study of the effect of sensor noise on activity recognition models. In *International Joint Conference on Ambient Intelligence* (pp. 151-162). Springer, Cham.

Schneiderman, H. and Kanade, T., 2000. A histogram-based method for detection of faces and cars. In *Image Processing, 2000. Proceedings. 2000 International Conference on* (Vol. 3, pp. 504-507). IEEE.

Schroff, F., Criminisi, A. and Zisserman, A., 2008, September. Object Class Segmentation using Random Forests. In *BMVC* (pp. 1-10).

Sergyan, S., 2008, January. Colour histogram features based image classification in content-based image retrieval systems. In *Applied Machine Intelligence and Informatics, 2008. SAMI 2008. 6th International Symposium on* (pp. 221-224). IEEE.

Shafarenko, L., Petrou, M. and Kittler, J., 1997. Automatic watershed segmentation of randomly textured color images. *IEEE transactions on Image Processing*, *6*(11), pp.1530-1544.

Shi, J. and Malik, J., 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(8), pp.888-905.

Singla, A., Yuan, L. and Ebrahimi, T., 2016, October. Food/non-food image classification and food categorization using pre-trained googlenet model. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management* (pp. 3-11). ACM.

Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U. and Böhm, A., 2017. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, *35*, pp.489-502.

Sirinukunwattana, K., Snead, D.R. and Rajpoot, N.M., 2015. A stochastic polygons model for glandular structures in colon histology images. *IEEE transactions on medical imaging*, *34*(11), pp.2366-2378.

Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U. and Böhm, A., 2017. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, *35*, pp.489-502.

Skibbe, H. and Reisert, M., 2012, May. Circular Fourier-hog features for rotation invariant object detection in biomedical images. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on* (pp. 450-453). IEEE.

Sumengen, B. and Manjunath, B.S., 2005, September. Multi-scale edge detection and image segmentation. In *Signal Processing Conference, 2005 13th European* (pp. 1-4). IEEE.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

Tan, K.S. and Isa, N.A.M., 2011. Color image segmentation using histogram thresholding–Fuzzy C-means hybrid approach. *Pattern Recognition*, *44*(1), pp.1-15.

Tan, X. and Triggs, B., 2010. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, *19*(6), pp.1635-1650.

Theodoridis, S., Theodoridis, S. and Koutroumbas, K., 2014. *Pattern Recognition*. Burlington: Elsevier Science.

Theriault, C., Thome, N. and Cord, M., 2013. Extended coding and pooling in the hmax model. *IEEE Transactions on Image Processing*, *22*(2), pp.764-777.

Tobias, O.J. and Seara, R., 2002. Image segmentation by histogram thresholding using fuzzy sets. *IEEE transactions on Image Processing*, *11*(12), pp.1457-1465.

Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J. and Jemal, A., 2015. Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, *65*(2), pp.87-108.

Van Den Boomgaard, R. and Van Balen, R., 1992. Methods for fast morphological image transforms using bitmapped binary images. *CVGIP: Graphical Models and Image Processing*, *54*(3), pp.252-258.

Van Putten, P.G., Hol, L., Van Dekken, H., Han van Krieken, J., Van Ballegooijen, M., Kuipers, E.J. and Van Leerdam, M.E., 2011. Inter-observer variation in the histological diagnosis of polyps in colorectal cancer screening. *Histopathology*, *58*(6), pp.974-981.

Veta, M., Van Diest, P.J., Willems, S.M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A.B., Vestergaard, J.S., Dahl, A.B. and Cireşan, D.C., 2015. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical image analysis*, *20*(1), pp.237-248.

Vincent, O.R. and Folorunso, O., 2009, June. A descriptive algorithm for sobel image edge detection. In *Proceedings of Informing Science & IT Education Conference (InSITE)* (Vol. 40, pp. 97-107). California: Informing Science Institute.

Wang, D., 1997. A multiscale gradient algorithm for image segmentation using watershelds. *Pattern recognition*, *30*(12), pp.2043-2052.

Wang, X., Han, T.X. and Yan, S., 2009. An HOG-LBP human detector with partial occlusion handling. In *IEEE 12th International Conference on Computer Vision, Results* (pp. 32-39). Los Alamitos CA: IEEE.

Wang, X.Y., Wang, T. and Bu, J., 2011. Color image segmentation using pixel wise support vector machine classification. *Pattern Recognition*, *44*(4), pp.777-787.

Warwick.ac.uk. (2016a). *Results*. [online] Available at: https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest/results/ [Accessed 12 December 2017]

Warwick.ac.uk. (2016b). *Download*. [online] Available at: https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest/download/ [Accessed 12 December 2017]

Wu, H.S., Xu, R., Harpaz, N., Burstein, D. and Gil, J., 2005a. Segmentation of microscopic images of small intestinal glands with directional 2-d filters. *Analytical and quantitative cytology and histology*, *27*(5), pp.291-300.

Wu, H.S., Xu, R., Harpaz, N., Burstein, D. and Gil, J., 2005b. Segmentation of intestinal gland images with iterative region growing. *Journal of Microscopy*, *220*(3), pp.190-204.

Wu, Z. and Leahy, R., 1993. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *15*(11), pp.1101-1113.

Xiaoling, W., 2009, March. A novel circular ring histogram for content-based image retrieval. In *Education Technology and Computer Science, 2009. ETCS'09. First International Workshop on* (Vol. 2, pp. 785-788). IEEE.

Xu, Y., Li, Y., Liu, M., Wang, Y., Lai, M., Eric, I. and Chang, C., 2016, October. Gland instance segmentation by deep multichannel side supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 496-504). Springer, Cham.

Yang, L., Zhang, Y., Chen, J., Zhang, S. and Chen, D.Z., 2017, September. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 399-407). Springer, Cham.

Yang, W., Guo, L., Zhao, T. and Xiao, G., 2007, May. Improving watersheds image segmentation method with graph theory. In *Industrial Electronics and Applications, 2007. ICIEA 2007. 2nd IEEE Conference on* (pp. 2550-2553). IEEE.

Yuheng, S. and Hao, Y., 2017. Image Segmentation Algorithms Overview. *arXiv preprint arXiv:1707.02051*.

Zahn, C.T., 1970. Graph theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.*, *20*(SLAC-PUB-0672-REV), p.68.

Zhang, Y., Matuszewski, B.J., Shark, L.K. and Moore, C.J., 2008, July. Medical image segmentation using new hybrid level-set method. In *Fifth International Conference BioMedical Visualization: Information Visualization in Medical and Biomedical Informatics* (pp. 71-76). IEEE.

Zhao, G., Ahonen, T., Matas, J. and Pietikainen, M., 2012. Rotation-invariant image and video description with local binary pattern features. IEEE Transactions on Image Processing, 21(4), pp.1465-1477.

Zhu, R., Sang, G., Cai, Y., You, J. and Zhao, Q., 2013. Head pose estimation with improved random regression forests. In *Biometric Recognition* (pp. 457-465). Springer, Heidelberg.