



# UNIVERSITY OF LEEDS

## **A Deep-learning Approach to Aid in Diagnosing Barrett's Oesophagus Related Dysplasia**

By

Deema Alabdullatif

Submitted in accordance with the requirements for the degree of  
Doctor of Philosophy

The University of Leeds  
Faculty of Engineering and Physical Sciences  
School of Computing

April, 2022

The candidate confirms that the work submitted is her own and that appropriate credit was given where reference was made to the work of others.

This copy was supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Deema Alabdullatif to be identified as Author of this work has been asserted by her in accordance with the Copyright, Designs and Patents Act 1988.

"To my beloved late grandfather,  
Abdurrahman Alabdullatif,  
for his love of knowledge."

## **Acknowledgements**

First and foremost, I am grateful to God the Almighty for His Blessings of health and patience.

I wish to express my gratitude and thanks to my country, the government of Saudi Arabia and the Saudi Cultural Bureau in the United Kingdom for their financial support and generosity.

I want to thank Mr Abdullah Abahusseini, who supported me and stood by my side during the stressful phases.

I extend my sincere appreciation to my supervisors, Dr Andy Bulpitt and Dr. Marc de Kamps, who have the substance of genius to help me confront problems and obstacles. Thank you for your clear guidance and encouragement. Without your persistent help, the goal of this research would not have been achieved.

I thank Dr Darren Treanor, this research consultant pathologist, who enriched me with his knowledge by explaining the process of Barrett's Oesophagus diagnosis, besides the process of producing histological images.

Also, I would like to acknowledge: Dr Shadi Albarquni for providing me with his feedback and sound advice regarding my research, Mr David Turner, who helped me access the dataset and the incredible ARC3 team for their technical support.

I want to express my heartfelt gratitude and profound thanks also to the following people for their constant encouragement and support throughout my journey:

♥ To my wonderful parents, my mother Lulwah and my father Abdullah, for their love and encouragement in every step of my life. You are my idols!

♥ To my wonderful siblings Reema, Nora, Maha, Sara, Abdurrahman and Fatimah and my dear sister-in-law Rahaf for their love and help. I am blessed to have you!

♥ To Victoria Masters, the one person I could talk to through my tough times. Thank you for being understanding, patient, kind and friendly.

♥ To the wonderful staff at the University of Leeds, who are always welcoming.

♥ To my friends and colleagues for all the lovely moments we shared.

Finally, this research would not have been accomplished without their blessings. I owe you a debt of gratitude.



## Abstract

Barrett's oesophagus is the only known precursor to oesophagus carcinoma. Histologically, it is defined as a condition of columnar cells replacing the standard squamous lining. Those altered cells are prone to cytological and architectural abnormalities, known as dysplasia. The dysplastic degree varies from low to high grade and can evolve into invasive carcinoma or adenocarcinoma. Thus, detecting high-grade and intramucosal carcinoma during the surveillance of Barrett's oesophagus patients is vital so they can be treated by surgical resection. Unfortunately, the achieved interobserver agreement for grading dysplasia among pathologists is only fair to moderate. Nowadays, grading Barrett's dysplasia is limited to visual examination by pathologists for glass or virtual slides. This work aims to diagnose different grades of dysplasia in Barrett's oesophagus, particularly high-grade dysplasia, from virtual histopathological slides of oesophagus tissue.

In the first approach, virtual slides were analysed at a low magnification to detect regions of interest and predict the grade of dysplasia based on the analysis of the virtual slides at 10X magnification. Transfer learning was employed to partially fine-tune two deep-learning networks using healthy and Barrett's oesophagus tissue. Then, the two networks were connected. The proposed model achieved 0.57 sensitivity, 0.79 specificity and moderate agreement with a pathologist.

On the contrary, the second approach processed the slides at a higher magnification (40X magnification). It adapted novelty detection and local outlier factor alongside transfer learning to solve the multiple instances learning problem. It increased the performance of the diagnosis to 0.84 sensitivity and 0.92 specificity, and the interobserver agreement reached a substantial level.

Finally, the last approach mimics the pathologists' procedure to diagnose dysplasia, relying on both magnifications. Thus, their behaviours during the assessment were analysed. As a result, it was found that employing a multi-scale approach to detect dysplastic tissue using a low magnification level (10X magnification) and grade dysplasia at a higher level (40X magnification). The proposed computer-aided diagnosis system was built using networks from the first two approaches. It scored 0.90 sensitivity, 0.94 specificity and a substantial agreement with the pathologist and a moderate agreement with the other expert.

## Table of Contents

<b>Acknowledgements</b> .....	<b>IV</b>
<b>Abstract</b> .....	<b>V</b>
<b>Table of Contents</b> .....	<b>VI</b>
<b>List of Tables</b> .....	<b>X</b>
<b>List of Figures</b> .....	<b>XIII</b>
<b>Chapter 1. Introduction</b> .....	<b>1</b>
1.1 An overview.....	1
1.2 Research motivation, aim and objectives.....	3
1.3 Thesis structure and research framework .....	4
1.4 Research questions .....	6
1.5 Research contributions to current knowledge .....	6
<b>Chapter 2. Literature Review</b> .....	<b>9</b>
2.1 Histopathology .....	9
2.1.1 Histopathology slides .....	9
2.1.2 Anatomy of the normal oesophagus.....	12
2.1.3 Barrett's oesophagus .....	15
2.1.4 Dysplasia in Barrett's oesophagus.....	17
2.1.5 Virtual slides.....	21
2.1.6 Pathology of different degrees of dysplasia .....	23
2.1.6.1 Negative for dysplasia (NFD) in Barrett's oesophagus (metaplasia).....	28
2.1.6.2 Indefinite for dysplasia .....	28
2.1.6.3 Low-grade dysplasia (LGD) .....	29
2.1.6.4 High-grade dysplasia (HGD).....	29
2.1.6.5 Intramucosal carcinoma (IMC).....	30
2.1.6.6 Summary.....	30
2.1.7 Clinical challenges .....	30
2.1.7.1 Intraobserver and interobserver variation in diagnoses.....	31
2.1.7.2 Costs of misclassifying dysplasia in Barrett's oesophagus .....	32
2.1.8 Colour normalisation .....	33
2.2 Related works in diagnosing dysplasia in Barrett's oesophagus.....	33

2.3	Performance metrics for medical tasks .....	35
2.4	Deep-learning and its biological inspiration .....	37
2.5	Deep-learning architectures .....	39
2.5.1	Convolutional Neural Network (CNN).....	39
2.5.2	“We need to go deeper!” .....	44
2.6	Learning approaches of deep-learning architectures .....	49
2.6.1	Supervised learning .....	49
2.6.2	Unsupervised learning .....	52
2.6.3	Weakly supervised learning .....	54
2.6.4	Transfer learning .....	61
2.7	Deep one-class classification .....	64
2.8	Discussion and conclusion .....	70
<b>Chapter 3.</b>	<b>Histopathology Dataset of Barrett’s Oesophagus .....</b>	<b>75</b>
3.1	Research material .....	75
3.2	Ground truth of virtual slides .....	76
3.3	Ground truth of annotated regions .....	79
3.4	Virtual slides and annotated regions selections .....	81
3.5	Tissue segmentation and Noise Reduction .....	85
3.6	Annotation masks generation.....	89
3.7	Sampling patches from annotated regions .....	90
3.8	Sampled patches pre-processing.....	96
3.9	Dataset challenges.....	96
3.10	Summary.....	98
<b>Chapter 4.</b>	<b>Regions of Interest Detection and High-level Analysis and Classification.....</b>	<b>100</b>
4.1	Introduction .....	100
4.2	An overview of the proposed model .....	103
4.3	Methodology.....	104
4.3.1	Regions of interest detection and dysplastic classification.....	105
4.3.2	Annotation-level and slide-level inference.....	107
4.4	Experimental design.....	107
4.5	Datasets .....	110
4.6	Experiments and results.....	112
4.6.1	Regions of interest detection.....	112
4.6.2	Dysplastic high-level analysis and annotation grading .....	115

4.6.3	The proposed model .....	120
4.7	Discussion.....	121
4.8	Conclusion .....	130
<b>Chapter 5.</b>	<b>Histopathology Low-level Analysis and Classification.....</b>	<b>132</b>
5.1	Introduction .....	132
5.2	An overview.....	134
5.3	Methodology.....	137
5.3.1	Potential dysplastic tissue detection.....	137
5.3.1.1	Feature extraction .....	141
5.3.1.2	Feature classification .....	143
5.3.2	Feature classification for the unfiltered patches .....	145
5.4	Experimental design.....	146
5.4.1	Novelty detection training.....	146
5.4.2	Novelty detection testing .....	148
5.4.3	Dysplasia classification .....	151
5.4.4	The proposed model assembly .....	151
5.5	Experiment datasets .....	152
5.5.1	Target dataset.....	152
5.5.2	Reference dataset.....	153
5.5.3	Signature dataset.....	153
5.5.4	Low-level based classification dataset .....	154
5.6	Results .....	154
5.6.1	Potential dysplastic tissue detection.....	154
5.6.2	Low-level based classification .....	157
5.6.3	The proposed model .....	159
5.7	Discussion.....	160
5.8	Conclusion .....	168
<b>Chapter 6.</b>	<b>Automated Dysplasia Detection and Grading in the Whole Virtual Slides.....</b>	<b>170</b>
6.1	Introduction .....	170
6.2	Methodology.....	171
6.2.1	A potential CAD system based on the consensus grading for the analysis at 10X and 40X magnifications....	179
6.2.2	The proposed CAD system .....	183
6.3	Empirical evaluation and results comparison .....	184

6.4	Datasets .....	190
6.5	Results of the selected CAD system .....	191
6.5.1	Background elimination and tissue detection .....	191
6.5.2	NFD classification and dysplasia detection (10X) .....	191
6.5.3	Detected dysplastic tissue classification (40X).....	194
6.5.4	Computational time .....	200
6.6	Discussion .....	201
6.7	Conclusion .....	205
<b>Chapter 7.</b>	<b>Conclusion and Future Work .....</b>	<b>208</b>
7.1	Thesis summary .....	208
7.2	Key contributions and findings .....	210
7.3	Research strengths, limitations and opportunities for future research .....	214
	<b>List of References .....</b>	<b>218</b>
	<b>List of Abbreviations.....</b>	<b>228</b>
	<b>Appendix A .....</b>	<b>230</b>
	<b>Appendix B .....</b>	<b>234</b>
	<b>Appendix C .....</b>	<b>240</b>

## List of Tables

<b>Table 2.1</b> Comparison of different classification systems of Barrett's associated dysplasia.....	<b>19</b>
<b>Table 2.2</b> Cytological and architectural abnormalities in Barrett's oesophagus associated dysplasia.....	<b>24</b>
<b>Table 2.3</b> Explanation of the cytological and architectural changes of each grade in dysplasia .....	<b>26</b>
<b>Table 2.4</b> Summary of the performance metrics and their corresponding equations.....	<b>35</b>
<b>Table 2.5</b> Showing the confusion matrix for observer1 against obserever2 .....	<b>37</b>
<b>Table 2.6</b> A summary of the related works that employ supervised learning in the field of histology.....	<b>51</b>
<b>Table 2.7</b> A summary of the unsupervised related works in the field of histology.....	<b>53</b>
<b>Table 2.8</b> A summary of the MIL weakly supervised related works in the field of histology .....	<b>60</b>
<b>Table 2.9</b> An overview of the best performing works that adapt the deep transfer learning approach in CAMELYON 16, CAMELYON 17 and BACH competitions .....	<b>63</b>
<b>Table 2.10</b> A summary of the semi-supervised deep one-class classification related works .....	<b>70</b>
<b>Table 3.1</b> The ground truth labels for the whole virtual slides within different subsets of the data provided by "Expert_B" and "Expert_E".....	<b>78</b>
<b>Table 3.2</b> The ground truth labels for the annotations within different subsets of the data provided by "Expert_B" and "Expert_E" .....	<b>80</b>
<b>Table 3.3</b> Details about the filtered annotations through phases one and two .....	<b>83</b>
<b>Table 3.4</b> Standard histological virtual slides magnifications and their associated Aperio levels .....	<b>89</b>
<b>Table 3.5</b> The number of extracted patches from annotations at 40X and 10X.....	<b>92</b>
<b>Table 4.1</b> The confusion matrices for the model based on the analysis of the patches, annotations and slide levels at 10X magnification ....	<b>117</b>
<b>Table 4.2</b> The performance measurements for the model based on the analysis of the patches, annotations and slide levels at 10X magnification (three-tier classification).....	<b>117</b>
<b>Table 4.3</b> The confusion matrices for the proposed model at the annotation and slide levels (three-tier) .....	<b>120</b>

<b>Table 4.4</b> The performance measurements for the proposed model at the annotation and slide levels (three-tier) .....	121
<b>Table 4.5</b> The confusion matrices for the model based on the analysis of the patch, annotation and slide levels at 10X magnification (two-tier classification).....	124
<b>Table 4.6</b> The performance measurements for the model based on the analysis of the patch, annotation and slide levels at 10X magnification (two-tier classification).....	124
<b>Table 4.7</b> Performance measurements for the high-level analysis and classification solely against it coupled with regions of the interest detection model.....	125
<b>Table 4.8</b> The three-tier and two-tier classification for the proposed model against other works at the slide-level with 95% confidence intervals.....	126
<b>Table 4.9</b> The proposed model slide-level and annotation-level agreements with experts .....	129
<b>Table 5.1</b> The number of patches within the training, validation, and testing sets in the target dataset .....	152
<b>Table 5.2</b> The number of patches within the training and the validation sets that were drawn from the sampled patches at 40X magnification .....	154
<b>Table 5.3</b> Obtained results for design decisions relating to the one-class classifier .....	156
<b>Table 5.4</b> The confusion matrices for the features classification model based on analysing slides at 40X magnification at the patch, annotation and slide levels (three-tier) .....	158
<b>Table 5.5</b> The performance measurements for the features classification model based on analysing slides at 40X magnification at the patch, annotation and slide levels (three-tier)....	158
<b>Table 5.6</b> The confusion matrices for the proposed model on the annotation and slide levels (three-tier) .....	160
<b>Table 5.7</b> The performance measurements for the proposed model on the annotation and slide levels (three-tier) .....	160
<b>Table 5.8</b> Performance measurements for features classification model based on analysing slides at 40X magnification solely against it is coupled with potential dysplastic tissue detection model .....	164
<b>Table 5.9</b> The three-tier and two-tier classification performances for the proposed model against the work proposed by Sali et al. (2020) at the annotation-level.....	167
<b>Table 6.1</b> Some of the inspiring pathology guidelines in designing the consensus system and their correspondence with a logical expression.....	180
<b>Table 6.2</b> The propositional logic table for the proposed consensus grading system.....	182

<b>Table 6.3</b> The architecture compositions for the tested approaches in the empirical evaluation. ....	<b>185</b>
<b>Table 6.4</b> The interobserver agreements for different approaches at the annotation and slide levels.....	<b>188</b>
<b>Table 6.5</b> List of grades for the test set provided by two pathologists for the virtual and glass slides and their associated diagnosis by the proposed CAD system .....	<b>195</b>
<b>Table 6.6</b> Results for different modules in the proposed CAD system.....	<b>196</b>
<b>Table 6.7</b> The interobserver agreements for trainee pathologists, expert pathologists and the proposed CAD system against the expert pathologists .....	<b>204</b>
<b>Table 6.8</b> The interobserver agreements in the diagnosing of dysplasia per biopsy and per patient in (Salomao et al., 2018) paper and the proposed CAD system .....	<b>205</b>



## List of Figures

Figure 1.1 An overview diagram for the work in chapters 3,4,5 and 6.....	3
Figure 2.1 The differences in endoscopies and histologic examinations between a healthy person and Barrett's oesophagus patients .....	13
Figure 2.2 Oesophagus four layers (mucosa, submucosa, muscularis propria, and adventitia).....	14
Figure 2.3 A transverse section histology image of the oesophagus with its four different layers at 500 $\mu\text{m}$ power (Peckham et al., 2003).....	15
Figure 2.4 (a) Squamous cells in healthy oesophagus epithelium, and (b) columnar cells in Barrett's oesophagus epithelium .....	17
Figure 2.5 Chart summarises the staging manual for cancer by the American joint committee .....	20
Figure 2.6 (A) the pyramid structure of a virtual pathology slide, (B) different zoom level representations for the same part of virtual pathology tissue.....	23
Figure 2.7 Recommended treatment plan for the patients with Barrett's oesophagus by the Practice Parameters Committee of the American College of Gastroenterology .....	32
Figure 2.8 A nervous cell.....	38
Figure 2.9 A building inception block in GoogleNet (inception V1).....	45
Figure 2.10 A residual building block in ResNet architecture.....	46
Figure 2.11 A comparison between different famous CNN architectures.....	47
Figure 2.12 Illustration for the Inception-ResNet-V2 architecture (Szegedy et al., 2017).....	48
Figure 2.13 An overview of the unsupervised convolutional auto-encoder model.....	52
Figure 2.14 Illustrations for the scenarios of supervised, unsupervised and the different weakly supervised tasks .....	55
Figure 2.15 A tree diagram for the different categorisation schemes of MIL approaches .....	57
Figure 2.16 Fine-tuning scenarios for deep-transfer learning based on the available dataset.....	63
Figure 2.17 Different one-class classification approaches based on the availability of the ground-truth labels of the training dataset .....	65

Figure 2.18 Samples from Barrett's oesophagus related dysplasia histological images dataset and the CIFAR-10 dataset .....	72
Figure 2.19 The expected distribution of inliers and outliers for CIFAR-10 and Barrett's oesophagus related dysplasia histological images datasets .....	73
Figure 3.1 Examples for accepted and rejected annotations .....	82
Figure 3.2 Histological artefacts examples .....	85
Figure 3.3 Tissue detection .....	86
Figure 3.4 The applied tissue detection and segmentation methods on a whole virtual slide from the train set.....	87
Figure 3.5 Visualisation for the regenerated curved NFD annotation mask.....	90
Figure 3.6 Patches sampling at two different magnifications .....	91
Figure 3.7 Demonstration of sampling 256x256 patches from the same tissue at four different low magnifications .....	92
Figure 3.8 Column charts for the number of annotated regions at the available magnifications .....	93
Figure 3.9 Samples for the extracted patches at different magnifications from different grades of dysplasia .....	95
Figure 3.10 An annotated region from the "CAMELYON16" dataset....	97
Figure 4.1 An overview of the annotation grade inference submodel .....	101
Figure 4.2 The proposed model for regions of interest detection and dysplasia classification.....	102
Figure 4.3 An overview of the architecture of Keras Inception-ResNet-v2 .....	104
Figure 4.4 Samples of the unimportant regions: (a) healthy oesophagus biopsy, (b) healthy epithelium, (c) Unknown but was never annotated by the pathologists, (d) muscular mucosa, and (e) wax .....	111
Figure 4.5 Samples of the epithelial layer mask dataset .....	112
Figure 4.6 The losses and accuracies evolutions for the training (in pink) and the validation (in purple) sets during training regions of interest subnetwork.....	113
Figure 4.7 Examples of the regions of interest detection results for some of the provided test annotations .....	113
Figure 4.8 Pie charts show the percentage of detected regions within the provided test annotations and the number of detected patches within each grade.....	114
Figure 4.9 The results of the region of interest detection model for IMC annotations from an IMC slide .....	114

Figure 4.10 The losses and accuracies evolutions for the training (in pink) and the validation (in purple) sets during the training of the subnetwork to classify Barrett’s oesophagus related dysplasia based on the analysis at 10X magnification.....	116
Figure 4.11 Examples of the correctly classified regions and the misclassified regions by the analysis of the annotations at 10X magnification from each grade.....	118
Figure 4.12 Example of a region labelled differently by two pathologists.....	119
Figure 4.13 Examples for small annotations that their predictions do not influence the label of their container slides.....	120
Figure 4.14 Comparison between the grade of the glands arrangements in annotation (b) and the precancerous progression form in (a).....	123
Figure 4.15 The slide-level confusion matrices for the first and second benchmarks follow the three-tier and two-tier classifications .....	127
Figure 5.1 Illustration for weakly supervised MIL problem .....	133
Figure 5.2 A diagram for the unfilled gap in the literature.....	134
Figure 5.3 The proposed model for dysplasia classification based on the analysis at 40X magnification .....	136
Figure 5.4 Feature space obtained using deep SVDD features: (a) expected, (b) real.....	139
Figure 5.5 Training framework for the potential dysplastic tissue detection model.....	142
Figure 5.6 The losses and accuracies evolutions for the training (in pink) and the validation (in purple) sets during training dysplasia low-level based classification subnetwork.....	146
Figure 5.7 Different loss functions during training the potential dysplastic tissue detection subnetwork. (a) Compactness loss, (b) descriptiveness loss, and (c) total loss for the validation set. ....	147
Figure 5.8 The obtained feature space projection, using t-SNE visualisation, using the novelty detection method for the testing set patches. Non-dysplastic (in blue) and dysplastic patches (in red) are relatively separated.....	148
Figure 5.9 Novelty detection with LOF .....	150
Figure 5.10 Testing framework for the potential dysplastic tissue detection model.....	151
Figure 5.11 Samples from the PCam dataset (Veeling et al., 2018).....	153
Figure 5.12 Pie charts show the percentage of detected potential dysplastic tissue within the initial testing set and the number of detected patches from each grade.....	155

Figure 5.13 Confusion matrices for different one-class classifiers ....	156
Figure 5.14 Samples from the test annotations show the detected non-dysplastic tissues (in light blue) .....	157
Figure 5.15 The prediction maps for annotations from the test set from each grade, following the Vienna classification .....	159
Figure 5.16 Examples of misleading overlapped annotations .....	163
Figure 5.17 The analysis provides the prediction maps for three test annotations from each grade at the 40X magnification model (on the left) against its predictions coupled with the potential dysplastic tissue detection model (on the right) .....	165
Figure 5.18 The annotation-level confusion matrices for Sali et al. (2020) .....	166
Figure 6.1 The relation between annotation (the red box), biopsies (the black boxes) and WSI (the blue box that surrounds the whole image) .....	170
Figure 6.2 The architecture of the potential CAD system following the first approach .....	173
Figure 6.3 Confusion matrices for different models at the annotation and the slide levels .....	175
Figure 6.4 Bar charts for the performance measurements for each grade and the overall grades at the annotation-level for different models .....	176
Figure 6.5 Bar charts for the performance measurements for each grade and the overall grades at the slide-level for different models .....	177
Figure 6.6 The general pyramidal architecture of the first and second approaches .....	178
Figure 6.7 The proposed CAD system architecture following the second approach .....	184
Figure 6.8 NFD, LGD and HGD annotations with their grades using different architectures .....	186
Figure 6.9 Examples for the overall region of interest detection measurement categories for annotations .....	189
Figure 6.10 The results comparison for the region of interest's module for the first and second approaches .....	190
Figure 6.11 Samples of the epithelial layer mask dataset for the tissues .....	191
Figure 6.12 Samples of the misclassified biopsies in slide "11040.svs" using the analysis at 10X magnification .....	192
Figure 6.13 Prediction maps for the test set slides produced by the analysis at 10X magnification submodel .....	193

<b>Figure 6.14 The slide-level confusion matrices for the NFD classification and dysplasia detection submodel.....</b>	<b>194</b>
<b>Figure 6.15 The distribution of each dysplasia grade within each test slide using the proposed CAD system .....</b>	<b>198</b>
<b>Figure 6.16 Visualised prediction maps for some test slides using the proposed CAD system (the second approach) .....</b>	<b>199</b>
<b>Figure 6.17 The slide-level confusion matrices for the CAD system..</b>	<b>200</b>
<b>Figure 6.18 The confusion matrices for the CAD system in the term of biopsies following three-tier classification.....</b>	<b>204</b>
<b>Figure 7.1 NFD, LGD, and HGD samples show similarities between Barrett's oesophagus tissue and colon tissue.....</b>	<b>215</b>

## **Chapter 1. Introduction**

### **1.1 An overview**

Barrett's oesophagus is a medical condition that results from the growth of abnormal cells in the oesophagus lining, as a columnar lining replaces the usual squamous lining. The transformation process could evolve into different forms of dysplasia, which are considered precancerous changes that could eventually lead to oesophageal cancer. Worldwide, oesophageal cancer is one of the deadliest types of cancer, with a dismal survival rate (Delpisheh et al., 2014). That disease could be diagnosed and monitored via endoscopic sessions with biopsies extraction to be examined histologically by pathologists. Unfortunately, monitoring and examining different types of dysplasia are critical and expensive, and many studies showed that diagnosing dysplasia in Barrett's oesophagus suffers from low agreement among pathologists (Wani et al., 2016) and (Vennalaganti et al., 2017). Those limitations create a gap in diagnosing and classifying different grades of dysplasia in Barrett's oesophagus that suggests and encourage artificial intelligence researchers to develop and improve accurate and reliable computer-aided diagnosis (CAD) to assist in the diagnosing by either detecting the region of interest where the pathologists should examine or localising and classifying different grades of dysplasia.

Developing CAD systems to analyse histological images requires learning models to extract features from virtual tissue slides. The learning approaches could follow conventional machine-learning or deep-learning feature engineering methods. The conventional approaches require the manual design of feature extractors, which involves domain experts. Those techniques are expensive and introduce cognitive bias. Therefore, deep-learning approaches are considered alternative feature extractors that do not need to be handcrafted (Sali et al., 2020). Although those approaches do not involve domain experts in designing the feature extractors, they still require them to provide massive and precise annotations from the tissue to train models in a supervised manner. Unsupervised and weakly supervised approaches offer a low-cost dataset annotation process to overcome this obstacle. On the one hand, unsupervised learning does not require a labelled dataset.

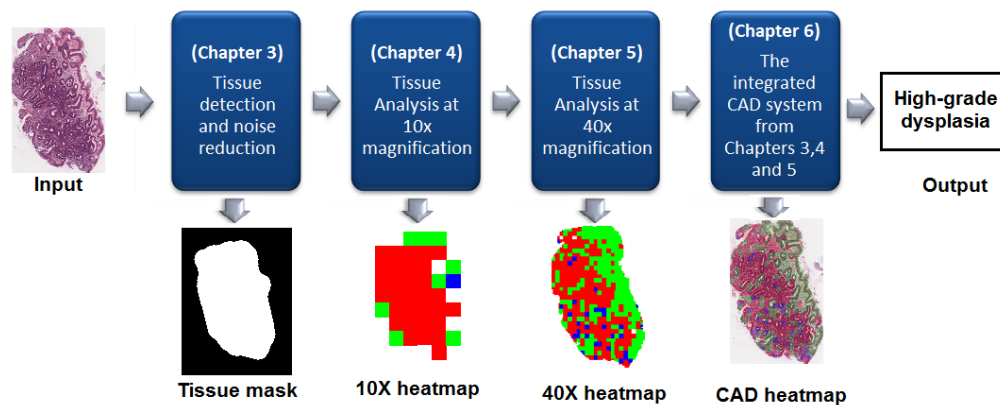
On the other hand, weakly supervised learning can handle datasets with multiple labels, inaccurate labels and labels for bags consisting of unlabelled instances, also known as multiple instances learning (MIL). Furthermore, transfer learning offers a solution to cases where insufficient dataset sizes are available to train models. Transfer learning is based on using a trained model on a massive related or unrelated field problem as an initial model for a new case. Then, the model is fine-tuned using the dataset of the new case.

The available dataset for this thesis is composed of whole virtual slides histological images for biopsies that were extracted from Barrett's oesophagus patients with different grades of dysplasia. From each whole virtual slide, different annotations, sometimes overlapped annotations, were labelled by domain pathologists and the whole virtual slide has the label of the highest grade of any of its contained annotations. Furthermore, each annotation contains multiple unlabelled patches. That dataset is a perfect case that represents the MIL problem, which MIL and deep-learning could analyse.

MIL deep-learning approaches generally train a model at the instance level, assuming that each instance in the bag has the bag's label. Usually, at the prediction phase of the bag, a spatial pooling over the instances is performed to infer the bag label. To the best of our knowledge, no studies have been conducted on employing a deep-learning one-class classifier to cleanse the bags from the instances that do not belong to them. This thesis will introduce a framework trained in a weakly supervised manner, particularly MIL, and it employs a one-class classifier to overcome this challenge. Section 1.3 provides information about the research framework.

This thesis will demonstrate the development of the proposed CAD system to detect dysplasia in Barrett's oesophagus and grade it into low-grade or high-grade dysplasia (refer to Chapter 6). Processing and analysing the WSIs is accomplished at different magnifications, starting from the lowest power magnification and increasing the power to perform more analysis that is complicated. The proposed work starts with tissue detection (background elimination) and noise reduction at the available thumbnail magnification (1.25X or 2.5X) for tissue detection and at 5X for noise reduction, as

discussed in Chapter 3. Then, the magnification power is increased to detect dysplastic tissue at 10X magnification (refer to Chapter 4). Finally, the detected dysplastic tissue at 10X will be analysed at higher power magnification (40X) to discriminate low-grade dysplasia from high-grade dysplasia (refer to Chapter 5). Figure 1.1 provides an overview diagram that illustrates the workflow of the CAD system over Chapter 3, Chapter 4, Chapter 5 and Chapter 6.



**Figure 1.1** An overview diagram for the work in chapters 3,4,5 and 6

## 1.2 Research motivation, aim and objectives

Grading dysplasia in Barrett's oesophagus suffers from a suboptimal interobserver agreement even between expert gastrointestinal pathologists. There is an intraobserver disagreement when a pathologist assesses a slide on different occasions. The interobserver and intraobserver disagreements are attributed to the lack of clearly defined guidelines for the grading process. Moreover, the dysplastic changes in Barrett's oesophagus are continuous, with undefined boundaries between each grade and its adjacent grades. This research is motivated mainly by the previously mentioned facts and the need for automating the process of analysing the available virtual pathology slides to save the time and cost of the manual process.

Thus, this research aims to shed light on the grey area where pathologists disagree by developing a CAD system that detects and grades dysplasia in Barrett's oesophagus. Furthermore, the developed system should increase the diagnosing performance for high-grade dysplasia, mainly because it



needs a surgical intervention to limit cancer progression, knowing that patients with oesophageal adenocarcinoma are predicted to have a dismal 5-year survival (Vennalaganti et al., 2017).

Therefore, this research aims to aid in measuring the degree of dysplasia in Barrett's oesophagus from the virtual slides using the cytological and architectural abnormality changes. In addition, this research should help identify the regions where a pathologist or the developed analyser examines to pave the way for the whole virtual slide analysis.

### **1.3 Thesis structure and research framework**

The remainder of this thesis is structured as follows: Chapter 2 provides detailed information about the histological anatomy of the normal oesophagus and Barrett's oesophagus and the cytological and architectural abnormalities that occur in each grade of dysplasia in Barrett's oesophagus. Having greater insight into the clinical guidelines in the diagnosing process would yield valuable information that helps design the CAD system and understand its performance. Additionally, it discusses the approaches of other attempts to diagnose this disease and the key papers concerning the current work, such as the conducted work in the deep-learning architectures and learning approaches, including weakly supervised learning, transfer learning and one-class classification.

Based on the reviewed literature and the conducted experiments, the developmental framework for the proposed CAD system is summarised in Figure 1.1. The framework shows that diagnosing dysplasia starts with whole slide pre-processing, including tissue detection, noise reduction, and patches sampling at low-power magnification. Then, the dysplastic regions are detected based on the analysis of the whole slide images (WSIs) at a higher power magnification (10X). After that, another higher-level analysis is performed at 40X magnification to grade the severity of dysplasia in the detected regions. Finally, the whole system is run in sequence to produce a heatmap representing the local-level classification of dysplasia for tissues within the inputted virtual slide. An inference histogram-based system is employed to grade the whole tissue.

Chapter 3 explains the used dataset and provides detailed explanations of the annotation process and the available ground-truth labels. In addition, it highlights the dataset challenge. The adopted image pre-processing algorithms to detect tissues (Foreground) and reduce the noise are also discussed. It contains the technique followed in sampling patches.

Chapter 4 explains the experiment that attempts to detect regions of interest by training a unique Convolutional Neural Network (CNN) architecture to discriminate between the structure of a normal oesophagus and Barrett's oesophagus tissues. Moreover, the experiment of extracting features at 10X magnification and grade dysplasia at the patch-level using them, then the annotation-level grading inference sub-model that employs the generated heatmaps to predict the annotation grade are included. Finally, the two networks are combined to grade the detected regions of interest only to reduce the computational cost, and the proposed model results are discussed.

Chapter 5 discusses the weakly supervised problem of MIL that is introduced by the nature of the histological images. It provides the implementation to manipulate this issue using a one-class classifier. The proposed solution can be used in two ways:

- To prepare the training dataset by filtering the non-dysplastic patches as much as possible from the dysplastic annotations. As a result, the cleaned training dataset can be used to train the low-level-based classification network in a supervised manner without worrying about confusing the classifier.
- Add the proposed solution to the previously mentioned network to boost performance by filtering the non-dysplastic patches from the test set.

Chapter 6 analyses the method pathologists follow in diagnosing the disease to design a consensus diagnosis between the architectural and the cytological analysis by summarising the clinical guidelines for diagnosing and grading Barrett's related dysplasia to design a logical system and also relying on the assumption that the extracted features at 10X and 40X magnifications represent architectural and cytological features, respectively. Furthermore, it analyses their behaviours in the annotations process by

observing the magnification level of each annotation and its corresponding label. As a result of the observation, the analysis of WSIs at a 10X magnification-based classifier was employed as a region of interest detector to detect dysplasia and to employ the analysis at a 40X magnification-based classifier as a dysplasia classifier. Lastly, it provides empirical experiments for different combinations of networks to decide the best performing CAD system.

Finally, Chapter 7 concludes the research work, provides a wrap-up discussion of the research outcomes, highlights the limitation of the proposed work, suggests further work to improve the work in the future, and discusses the possibility of extending it to be used in other related problems such as grading colon dysplasia.

## **1.4 Research questions**

This research aims to propose a CAD system that aids pathologists in detecting and grading dysplasia in Barrett's oesophagus. That is achieved by answering the following research questions:

**Question 1:** When testing performance for grading dysplasia in Barrett's oesophagus at 10X magnification, how do a conventional machine learning-based model proposed by Adam (2015) and a weakly supervised deep-learning model compare?

**Question 2:** How effective is employing the deep-learning one-class classification algorithm in addressing the multiple instances problem in histological images?

**Question 3:** How do the analysis and imitation of the pathologists' behaviours while designing a CAD system in selecting the magnification to grade levels of dysplasia in Barrett's oesophagus affect the performance of the CAD system?

## **1.5 Research contributions to current knowledge**

The contributions of this thesis can be summarised and presented in chronological order as follows: Chapter 3 contributes to the field of virtual

histopathology image pre-processing by providing a tool to pre-process H&E stained virtual slides. The tool integrated two approaches from other works. It includes the approach used by Haggerty et al. (2014) to detect the tissue and reduce noise. Some rules from the work of Adam (2015) and new rules from our observation were added for artefacts elimination. The tool samples patches from the detected tissue or the annotations.

Chapter 4 implemented a fork-style model to detect the region of interest (normal versus Barrett's tissue) and classify them using low-power magnification. It contributes to the pathology society by implementing a fast and low-cost, weakly supervised deep-learning model to grade dysplasia. Another contribution is a novel inference approach for grading the annotations and slides of dysplasia in Barrett's oesophagus. The inference approach relies on the histogram of the patch-level prediction. The location of each patch is taken into consideration as the contribution of the disease diagnosis varies from the epithelial layer to the lamina propria layer.

Chapter 5 developed a novel solution to tackle Barrett's oesophagus dataset's weakly supervised MIL problem. This solution contributed to the deep-learning community by proposing a one-class classifier to fill the gap in addressing the MIL problem following the object detection approach. This solution was used mainly to clean the training dataset before training the supervised network, which analysed the virtual slides based on the cytological abnormalities. In addition, it can be used to boost the performance of the low-level-based classification network by detecting the non-dysplastic patches from the virtual slides.

Chapter 6 contributes to the histopathology community by doing the following:

- It provides a logical system to compute the consensus diagnosis between the prediction of the networks that analyse WSIs at 10X and 40X magnifications.
- It provides a solution that emulates the pathologists' behaviours in detecting the region of interest. The solution uses the 10X magnification analysis network to detect dysplasia because empirical experiments showed that classifying dysplasia at that level has high

precision in grading NFD. Indicates that the system rarely predicts dysplastic tissue as not dysplasia.

- It provides a fully automated CAD that localises different dysplasia grades in tissues and classifies the slide into one of the three grades of dysplasia.

## **Chapter 2. Literature Review**

This Chapter presents the main research areas: histopathology, deep-learning architectures and approaches for learning such architectures, and the best performance metrics to evaluate the performance of the learnt deep models in the medical field. Finally, it discusses the one-class classification. All the discussion in this chapter is held in the light of that were used with histological images. Besides, it will highlight the limitation of the current works leading the discussion to possible approaches that will inform the methodology of this research.

### **2.1 Histopathology**

This section will tackle histology, particularly the virtual slides, the anatomy of the normal and the abnormal oesophagus, and the pathological criteria for diagnosing and grading dysplasia in Barrett's oesophagus. Finally, some of the clinical challenges that increase the difficulty in grading Barrett's oesophagus dysplasia will be discussed.

#### **2.1.1 Histopathology slides**

When a treatment of a patient from disease cannot be provided unless histology is conducted, histology takes place for further verification to enable the physician to diagnose by getting adequate tissue from the patient after the physician has already examined the patient physically, referred to the patient's history, or conducted the necessary imaging and laboratory examinations. That is the first stage in histology, where such tissue can be taken by fine-needle aspiration, needle biopsy, excisional biopsy, or complete damaged-area removal. Those procedures are arranged in ascending order according to sensitivity and specificity. The rates increase since large samples assist in understanding the contextual relationship of a cell and help pathologists examine different tissue slides. The tissue is examined closely using a microscopic tool for scaling by a pathologist, specifying its colour and features. However, large samples are handled by cutting them into smaller pieces to suit the tissue-holding cassette, i.e. about 10 × 10 × 3 millimetres.

After the tissue is collected, processing it (chemical and physical preparation) comes next, the second stage. In this stage, the tissue is first dipped in a solution preventing cell breakdown and the growth of microorganisms. This preparation process ranges from a few hours to 24 hours based on the size of the biopsy, as more extensive biopsies require a longer time. It is imperative to conduct that process for better sectioning and microscopic morphology. The chemical preparation is followed by the physical one in which the cellular morphology is retained using different means, including freeze-drying, microwave, and chemical means. The use of alcohols and xylene is commonly seen in many labs as the tissue is dried from the water and the fixing solution using a dehydrator. After that, the dehydrator is removed to prepare for the next step, which requires infiltrating the tissue with paraffin. In the last step, the tissue is altered to be firm during a process that takes time ranging from nine to a few hours, which is the case in many labs. The process can be accomplished by paraffin that gets heated in the processor to turn it into a liquid. Then by a vacuum, it infiltrates into the tissue and turns the processed tissue into a firm object after it cools down. This process leads to having a smaller sample than the original one.

The third stage is embedding, and in this stage, the previously prepared tissue is positioned in a mould, covering that tissue with paraffin. After that, it will be left on top of the cooling surface to harden it. This embedded tissue will be ready for sectioning since the solidified paraffin wax covers it. It is crucial to consider the positioning of the tissue when placed over the cassette since the tissue will be cut based on the holding cassette.

After embedding is completed, sectioning, the fourth stage, comes next. The tissue will be cut into slices during this stage and then placed on a flat microscope glass. Like a meat-slicing machine, a microtome is used to slice the tissue, whether manually, semi-automatically, or automatically. The manual slicer, for instance, is a microtome with a rotating handler that, when rotated by the handler, creates a thin tape-like-shaped tissue. Usually, the tissue is cut at a thickness ranging from three to four micrometres. It is somewhat difficult to cut the tissue into small slices because it may ruin the tissue, and it is essential to note that thick slices darken the stain and conceal the nuclear properties. After cutting the tissue into thin slices, the slices are placed over around 10°C hot water, avoiding reaching the melting degree of paraffin to remove any wrinkles caused by the microtome during

the slicing process. The final step is to put the tissue over the flat microscope glass (dimensions: 25 × 75 × 1 millimetre).

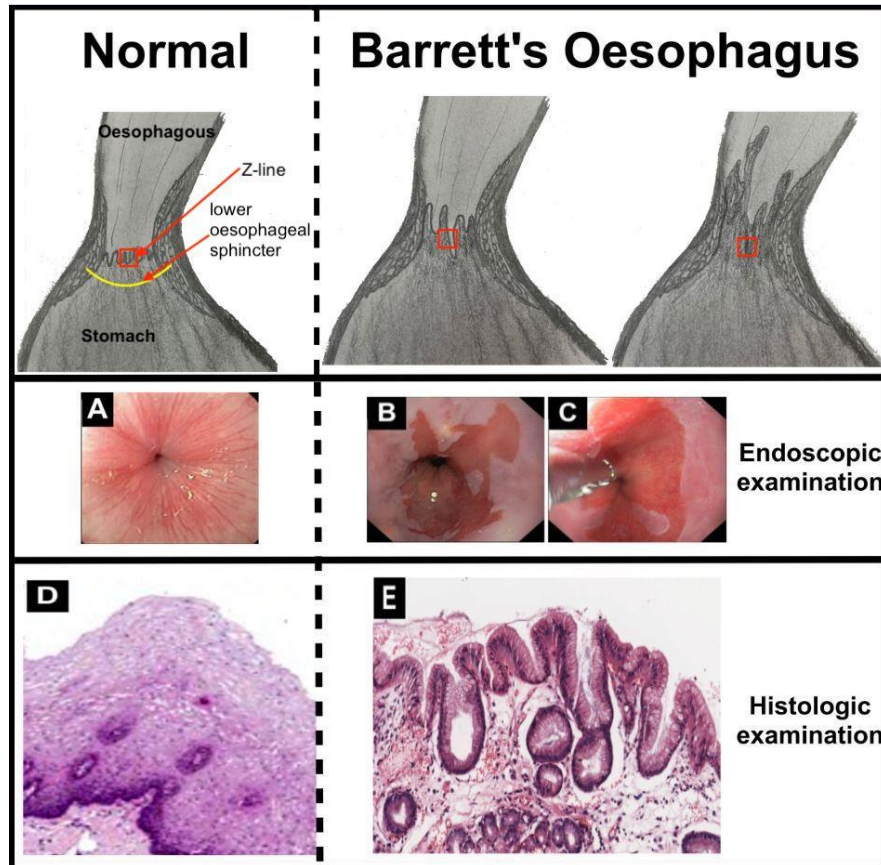
Staining is the fifth stage, during which staining techniques stain the slices, and the most common ones are haematoxylin and eosin (H&E). The staining is of great importance since the tissue will not be visible by merely using the microscope. When haematoxylin is used for staining, the nucleic acids turn blue, unlike eosin, which turns the proteins such as cytoplasm and connective tissue into pink (Gurcan et al., 2009). Those colours will appear to the eye when a bright-field microscope is used. In most cases, the blue colour represents cell nuclei, and the cytoplasm, based on its components, can be represented by either a bright red or purple colour. Since H&E stains most of the cells' constituents, those stains are still used until this day in pathology. In addition, they clearly differentiate between cells' constituents as they have chemical features staining those constituents with colours located on the opposite ends of the visible spectrum. The differences in those colours assist in the diagnosis process as they help in spotting tissue differences. The final step is to cover the sample slide with another smaller adhesive cover glass to prepare it for microscopic or digital visualisation, which is the sixth stage.

Recently, pathologists have preferred to turn samples into virtual ones, making diagnosis easier. Thus, whole-slide scanners are used to facilitate easier scanning of high quality. When samples are digitised, they become easier to save in the records, hence easier to recover. With the digital approach, samples are explored, note-recorded, and shared effectively. They can be used for different purposes, including education and discussions. Those samples may be detected efficiently and quantitatively in the future, whether entirely or partly automatically, to spot any problems in the tissue. Digital visualisation may take over the typical microscopic process due to its prominent benefits. Therefore, different companies worldwide provide such techniques with a spatial resolution of about 0.25 µm per pixel using an objective lens with a magnification of 40X. Saving the virtual samples would be better than the glass ones since they do not occupy physical storage spaces; besides, they are less likely to be damaged. However, this does not indicate that the concerned organisations will dispense the glass slides, as they should be kept minimally for ten years.



### **2.1.2 Anatomy of the normal oesophagus**

The oesophagus links the pharynx to the stomach as a channel with a length ranging from around 25 to 28 cm in adults, differing from one person to another depending on age, height, physical condition, and gender. In contrast, men tend to have a longer oesophagus than women (Ferhatoglu and Kivilcim, 2017). It moves food caudally towards the stomach and stops any stomach or oesophagus contents retrogression. The oesophagus is closed at its opposite ends like an empty pipe, with the upper oesophageal sphincter at the top and the lower oesophageal sphincter at the bottom (see Figure 2.1) (Fisichella and Patti, 2001). The gastrointestinal tract, which includes the oesophagus, is of a histologically distinguished structure. Like a tube with a various-diameter lumen, this tract consists of a four-layer wall; such layers include the mucosa, submucosa, muscularis propria (externa), and adventitia (Ferhatoglu and Kivilcim, 2017). Due to the absence of a serosa layer in the oesophagus, which was substituted with adventitia that works as a holder for the oesophagus and it binds it to the adjacent tissue and organs, infections and tumours tend to spread widely and quickly to the other organs once they start in the oesophagus (Shaheen et al., 2017).

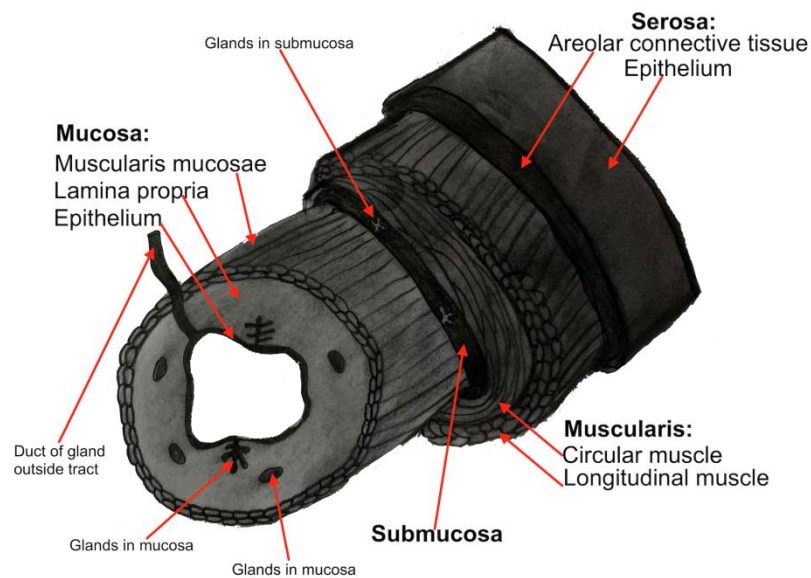


**Figure 2.1** The differences in endoscopies and histologic examinations between a healthy person and Barrett's oesophagus patients

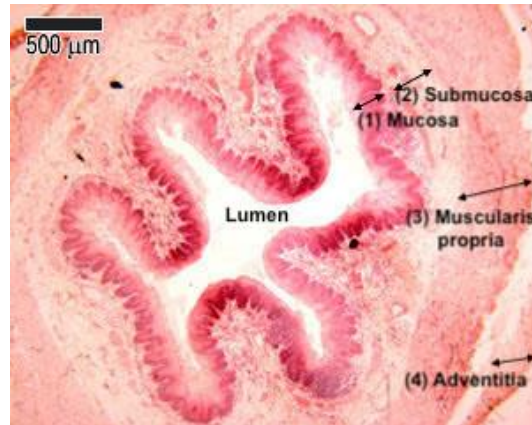
A and D belong to a healthy person, while B and C show salmon-coloured velvety mucosa appearing over the Z-line in a circumferential (C) and a tongues-like form (B). E shows the histologic image for how the histologic image of the biopsies should like if they are taken from the salmon regions in B and C.

According to (Ferhatoglu and Kivilcim, 2017) and (Peckham et al., 2003), a healthy oesophagus has the following histologic structure. The lumen wall of the oesophagus is covered by non-keratinising stratified squamous epithelium. The epithelium's basal layer contains columnar cells with a spherical cell nucleus. Cellular regeneration occurs in the basal layer as new cells disconnect from the basement membrane (i.e., an extracellular matrix of thin thickness, splitting the lamina propria from the epithelial layer). Those cells rise, reshaping and substituting the epithelium's inside layer. There is a layer under the epithelium providing vascular support. It consists of lymphatic capillaries, blood capillaries, and the lamina propria, a loose connective tissue. Such a supporting layer is vital to the epithelium as it reaches it by the papillae, the finger-like extensions.

Regarding histology representations, lamina propria cells of dark colour are considered lymphoid aggregations. The third layer of the mucosa is muscularis mucosa which consists of two layers that are thin and smooth longitudinally shaped muscles that assist in mucosa movement. The second layer in the oesophagus is the submucosa, characterised by being prominently vascular. It has loose connective tissue and oesophageal glands producing mucus that assists in food movement. The third layer is the muscularis propria, which has different muscles, as it has a skeletal muscle in the upper part, both a smooth and skeletal one in the centre and a smooth one in the lower part. Finally, the last external layer in the oesophagus is the adventitia, which has loose connective tissue covered with the visceral peritoneum. Such tissue includes blood vessels, lymph, and nerves. The oesophagus's layers are illustrated in Figure 2.2, while a transverse section of the oesophagus is illustrated in Figure 2.3, showing the four layers in histology representation at low power.



**Figure 2.2** Oesophagus four layers (mucosa, submucosa, muscularis propria, and adventitia)



**Figure 2.3** A transverse section histology image of the oesophagus with its four different layers at 500 μm power (Peckham et al., 2003)

As the concerned subject in this study, the first layer mentioned earlier in the oesophagus, the mucosa, consists of squamous cells like those in the skin or mouth. The colour of the normal squamous mucosal surface seems whitish-pink, unlike the gastric mucosa, containing columnar cells, as its colour ranges from salmon-pink to red. Figure 2.1 demonstrates where the Z-line is located in a typical case. The Z-line (also known as the squamocolumnar junction) is the line that separates and marks the meeting point of the oesophagus squamous mucosa and the gastric columnar mucosa (Ferhatoglu and Kivilcim, 2017).

### 2.1.3 Barrett's oesophagus

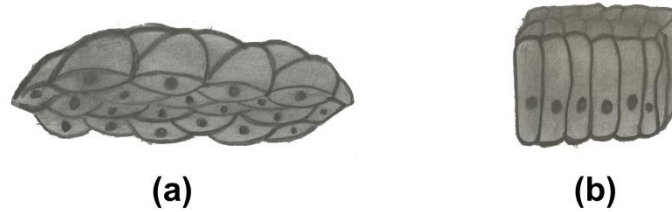
Globally, Barrett's Oesophagus is defined differently. Moreover, there is a conflict over differences in the stances regarding the necessity of intestinal metaplasia identification. Intestinal metaplasia can be identified with histological means once goblet cells appear in the gastric mucosa. As stated in the American College of Gastroenterology's guidelines, Barrett's Oesophagus is defined as a case in which the distal oesophagus shows changes that appear to be columnar using endoscopy, confirmed by the presence of intestinal metaplasia in the taken biopsy. Such a definition is approved as well by the American Gastroenterological Association.

On the other hand, the British and Japanese gastroenterologists do not see the need for the goblet cells' exploration using histological means, as the British Society of Gastroenterology states that Barrett's Oesophagus is the metaplastic columnar mucosa which can be explored using endoscopy (i.e.,

above the gastroesophageal junction by  $\geq 1$  cm). A biopsy can support such detection to decide whether it is metaplastic or not. However, histology is not necessary to recognise goblet cells, whether the gastric-type mucosa involves intestinal metaplasia or not (Grin and Streutker, 2014). The British Society of Gastroenterology's definition is endorsed in several studies (Kelty et al., 2007), (Gatenby et al., 2008) and (Liu et al., 2009)) as they indicate that metaplastic columnar epithelium that does not involve goblet cells could involve molecular abnormalities being somehow similar or the same to the ones in cases of goblet cells (Naini et al., 2016).

Previously, Barrett's oesophagus used to be deemed as a congenital abnormality. However, now it is believed that Barrett's oesophagus is a resisting effect of chronic acid exposure from reflux esophagitis (Gastroesophageal reflux disease (GERD)). The regurgitation of gastric contents to the oesophagus is known as Gastroesophageal reflux. Cases of acute GERD could result in erosions or ulcers, even though GERD rarely causes oesophagitis. Moreover, in healthy cases, ordinary squamous mucosa cells are regenerated to heal the effects of erosions. Nevertheless, in the case of Barrett's oesophagus patients, their cells are replaced with mucus-producing columnar cells. It is important to note that GRED is not the main reason behind that, as the severity of bile reflux could be another reason. However, it is a means of adaptation to the changes in an acidic environment (Basu and de Caestecker, 2002).

At the beginning of the smooth-surfaced healthy squamous cells' transformation (Figure 2.1 (D) and Figure 2.4 (a)) into the villiform-like metaplastic columnar cells (Figure 2.1 (E) and Figure 2.4 (b)), mucus-producing goblets and glands are generated in the oesophagus tissue, particularly in the epithelial layer. After that, such generation of goblets and glands occurs in the lamina propria, resulting in the change of cells and nuclei's size, shape and different cytological features. Those changes have different references depending on how acute they are (Naini et al., 2016). The different types are Barrett's Oesophagus, "indefinite for dysplasia", "low-grade dysplasia" (LGD), "high-grade dysplasia" (HGD), and intramucosal carcinoma (IMC).



**Figure 2.4** (a) Squamous cells in healthy oesophagus epithelium, and (b) columnar cells in Barrett's oesophagus epithelium

Barrett's oesophagus can be diagnosed with endoscopy when a salmon-coloured velvety mucosa appears over the Z-line in either circumferential or tongues-like form. Such a diagnosis should be confirmed with histological means that detect metaplastic mucosa in the lower oesophagus (Garud et al., 2010). Figure 2.1 shows endoscopies for a healthy person and Barrett's oesophagus patients and their corresponding histology microscopies.

#### **2.1.4 Dysplasia in Barrett's oesophagus**

Cancers in Barrett's oesophagus patients go through different phases of genetic and epigenetic changes in nature. Such changes result in activating oncogenes, silencing tumour suppressor genes, and freeing cells from the controls of healthy growth. Before an individual's cells turn cancerous, the DNA strange alterations could affect the oesophagus histologically, and the effect has ranging levels of severeness. That is known as dysplasia by pathologists. The first type is LGD, in which some alterations appear in several cells, yet they are not of great seriousness as such a problem could disappear; nevertheless, it could worsen over time. The second type is HGD, in which cells go through profound changes that could lead to cancers; therefore, treatment must be given to fight those cells (Shaheen and Richter, 2009).

Dysplasia is the neoplastic epithelium that grows and groups under the gland's basement membrane (Rice et al., 2005). Abnormal changes in dysplasia conditions are considered the second phase that follows metaplasia and precedes carcinoma. Dysplasia has to be diagnosed regularly by taking a biopsy from the patient for endoscopic use. Histologically, diagnosing dysplasia requires checking for abnormalities on both the architectural and cytological levels. Checking for architectural

abnormalities requires the consideration of glandular distortion and crowding, the possibility of papillary extensions in the gland lumen, and finally, the mucosal surface's villiform configuration.

On the other hand, nuclear changes should be considered when checking for cytological abnormalities. Such changes could be that the nuclear or nucleoli get more extensive and change in shape, resulting in an increased nuclear to cytoplasmic ratio and hyperchromatism and an increase in abnormal mitoses. The majority of pathologists believe it is better to consider the mucosal surface to confirm the diagnosis of dysplasia. Dysplasia has different ranging grades specified depending on the severity of changes in a case (Flejou, 2005).

Dysplasia can evolve into an invasive carcinoma when it spreads from the mucosa to the submucosa or even deeper. When abnormalities occur in a gland or lymph, it will be referred to as adenocarcinoma, which rarely appears in Barrett's oesophagus cases (Haggitt, 1994). Dysplasia in the gastrointestinal tract or Barrett's oesophagus cases can be categorised according to two different classification systems used globally: the inflammatory bowel disease (IBD) and the Vienna classification (Odze, 2006). The first classification, IBD, which is more prevalent in the United States, has three different results which are negative for dysplasia, positive for dysplasia (PFD) (either being of HGD or LGD), and indefinite to dysplasia (whether being dysplastic or inflammatory) (Riddell et al., 1983). The second classification system, the Vienna classification, is applied in several European countries besides many far Eastern ones, yet this system is not popular in the United States (Odze, 2006). Both systems are somehow alike; however, the Vienna one replaces the term "low/high-grade dysplasia" with "non-invasive low/high-grade neoplasia". Moreover, the second classification expands its non-invasive high-grade neoplasia to include three subclasses when tissue invasion is observed cytological or architectural: "high-grade adenoma/dysplasia", "non-invasive carcinoma" and "suspicious for invasive carcinoma". This system has a fifth category which is "invasive neoplasia" that includes IMC, submucosal carcinoma or beyond (Schlemper et al., 2000) (Table 2.1 for the category of the two classifications).

**Table 2.1** Comparison of different classification systems of Barrett's associated dysplasia

Severity degree	(a)	(b)	(c)	(d)	(e)
NFD	NFD	NFD	1-NFD (Barrett's only)	Negative	Category 1: Negative for neoplasia/ dysplasia
			2-Atypia, probably negative for dysplasia	Indefinite: 1-probably negative ( probably inflammatory)	Category 2: Indefinite for neoplasia/dysplasia
LGD	LGD	LGD	3-Atypia, probably positive for dysplasia	Indefinite: 2-probably positive ( probably dysplastic)	
			4-LGD	Positive: LGD	Category 3: Non-invasive low grade neoplasia (low grade adenoma/ dysplasia)
HGD	dysplasia	HGD	5-HGD	Positive: HGD	Category 4 Non-invasive high-grade neoplasia: 1-High-grade adenoma/dysplasia
					Category 4 Non-invasive high-grade neoplasia: 2-Non-invasive carcinoma (carcinoma in situ)
					Category 4 Non-invasive high-grade neoplasia: 3-Suspicion of invasive carcinoma
Cancer			6-IMC		Category 5 Invasive neoplasia: 1-IMC
	Not available	Not available	Not available (adenocarcinoma)		Category 5 Invasive neoplasia: 2-Submucosal carcinoma or beyond

(a) Modified classification (two groups), (b) Modified classification (three groups), (c) The available classification for the thesis dataset classified by Pathologists, (d) IBD study group and (e) Vienna Classification of Dysplasia



Being one of the types of cancer, carcinoma is a condition in which cancer affects the cells in the skin or the tissue lining organs. Those cancerous cells may divide and increase in number out of control and possibly spread to different places in an individual's body. Therefore, carcinoma is classified according to the spread of cells. The first classification is Non-invasive carcinoma (also known as carcinoma in situ), and it applies to cases when cancerous cells remain in the same place of their formation without spreading to other places in the body (e.g., remaining in the epithelial layer when they start in Barrett's oesophagus). The second classification is the IMC, and it applies to cases in which carcinoma penetrates the basement membrane of the glands into the lamina propria. Nevertheless, it does not affect the muscularis mucosa or the submucosa. However, the third classification applies when the submucosa is affected, which is submucosal carcinoma (i.e., invasive carcinoma). Finally, when those cells spread to the lymph nodes and glands in the lamina propria or submucosa, IMC and "submucosal adenocarcinoma" are used for classification. Figure 2.5 shows the invasion of different types of carcinoma. (Washington, 2010) and (Odze, 2006).

The American Joint Committee on Cancer (AJCC)

		Case 1				Case 2				Case 3				Case 4			
		GX	G1	G2	G3	GX	G1	G2	G3	GX	G1	G2	G3	GX	G1	G2	G3
Epithelium		0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1. Mucosa: Lamina Propria		IA	IA	IB	IB	IIB	IIB	IIB	IIB	IIIA	IIIA	IIIA	IIIA	IVA	IVA	IVA	IVA
	Muscularis Mucosa	IA	IA	IB	IB	IIB	IIB	IIB	IIB	IIIA	IIIA	IIIA	IIIA	IVA	IVA	IVA	IVA
2. Submucosa		IB	IB	IB	IB	IIB	IIB	IIB	IIB	IIIA	IIIA	IIIA	IIIA	IVA	IVA	IVA	IVA
	3. Muscularis Propria	IB or IIA	IB or IIA	IB or IIA	IB or IIA	-	-	-	-	IIIA or IIIB	IIIA or IIIB	IIIA or IIIB	IIIA or IIIB	IVA	IVA	IVA	IVA
4. Adventitia		IIA or IIB	IIA or IIB	IIA or IIB	IIA or IIB	-	-	-	-	IIIB	IIIB	IIIB	IIIB	IVA	IVA	IVA	IVA
Adjacent Tissues		-	-	-	-	IIIB	IIIB	IIIB	IIIB	IVA	IVA	IVA	IVA	-	-	-	-
Distal Lymph Nodes and/or Distal Organs		-	-	-	-	IVB	IVB	IVB	IVB	IVB	IVB	IVB	IVB	IVB	IVB	IVB	IVB

**Figure 2.5** Chart summarises the staging manual for cancer by the American joint committee

GX: The grade cannot be assessed (because of incomplete information), G1: Grade 1, G2: Grade 2 and G3: Grade 3. Case 1: cancer has not spread to nearby lymph nodes or distant sites. Case 2: cancer has spread to nearby lymph nodes but has not spread to distant sites. Case 3: cancer has spread to more than one nearby lymph node but has not spread to distant sites. Case 4: cancer has spread to nearby lymph nodes and distant sites.

Nowadays, only expert pathologists can classify Barrett's oesophagus dysplasia by examining glass slides with a microscope or virtual ones via computers. It is important to note that the Vienna classification's dysplasia categories are used for Barrett's oesophagus conditions. However, subjectivity sometimes affects the interpretation of the criteria set, and thus notable intraobserver and interobserver variability happens when dysplasia is graded. Usually, and since the categories of dysplasia range, controversies arise over some definitions, for example, LGD and "indefinite for dysplasia", LGD and HGD, and HGD and IMC. However, controversies are less common over definitions of terms at different ends of the grading scale. For instance: "Negative for dysplasia" (NFD) and HGD (Grin and Streutker, 2014). Both classification systems, IBD and the Vienna ones, have a low-level interobserver agreement, which means that two different observers have different interpretations of the same case (Odze, 2006).

On the other hand, intraobserver disagreement refers to a case in which the same observer provides different interpretations when examining the same case at two different times. Fleiss kappa statistic measures observations' agreement (Salomao et al., 2018). Most cases of interobserver disagreement occur in cases of epithelial lesions as confusing features appear, whether for reactive lesions (where abnormal alterations occur due to repair effect) or LGD. The same abnormal changes in response to injury could affect the healthy tissue in these two cases. Nevertheless, both systems have high agreement among observers over clinically relevant high-grade lesions and carcinoma (Odze, 2006) (Eleftheriadis et al., 2014)

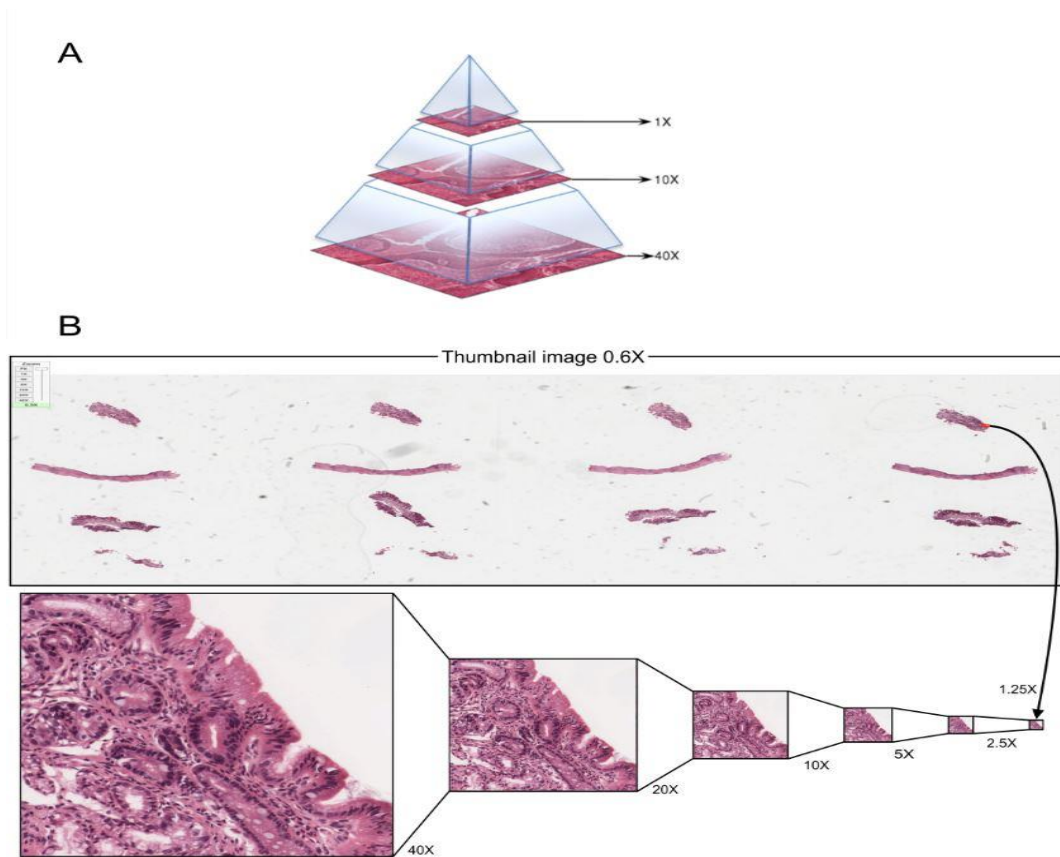
### **2.1.5 Virtual slides**

Visual microscopic inspection of tissue specimens from patients can be achieved using digital pathology images (e.g. WSIs) taken when microscope glass slides are scanned. Such images assist in conducting healthcare-related research and diagnosing patients (Wang et al., 2012). It is vital to have those images in high resolution and excellent colour depth to prepare them for research and diagnostic purposes. An image resolution is measured by microns per pixel; however, an image's colour depth is measured by the number of bits per pixel, identifying the number of various colours in the image. When a WSI is scanned with a magnification of X40, such an image will have almost a 0.25  $\mu\text{m}$  per pixel resolution and a colour

depth of 24 bits; therefore, one mm<sup>2</sup> area of the slide will have information of 384 million bits, and this in return will lead to having a 48 MB file taking into consideration that data were not addressed to be managed efficiently yet. Once the whole slide or more of the one z-plane is scanned, the size will increase, leading to difficulty in storing such data and having an obstacle with its bandwidth. This challenge shall be addressed to ensure that a user has a smooth experience (Zarella et al., 2018).

A WSI's size often reaches up to 1 GB, and when digital technologies for collaboration are used, downloading such a significant amount of data might not be allowed. The memory loading of that kind of data to view it might hinder. This matter was addressed by identifying how the field of view prominently relates to the image scale. In cases of large fields of view, computer screens limit the level of resolution; thus, it is not necessary to have a high-level resolution. However, only a small part of the field of view appears on the screen when tissue is inspected with strong magnification; therefore, the whole image is not necessarily loaded. Such restrictions when inspecting an image give a chance to enable a better image viewing experience for users (Zarella et al., 2018).

Usually, WSIs are sorted in an image pyramid representation that provides different resolution versions of that image that are arranged as a set of tiles. This technique facilitates how different resolutions are retrieved. The image at the bottom (i.e., the baseline one) is the image with the highest resolution version. When scanned for diagnostic purposes, WSIs get to be large as they often have 100,000x100,000 pixel sizes (Wang et al., 2012). It is illustrated in Figure 2.6 that when Aperio Scanscope whole slide scanner is used to scan a sample by a magnification of X40, it provides the same image at different resolutions, gradually decreasing to X10, X2.5, and X1.25. This scanner also provides a thumbnail image in which the whole sample is represented in a frame of 1 megapixel (Zarella et al., 2018). A window of two dimensions is provided by such a structure in which the area of interest is illustrated, for example, a tumour, a pseudopalisade and a distinguished nucleus (Wang et al., 2012). In most cases, the captured images are stored in one file even though it is unnecessary (Zarella et al., 2018).



**Figure 2.6** (A) the pyramid structure of a virtual pathology slide, (B) different zoom level representations for the same part of virtual pathology tissue

It is, to some extent, difficult to display those images using the standard tools in which a user can expand the compression of images file into RAM or swap. However, when “OpenSlide” is used, which is a “C” library, WSIs that are formatted in various ways can be explored easily with the friendly interface of “OpenSlide”. “OpenSlide” provides an interactive experience during navigation (OpenSlide, 2013).

### **2.1.6 Pathology of different degrees of dysplasia**

Histologically, there are two kinds of irregularities on which the degree of dysplasia is based, which are cytological and architectural abnormalities. The degree of irregularities determines whether Barrett’s Oesophagus Dysplasia is low or high (Haggitt, 1994). The cytological and architectural changes’ criteria are briefly demonstrated in Table 2.2. In Table 2.3, the degree of changes according to experts in pathology is illustrated (Flejou, 2005), (Haggitt, 1994), (Montgomery, 2005), (Odze, 2006) and (Spechler, 2002) specifying the difference in the degree of changes among different dysplasia stages.

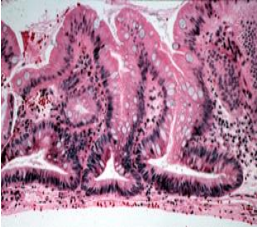
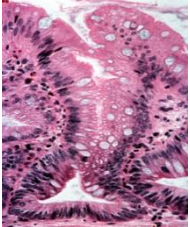
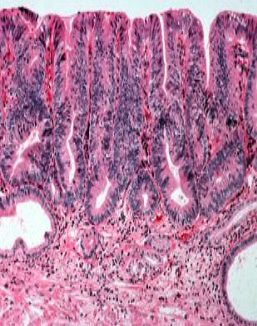
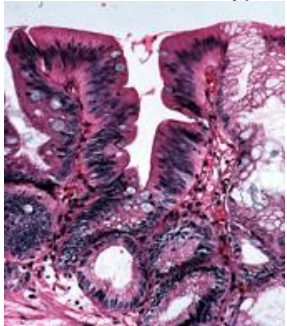
**Table 2.2** Cytological and architectural abnormalities in Barrett's oesophagus associated dysplasia

Condition	Normal oesophagus	Barretts's oesophagus	LGD	HGD	
		NFD or probably NFD	probably positive for dysplasia or LGD	HGD	IMC
<b>Cytological</b>					
Increase nuclei/cytoplasm ratio	-	-	+	++	+++
Loss of polarity	-	-	+/-	++	+++
Mitosis	-	+/-	+	++	++
Atypical mitosis	-	-	+/-	+	++
Full-thickness nuclei stratification	-	-	-	+	++
Hyperchromasia	-	-	+/-	++	+++
Multiple nucleoli	-	-	+/-	+/-	+
Large irregular nuclei	-	-	-	+/-	++
Irregular nuclei contour and variation of size	-	-	+	++	+++
Irregular cell size and shape	-	-	+/-	+	++
Necrosis/desmoplasia	-	-	-	-	+/-
Cell maturity	++	+	+/-	-	-
Glandular	-	+	++	+++	+++
Loss of mucin production	-	-	+	++	+++
<b>Architectural</b>					
Villiform change	-	+	+	++	++
Crypt budding/branching	-	-	+/-	+	++
Crowded (back-to-back) crypts	-	-	+/-	+	++

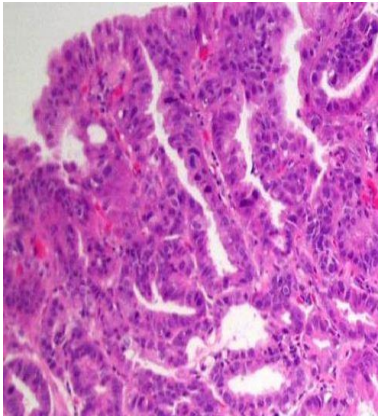
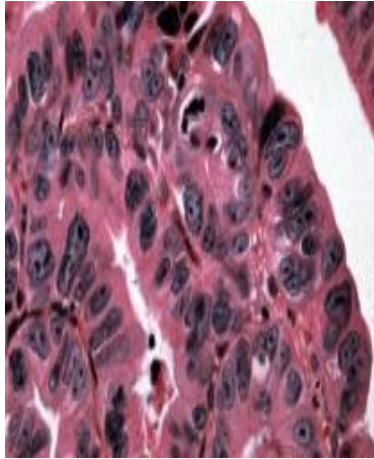
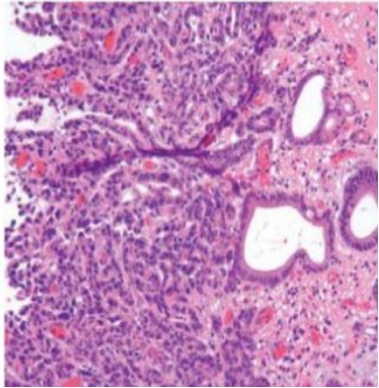
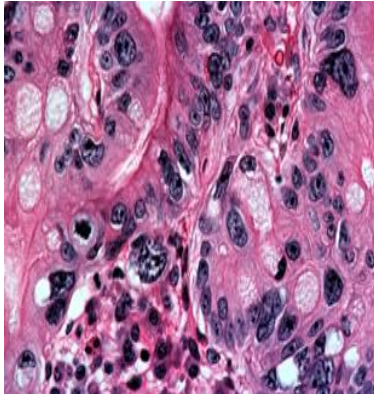
Irregular crypt shapes	-		+/-	++	+++
Crypts breach muscularis mucosa	-	-	-	-	+
Crowded glands	-	-	+/-	+	++
Intraluminal papilla/ridges		++	+	+	++
Lamina propria between glands	+++	++	+	-	--
Existence of infiltration	-	-	-	-	+/-

-: absent, +/-: might be present, +, ++ and +++: always present with different degrees.

**Table 2.3** Explanation of the cytological and architectural changes of each grade in dysplasia

		Histology	
	Architectural features		Cytological features
NFD	<p>preserved crypt architecture</p> 	<p>Preserved or relatively preserved cytological features:</p> <ol style="list-style-type: none"> <li>1-basally located nuclear</li> <li>2-regular</li> <li>3-mature</li> <li>4-oval or round shape</li> </ol> 	
LGD	<p>Relatively preserved crypt architecture</p> 	<p>The cytoplasm is generally mucin depleted</p> <p>Changes in the cytological features:</p> <ol style="list-style-type: none"> <li>1-Nuclear enlargement and elongation</li> <li>2-hyperchromasia</li> <li>3-irregular nuclear contours and a dense chromatin pattern either with or without multiple, small inconspicuous nucleoli</li> <li>4-Increased N/C ratio</li> <li>5-Nuclear stratification limited to the basal half of cell cytoplasm</li> <li>6-Preserved or only mild loss of nuclear polarity</li> <li>7-Increased mitoses, usually limited to crypts</li> <li>8-Few, if any, atypical mitoses limited to crypts</li> </ol> 	



<p>HGD</p>	<p>Changes in the crypt architecture: Irregular size and shape of crypts, crowded crypts, and intraluminal budding or cribriform</p>		<p>A higher degree of changes in the cytological features:</p> <ol style="list-style-type: none"><li>1-Nuclear enlargement</li><li>2-Full-thickness nuclear stratification</li><li>3-Mild to marked nuclear pleomorphism</li><li>4-Irregular nuclear contours with multiple, large nucleoli</li><li>5-Prominent loss of nuclear polarity</li><li>6-Mitoses on the epithelium surface</li><li>7-Increased number of atypical mitoses</li><li>8-little or no mucin cap or visible cytoplasm at the most luminal aspect of the cell</li></ol>	
<p>IMC</p>	<p>A higher degree of changes in the crypt architecture: Irregular size and shape of crypts, crowded crypts, intraluminal budding or cribriform, and the crypts may show little or no intervening lamina propria prominent back-to-back gland pattern</p>		<p>The highest degree of changes in the cytological features:</p> <ol style="list-style-type: none"><li>1-contain cells that are more epithelioid or cuboidal-shaped</li><li>2-high N/C ratio</li><li>3-round or oval highly irregular-shaped nuclei</li><li>4-an open chromatin pattern and prominent nucleoli</li></ol>	



### **2.1.6.1 Negative for dysplasia (NFD) in Barrett's oesophagus (metaplasia)**

The term NFD is given as a diagnosis for cases in which metaplastic columnar epithelium appears in ordinary and regenerative cases. When conducting histology on Barrett's Oesophagus cases that do not show inflammation, a flat mucosal surface is seen, and sometimes a villiform surface can be recognised. Moreover, according to the British Society of Gastroenterology, the epithelial layer could include columnar cells and goblet cells could appear. Goblet cells contain acid mucin; such mucin positively stains with Alcian blue. The columnar cells in the middle of goblet cells could be similar to normal gastric foveolar (Haggitt, 1994). The nuclear is of an oval or round figure with a basal location, being regular and mature (Odze, 2006).

Sometimes, epithelial regenerative abnormalities could be acute, especially in the mucosa next to the neo-squamocolumnar junction, or it may be so when there is an active inflammation or ulceration. In such cases, a deficient degree of cytological abnormalities might affect the newly generated epithelium. These changes include a slight increase in the nuclei/cytoplasm ratio, an increase in typical mitoses, hyperchromatic, slight loss of polarity, and pleomorphism. However, the nuclear stratification is retained with nuclei having standard size but prominent nucleoli (Odze, 2006). The tissue in NFD is expected to keep the regular architectural features of their crypts. Still, a considerable degree of crypt budding, branching, and distortion is accepted in inflamed areas (Odze, 2006).

### **2.1.6.2 Indefinite for dysplasia**

According to Grin and Streutker (2014), the term "indefinite for dysplasia" is not considered a degree as much as an indication that the biopsy diagnosis is extremely difficult. That is due to either technical factors like crush artefact, poor orientation and small biopsies or a severe inflammation that results in cytological changes in the epithelial layer that confuse the observer with LGD. Also, cases where changes similar to dysplasia are limited to the crypts' bases with the mature surface, are categorised (Odze, 2006). Generally, this category is a provisional diagnosis and not a type of dysplasia. Patients diagnosed with it are recommended to contact their clinic to plan further biopsy examination (Naini et al., 2016).

### **2.1.6.3 Low-grade dysplasia (LGD)**

LGD is the most common type in Barrett's oesophagus patients (Grin and Streutker, 2014). LGD cytological changes include: nuclear enlargement, as a nucleus may reach two to three times and three to four times the size of lymphocytes at the surface and the crypts, respectively, nuclei have one or multiple small nucleoli, stratification with a possible slight loss in polarity (the nucleus has a perpendicular orientation to the basement membrane) but generally the nuclear polarity is preserved, nuclei accumulate upon the mucosa surface leading to a loss in the surface maturation; still the maturation is preserved in deep crypts, nuclei hyperchromasia, nuclei irregular contour, and nuclei elongation (Naini et al., 2016) (Odze, 2006). Besides, the number of typical and atypical mitosis is increased at the crypts of the epithelium, the cytoplasm has a lower supply of mucin, and the goblet cells rarely appear. The architectural features are retained at the surface but might show some abnormalities in the crypts, such as crowded crypts with visible lamina propria separating the affected crypts. However, complex budding or angulation crypts are not expected to be witnessed in LGD tissues. In other words, a low degree of nuclear abnormalities affects the surface and increases at the crypts, where the dysplasia starts, with minor changes in the architectural features (Naini et al., 2016).

### **2.1.6.4 High-grade dysplasia (HGD)**

In HGD, the cytological abnormalities degree increases significantly, and they are not limited to the base of the crypts. These abnormalities spread to the epithelium surface with larger round nuclei that contain multiple prominent nucleoli. Both surface and crypts lose their nuclear polarity and have nuclear pleomorphism. In addition, more mitoses that are atypical appear. The cytoplasm disappears at cells close to the lumen, leaving the epithelium with no mucin cap because of nuclei stratification to the cytoplasm surface. Cytological changes are coupled with significant architectural abnormalities such as crowded (back-to-back) crypts, lack of lamina propria separation between the crypts, dilated glands, intraluminal budding, and villiform surface configuration. In cases where tissues have any architectural abnormalities and a lower degree of cytological abnormalities, they are considered HGD (Naini et al., 2016). In many cases, the cytological abnormalities are adequate in diagnosing HGD.

#### **2.1.6.5 Intramucosal carcinoma (IMC)**

When the dysplastic changes invade the epithelium's basement membrane and start to develop in the lamina propria or muscularis mucosa, invasive adenocarcinoma is diagnosed. Figure 2.5 shows different degrees of IMC. IMC patients should be treated more aggressively than HGD patients undergoing oesophageal resection. Thus, discriminating between the two categories is clinically crucial in deciding on further treatments. However, some clinics manage non-adenomatous dysplasia, similar to HGD. IMC has the highest degree of dysplastic abnormalities on a cytological and architectural basis, and the abnormalities are not limited to the epithelial layer but deeper layers (Odze, 2008).

#### **2.1.6.6 Summary**

Tissues that are diagnosed with NFD retain both cytological and architectural features. In some cases of regenerating epithelium, some cytological abnormalities may occur. Still, they never reach dysplastic features, which are nuclear pleomorphism, loss of cell polarity, a significant increase in nuclei to cytoplasm ratio, loss of surface maturation, and low mucin in the cytoplasm (Odze, 2006).

Also, to differentiate LGD from HGD, the latter category has some features that cannot be found in LGD, like full-thickness nuclear stratification, loss of cell polarity at the epithelium surface and higher crypts, atypical mitoses, or the architectural abnormalities that were discussed in the previous section (Odze, 2006).

Finally, to distinguish between HGD and IMC, when abnormalities are confined in the epithelial layer and not spread to any further layers, it is HGD; otherwise, it is diagnosed with IMC.

#### **2.1.7 Clinical challenges**

According to Haggitt (1994), there are many limitations in the systems of grading dysplasia in Barrett's oesophagus. These limitations are:

- Western and Eastern pathologists tend to have their unique criteria for grading dysplasia. In grading dysplasia for Barrett's oesophagus,

pathologists worldwide disagree on the criteria. For example, in distinguishing HGD from IMC, Japanese pathologists focus on cytological abnormalities to detect carcinoma. In contrast, western pathologists are content with invasion into the lamina propria to diagnose IMC (Naini et al., 2016).

- Some pathologists depend on morphological features that are not scientifically approved. For instance, in grading IMC, Western pathologists rely on dysplastic abnormalities to invade the lamina propria but not reach the muscularis mucosa. Those criteria have not been approved yet for lamina propria invasion (Naini et al., 2016).
- The effect of inflammation in non-dysplastic tissue that mimics dysplasia abnormalities
- Interobserver and intraobserver disagreement in diagnosing dysplasia
- Difficulties in diagnosing cases between LGD and HGD, and cases between HGD and IMC

#### **2.1.7.1 Intraobserver and interobserver variation in diagnoses**

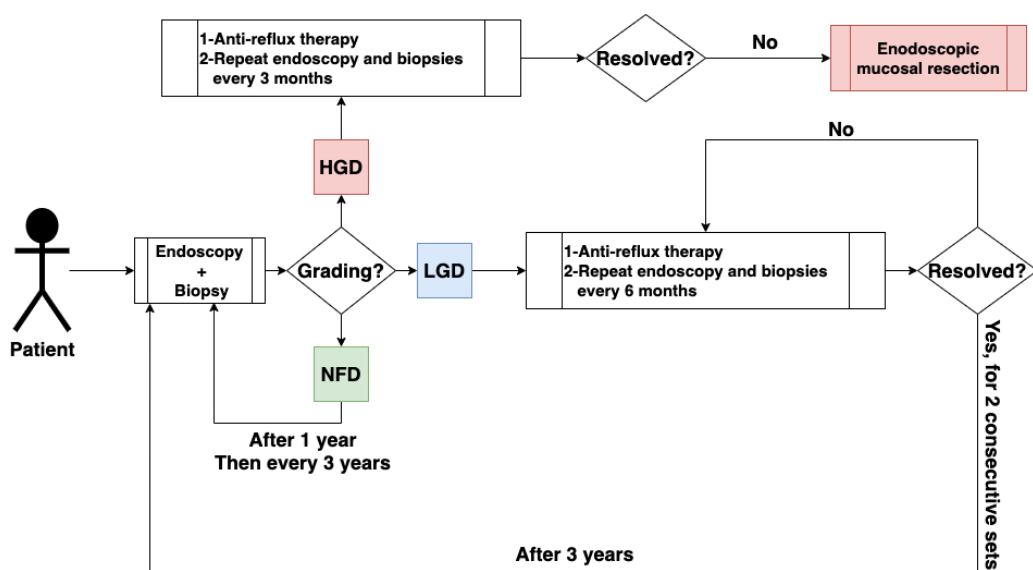
Grading dysplasia in Barrett's oesophagus is based on cytological and architectural abnormalities that follow a continuous spectrum from normal to low and high degrees of abnormalities. Thus, it suffers from a high intraobserver and interobserver disagreement due to the absence of sharp definitions of the boundaries that separate each category from the following higher category or the previous lower category. The highest level of disagreement occurs in the indefinite dysplasia versus LGD interface, even amongst experienced gastrointestinal pathologists, as most NFD cases are over-diagnosed (Odze, 2006). While at the higher end of the spectrum (at HGD versus IMC), the disagreement is at the lowest, as some HGD diagnoses are often downgraded after expert review (Naini et al., 2016). Fortunately, this agreement is concerning the most because the results of these diagnoses help in the decision of oesophagus resection (Haggitt, 1994).

In general, a study focused on finding the interobserver agreement for grading different grades of dysplasia among experienced gastrointestinal pathologists concluded that the agreement was moderate (Coco et al., 2011). Also, Bhat et al. (2011) conducted a study on eight experts to distinguish NFD from "indefinite for dysplasia" and LGD, and they found that

they have 40% disagreement. Moreover, (Odze, 2006) found that expert pathologists reached a “fair” agreement when they tried to detect LGD and a “slight” agreement for detecting “indefinite dysplasia”.

### 2.1.7.2 Costs of misclassifying dysplasia in Barrett's oesophagus

Oesophageal cancer is the sixth deadliest cancer worldwide, with less than 10% of patients who can survive for five years (Delpisheh et al., 2014). Downgrading the diagnosis of those patients increases the risk of developing adenocarcinoma and may lead to disqualifying them from life-sustaining treatment. For instance, downgrading a patient's diagnosis with LGD will increase the risk of progressing adenocarcinoma five times higher than the precautions for the mistaken NFD (Bird–Lieberman et al., 2012). Also, downgrading HGD will prevent the patient from having a necessary near-future mucosal resection to eliminate pre-cancerous tissue before it develops into cancer (Spechler, 2007). However, upgrading NFD to LGD will add a burden to the healthcare system as LGD is the worst grade when it comes to cost. LGD can be confirmed only by two gastrointestinal pathologists, and the diagnosed patient is offered surveillance every six months. Whereas upgrading LGD to HGD will lead to unnecessary surgical intervention (Vladimirov et al., 2013). Figure 2.7 provides the recommended treatment plan for patients with Barrett's oesophagus by the Practice Parameters Committee of the American College of Gastroenterology.



**Figure 2.7** Recommended treatment plan for the patients with Barrett's oesophagus by the Practice Parameters Committee of the American College of Gastroenterology

### **2.1.8 Colour normalisation**

Colour variations could be introduced to histological images during the biopsies' preparation, staining, and digitalisation. For instance, the variation of stain concentration, the different staining times, the pH of the solutions, and the use of different scanners (Tosta et al., 2019). This colour variation imposes obstacles to the process of analysing histological images. Thus, stain normalisation algorithms have been developed to overcome this issue. Employing colour normalisation techniques is covered by important studies in the literature. Generally, Shaban et al. (2019) classified those approaches under three classes: colour matching, stain-separation, and pure learning-based approaches. The colour matching based methods focused on matching the colour spectrum of a processed histological image to a target template image. For example, Reinhard et al. (2001) aligned the colour distribution of an image to a reference using a linear transformation. According to Shaban et al. (2019), the disadvantage of such methods is that one transformation is used for all the images regardless of the contribution of each stain dye to the concluding colour. The second class of methods is the stain-separation methods that separately apply normalisation on staining channels. For instance, Khan et al. (2014) proposed a nonlinear approach: a stain matrix estimator that employs a colour classifier to classify each pixel in the image to the related stain component. Finally, the pure learning-based approaches offer solutions considering it as a style-transfer problem. Shaban et al. (2019) introduced the StainGAN model based on Generative Adversarial Networks and was trained end-to-end. That solution does not require a reference template slide that usually is picked by an expert.

## **2.2 Related works in diagnosing dysplasia in Barrett's oesophagus**

This section discusses related works that use machine learning and deep-learning approaches to diagnose or detect dysplasia in Barrett's oesophagus. Although comprehensive literature in this field is available, it was mainly focused on the endoscopic and volumetric laser endomicroscopy images. Relying on the outcomes of our search using "Web of Science" and "Google Scholar" using the keywords "deeplearning", "deep learning", "machine learning", "Barrett's oesophagus", "Barrett's oesophagus", and "histology", there were five related works. (Adam et al., 2011) and (Adam et al., 2012) were conducted by a previous PhD. researcher at the University of

Leeds uses an unsupervised machine learning approach on the same dataset used in this research to classify whole virtual slides into the three-tier dysplasia grades. Their approach randomly sampled images across the whole virtual slides, converted them into grey-scale images, and calculated their grey-level co-occurrence matrices (GLCM) (Haralick et al., 1973). After that, GLCM features, namely contrast, energy, correlation and homogeneity, were calculated in four directions for each patch within the sampled image. Then, the calculated features were clustered using k-means into five clusters. Based on the clustering results, cluster coded images (CCI), known as heatmaps, were generated for the images and the spatial relationships for the texture features were calculated using cluster co-occurrence matrices (CCM), in a way similar to GLCM, to produce a prediction for the image using random forest and decision tree classifiers. Their approach achieved a 77.8% accuracy and a 0.54 Kappa Value (KV).

Kandemir et al. (2014) and Kandemir et al. (2015) proposed a weakly supervised machine learning model to solve the MIL problem in Barrett's oesophagus H&E stained images. They aimed to detect cancer in those images by tiling the biopsy and generating a feature vector for each tile. The set of nature vectors is considered instances in the whole slide bag. They employed the mi-Graph proposed by Zhou et al. (2009) to predict the label of the whole slide. They managed to achieve an accuracy of 87% and a 0.93 AUC at the bag-level and an 82% accuracy and a 0.89 AUC at the patch-level.

The previously mentioned studies proposed conventional machine learning approaches to develop the CAD system. Unlike Tomita et al. (2019), who proposed a weakly supervised deep-learning approach to classify tissues into "healthy oesophagus", "negative for dysplasia Barrett's oesophagus", "positive for dysplasia Barrett's oesophagus", and "adenocarcinoma". Their model processes biopsies in two stages. In the first stage, they divided each whole virtual image into  $r \times c$  grid. Then, they fed the grid cells into a CNN to produce feature representations assembled to build a grid-based feature representation with the size of  $r \times c \times k$ , where  $k$  is the length of the learnt vector representation. In the second stage, they applied 3D convolution with  $k \times d \times d$  sized kernel to generate  $r \times c$  size attention map  $\alpha$  that every row and column represents the weight value (the importance) of the corresponding grid cell. Finally, they computed the whole-slide global feature

vector using the dot product of the output of the first and second stages. The whole-slide global feature vector trains a fully connected layer and Softmax to predict the whole slide label. That approach is considered as MIL pooling on scores, and it achieved a 0.49 recall, a 0.93 specificity and an 87% accuracy in predicting negative and positive for dysplasia.

### 2.3 Performance metrics for medical tasks

Selecting the evaluation criteria is an essential step in evaluating the performance of any CAD system. Generally, the confusion matrix is calculated for assessing any classifier, which counts the classifier prediction against the actual classes. For this matrix, the error rates for each class are driven. True positive (TP) and true negative (TN) are the number of correctly predicted positive instances and negative instances, respectively. The false positive (FP) and false negative (FN) are the number of negative instances that are mislabelled as positive and the number of positive instances that are mislabelled as negative, respectively. Other metrics such as accuracy, precision, recall, fall-out, specificity, and F1-score can be calculated using the driven error rates, as shown in Table 2.4.

**Table 2.4** Summary of the performance metrics and their corresponding equations

Metric	Equation	Description
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	It represents the percentage of instances that are predicted correctly.
Precision	$\frac{TP}{TP + FP}$	It describes the ability of the classifier to classify instances correctly under a class as opposed to all instances which were correctly or incorrectly labelled with the same class.
Recall, sensitivity or true positive rate	$\frac{TP}{TP + FN}$	It describes the rate of instances classified correctly as positive amongst all of the same class.



Specificity or true negative rate	$\frac{TN}{FP + TN}$	It is the percentage of real negative instances that are predicted as negative.
F1-score	$2 \times \frac{Recall \times Precision}{Recall + Precision}$	It is the weighted average of recall and precision.
Fall-out or false positive rate	$\frac{FP}{FP+TN}$ or $1 - specificity$	It is the percentage of negative instances that are misclassified as positive.

Finally, descriptive statistics were used to evaluate the agreements between the proposed models and the pathologists, which are the Cohen kappa coefficient (KV) (Cohen, 1960), and the weighted Cohen kappa Coefficient (weighted KV) (Cohen, 1968). For computing KV, the first step is to find the confusion matrix of any two observers, as shown in Table 2.5. Then, calculate the observed agreement  $P_o$  and the expected agreement  $P_e$ , which is the agreement that occurs by chance, as shown in Equation 2.1 and Equation 2.2 with reference to labels provided in Table 2.5. Finally, KV is computed using Equation 2.3. KV is a robust statistic to measure the agreement that avoids the agreement by chance; however, it relies on the nominal categories and not the ordinal categories. For instance, in grading cancer into normal, low-grade and high-grade classes, KV will decrease the agreement score by a similar amount whether the two observers predict low and high grades or normal and high grades. In contrast, the weighted kappa computes the disagreement concerning the degree of a disagreement using a predefined table of weights. The strength of the agreement is categorised based on the KV score into “poor agreement” if it is less than zero or “slight agreement”, “fair agreement”, “moderate agreement”, “substantial agreement”, or “almost perfect agreement” if it falls in the following intervals [0.00-0.20], [0.21-0.40], [0.41-0.60], [0.61-0.80] or [0.81-1.00], respectively.

**Table 2.5** Showing the confusion matrix for observer1 against observer2

		Observer 1		
		Normal	Abnormal	total
Observer 2	Normal	TP	FN	$rm^1$
	Abnormal	FP	TN	$rm^2$
	total	$cm^1$	$cm^2$	$n$

TP and TN show the agreement between the observers, and FP and FN show their disagreement.  $n$  is the number of observed samples.

**Equation 2.1**

$$P_o = \frac{TP + TN}{n}$$

**Equation 2.2**

$$P_e = \frac{(cm^1 \times rm^1) + (cm^2 \times rm^2)}{n \times n}$$

**Equation 2.3**

$$KV = \frac{P_o - P_e}{1 - P_e}$$

**2.4 Deep-learning and its biological inspiration**

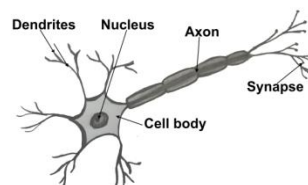
Deep-learning is considered one of the machine learning approaches, which employs neural networks in the architecture of non-linear and multiple layers. Learning distinct features from raw input data rather than designing detectors to extract them, which human experts accomplish, is the crucial feature of deep-learning. Moreover, the techniques used in learning the representation from raw data can detect its hierarchical representation as to the abstraction level increases (Xu et al., 2015). By way of illustration, learning the hierarchical representation in text recognition is represented by learning the words, clauses, sentences, and story.

According to (Deng and Yu, 2014), deep networks can be divided into three groups: the discriminative networks (supervised networks), the generative networks (unsupervised networks), and deep hybrid networks. This categorisation is based on the purpose of usage, such as generating

features, recognising objects, or classifying patterns. Discriminative networks are the deep networks utilised in supervised learning, where the input data are always labelled. This type of network is used for classification tasks. While generative networks are set to capture high-level features from the input data when there is a lack of information regarding the target class labels. The last type is the deep hybrid networks which almost always refer to the results of generative networks in discrimination. More details are provided in Chapter 2.

The biological feature of an organism to behave based on its perception of the environment was the inspiration for deep-learning. Between the action and perception, the collected information is processed. The nervous cells generate electrical signals in animals and humans to manipulate the collected information. On average, the human brain contains billions of nerve cells, of which each is connected to about 10,000 other neurons (Garrett, 2014).

The nervous cell consists of dendrites, cell body, axon, and synapses (see Figure 2.8). The dendrites are lengthy appendices that connect the cell with other cells to receive signals from them. Typically, signals travel from the cell body through the axon, causing action potential until it reaches the axon terminals. Other cells' dendrites receive these signals through a special connection or gap known as a synapse. When the post-synaptic cell receives the action potentials, it is either encouraged to fire action or prevented from firing an action. These cells are known as Excitatory synapses and inhibitory synapses, respectively (Garrett, 2014). The synapses are strengthened between two neurons if one of them repeatedly encourages the other to fire an action potential. That is a sign of the constant development of the nervous system. It is believed that the activity-dependent synaptic plasticity constitutes the underlying mechanism for learning (Hebb, 1952).



**Figure 2.8** A nervous cell

For example, the visual systems manifest the high ability and the high speed of the neural network. The experiment of Hubel and Wiesel demonstrates that. In 1959, they experimented on the vision system of a cat. They stimulated the receptive fields of the cells by presenting patterns, and they recorded activities of the primary visual cortex's cells. As a result, they found three types of cells: simple, complex, and hyper-complex. Simple cells fire mainly to edges and gratings of particular orientation as they have excitatory and inhibitory regions, which are well arranged in their almost rectangular receptive field. Similar to these cells are the complex cells; however, they are not sensitive to the position of the pattern as they have wider receptive fields than the simple cells. Also, some of them are fired by motion. Finally, the hyper-complex cells have the same features as the complex ones plus the sensitivity of the length of the stimuli; thus, they allow the perception of corners. These cells work dependently on each other so that the simple cells are connected to the complex cells, providing them with their inputs (Hubel and Wiesel, 1959).

## 2.5 Deep-learning architectures

This section briefly describes the most commonly employed deep-learning architecture in the computer vision field. Besides, it discusses the obstacles developers faced in their attempts to increase the architecture size and the proposed solution to overcome them. Finally, it overviews the used architecture in this research.

### 2.5.1 Convolutional Neural Network (CNN)

CNN is a multi-layer perceptron feed-forward neural network introduced to the computer vision community by LeCun et al. (1998). CNN adapts a successive feature extraction technique that learns the simple features from an input image in the earlier stages. Then, the complexity of learned features is increased in further stages. It consists of one or multiple stages of layers and a one-dimensional output layer. Each stage is a feature extractor and contains two to three two-dimensional layers. The first layer is **the convolutional layer**. It takes advantage of the main feature of images, which is the similarity of the corresponding statistics of different regions in the same image. It uses the concept of weight sharing to limit the number of neurones. The input of this layer could be the original image in the first stage

or the output of the other stages in the later stages (Abdel-Hamid et al., 2012). Each convolutional layer contains a filter bank, and each filter (weight) within the bank is a 2D array. The model learns filters through the processes of the forwards-propagation of the input and the back-propagation of the model's prediction error. The two strategies update those filters in the training phase until the model reaches convergence. In addition, each filter connects an input with a corresponding feature map by convolving the filter and the 2D input. During convolving the 2D filter and the input, the filter slides over each pixel or most of them, which can be specified by setting the stride parameter. The stride can be defined as the overlap pixels while applying the convolution operation. The convolution operation is achieved by summing the results from multiplying a filter's values by the corresponding pixels in the input. If the input image is an RGB image, then each filter will have three channels to convolve each image channel. Then the results of the three convolutional operations are accumulated, and a bias is added to the result. The convolution layer function is provided in Equation 2.4.

#### Equation 2.4

$$y = \alpha(x * W + b)$$

Where  $\alpha$  is the activation function,  $x$  is the input,  $y$  is the outputted representation feature,  $W$  is the weight set,  $b$  is the bias and  $*$  denotes the 2D convolution.

The second layer is **the activation function layer**, which gives the CNN two preferred properties. The nonlinear capabilities of the network and the production of zero-mean inputs for the next layers. The "Sigmoid" function is one of the most used functions with deep-learning architectures. This function transforms the real number to fit in the range [0,1]. As a result, every large negative number will be 0 and 1 for any large positive number. However, this function has two drawbacks: the activated neurones at 0 or 1 will not receive any signals because their gradients calculations will be almost zero when the network back-propagates the error and outputs a non-zero-centred output. That might introduce zigzag dynamics in updating the parameters when fed into the higher layers. The "Tanh" function could be used alternatively to overcome the second issue as its output range is [-1,1] (LeCun et al., 2012).

Recently, Rectified Linear Unit (ReLU) usage with deep-learning architectures increased. It behaves much better than the previously mentioned functions because it is cheaper, faster to compute, and increases the stochastic gradient descent speed convergence. This function outputs 0 if the input is negative or zero, or it outputs a number equal to the input number if it is a positive number. Only, it guarantees that it does not saturate for positive values because it rectifies most of its input to be a positive number. Unfortunately, its main disadvantage is that during the training, the flowing of large gradients may update the layer's parameters in a way that has negative numbers for the weighted sum of the input that will output 0. That will not allow any input to be activated later (Zeiler and Fergus, 2014). This problem is known as the vanishing gradient problem. To overcome this issue, Clevert et al. (2015) proposed a new activation function called the exponential linear unit (ELU) that behaves similar to ReLUs in the case of positive inputs. However, it saturates to the negative inputs as it has negative values, which pushes the mean of the activation to be close to zero. As a result, it decreases the learning time in deep neural networks and results in higher accuracies in classification tasks. ELU can be calculated using Equation 2.5.

**Equation 2.5**

$$f(x) = \begin{cases} \alpha(e^x - 1) & , x < 0 \\ x & , x \geq 0 \end{cases}$$

Where  $x$  is the input and  $\alpha$  is a stochastic variable sampled from a uniform distribution at training time. It is fixed to the expectation value of the distribution at the test time.

Moreover, it is optional to add a **pooling layer**, also known as a sampling operation. The pooling layer exploits the stationary property of images and aggregates statistics of features over regions of the image. The most common methods are mean-pooling and max-pooling, as they choose the mean features and winning features, respectively (Havaei et al., 2017).

Training deep neural network architectures is accomplished by the forward-propagation and back-propagation strategies. The forward-propagation is the process of feeding input data toward successive layers of the neural network to be processed by the convolutional, activation and pooling layers

until it reaches the classifier, which is usually a Softmax layer (see section 0 for more description). An error is computed for the classifier prediction by an objective function that will be discussed later. In contrast, the back-propagation decreases the error by computing the objective function's partial derivatives (gradients) concerning the network's parameters. The computed gradients update the network's parameter following the chain rule (LeCun et al., 1998). The learning process from the error involves different terminologies and techniques such as objective function, optimiser and learning rate.

In auto-encoders cases, **the objective function** calculates the model error from predicting the input's class or regenerating it, considering the ground-truth class of the input or the original input. The objective function is applied to all instances in the training dataset. In training the model, the goal is to reduce the sum of the resulting error. The cross-entropy error and categorical cross-entropy are the most common objective functions to train deep neural networks. The cross-entropy is discussed later in section 5.3.1.1, and it is used with binary tasks, while the categorical cross-entropy is used in multi-class tasks, as discussed later in section 4.3.1.

Optimising a model is the process of finding the optimal solution for a problem using one of **the optimisation methods**. Stochastic gradient descent (Robbins and Monro, 1951) is the most common algorithm to optimise deep-learning architectures. It is an iterative algorithm that aims to minimise the objective function and update the model's parameters in each iteration by subtracting the calculated gradients of the cost function with respect to the multiplication of the weights by the learning rate. In general, the previous gradients are not considered in updating the parameters, which may slow the learning process. Also, using the stochastic gradient descent optimiser to train a very large deep neural network can slow the training. Employing alternative algorithms can speed the learning for such networks. The most popular fast optimisers are momentum optimisation, the adaptive gradient algorithm (AdaGrad), root mean square propagation (RMSprop), and finally, adaptive moment estimation (Adam).

Contrary to the stochastic gradient descent, the stochastic gradient descent with momentum (Polyak, 1964) uses gradients as an acceleration rather

than a speed. This optimiser keeps the calculated gradients of each iteration in the momentum vector, and then it updates the weights by subtracting the momentum vector. Another disadvantage of the stochastic gradient descent algorithm is the inability of the early detection of the direction to the global optimum because it goes very fast along the steepest slope. Then, the pace slows down when the bottom of the valley is reached. This problem is addressed by the AdaGrad algorithm (Duchi et al., 2011), as it tends to head down toward the steepest slope. It also keeps the gradient vector scaled down. AdaGrad does not converge to the global optimum even though it detects its direction early. The RMSprop algorithm converges to the best solution by keeping track of the gradients from the latest iterations. RMSprop is an unpublished work; however, it is proposed by Hinton et al. (2012). Adam optimiser (Kingma and Ba, 2014) is an algorithm that combines the advantages of momentum and RMSProp algorithms. It uses the momentum vector and the tracked gradients from the most recent iterations to update the network parameters.

**The learning rate** represents the size of a step at each training iteration, while the optimisation algorithm seeks the minimum value of a loss function. Finding the ideal learning rate is a complex task. For instance, if a very small number is assigned to the learning rate, then training the model will converge after a long time. While if it is set to a very high value, then training the model will diverge. Also, a slightly high learning rate may show progress toward reaching the optimum initially, but after that, it will fluctuate around it without settling down (Géron, 2017). A good learning rate can be found empirically by running the experiments several times using different values for the learning rate; then comparing the learning curves. The corresponding learning rate for the curve that decreases faster and converges should be selected. Generally, fine-tuning models require a very small learning rate value.

Optimisation functions have different methods to update the parameters based on the amount of computed data in each iteration. The first method is **batch optimisation**, which updates the parameter set in each iteration after computing the gradients for the whole training set concerning the cost function. Batch optimisation costs time and memory capacity. Also, it is not used with huge datasets as they cannot be fit in memory, and it is hard to update the dataset while training the model (online) (Géron, 2017). In

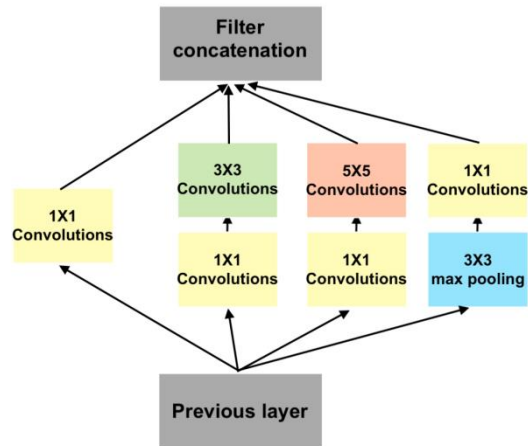


contrast, the **stochastic optimisation** that updates the model's parameters after each training instance makes updating the dataset online possible. Although this technique is fast in training, it suffers from the redundancy of information and convergence complexity (Géron, 2017). The **mini-batch optimisation** has the advantages of the previous methods. It overcomes the previous issues by dividing the dataset into small batches and updating the parameters once each mini-batch is trained (Géron, 2017).

### 2.5.2 “We need to go deeper!”

CNNs have achieved state-of-the-art performance in object classification, segmentation and detection tasks. For instance, they proved that by always being the winning architecture in the ILSVRC challenge (Russakovsky et al., 2015). The difference between the proposed architectures in the competition is their depth and width, as the performance is increased by increasing the network size. However, by increasing the width and depth, the computational cost increases dramatically, and by adding more layers, the performance starts to degrade due to the vanishing gradient issue.

Szegedy et al. (2015) proposed a solution to increase the network width and keep costs low. Their model managed to capture the optimal local sparse construction and detect its pattern in an input image by convolving different sizes of filters (1X1, 3X3 and 5X5) to the input. The output of those filters was then concatenated into an output vector that will be the input for the next layer in the model. Additionally, the output of a pooling operator (max-pooling with stride 2) on the input was appended to the output vector to benefit from the effect of the pooling operation. Additionally, dimension reductions and projections were applied before expensive filters and following the pooling layer to reduce the computational cost. The expensive filters are the ones with 3X3 and 5X5 pixel-sized. The best dimension reduction method that preserves the spatial sparsity representation and compresses the signals is achieved by applying 1X1 convolutions accompanied by rectified linear activation. The inception block structure is illustrated in Figure 2.9. The proposed model was named GoogleNet (also known as inception V1), and it won the ILSVRC 2014 competition with 6.67% top-5 error only.

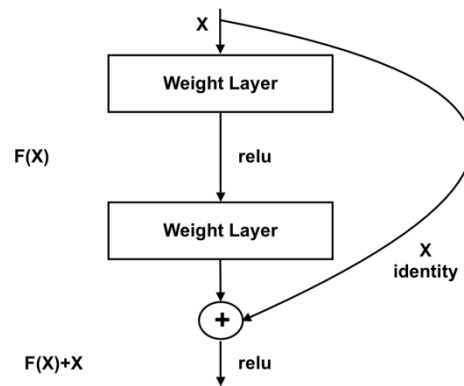


**Figure 2.9** A building inception block in GoogleNet (inception V1)

Furthermore, He et al. (2016) proposed another solution to increase the depth of the network without imposing the vanishing gradient problem. Before their proposed model, the conventional CNNs reached 30 layers in maximum before they fell into the vanishing gradient problem. He et al. (2016) proposed a deep residual network that is eight times deeper than VGG-16 (Simonyan and Zisserman, 2014), has better performance and yet has far lower complexity. Their architecture relies on the concept of adding more layers to increase the network's ability to discriminate different categories as the deep neural networks integrate variant levels of features starting from low-level features in the first part of the network, medium-level features in the middle part, and high-level features at the end. For example, layers in the first part of the network detect features like edges and corners and the activated map almost looks similar to the input image.

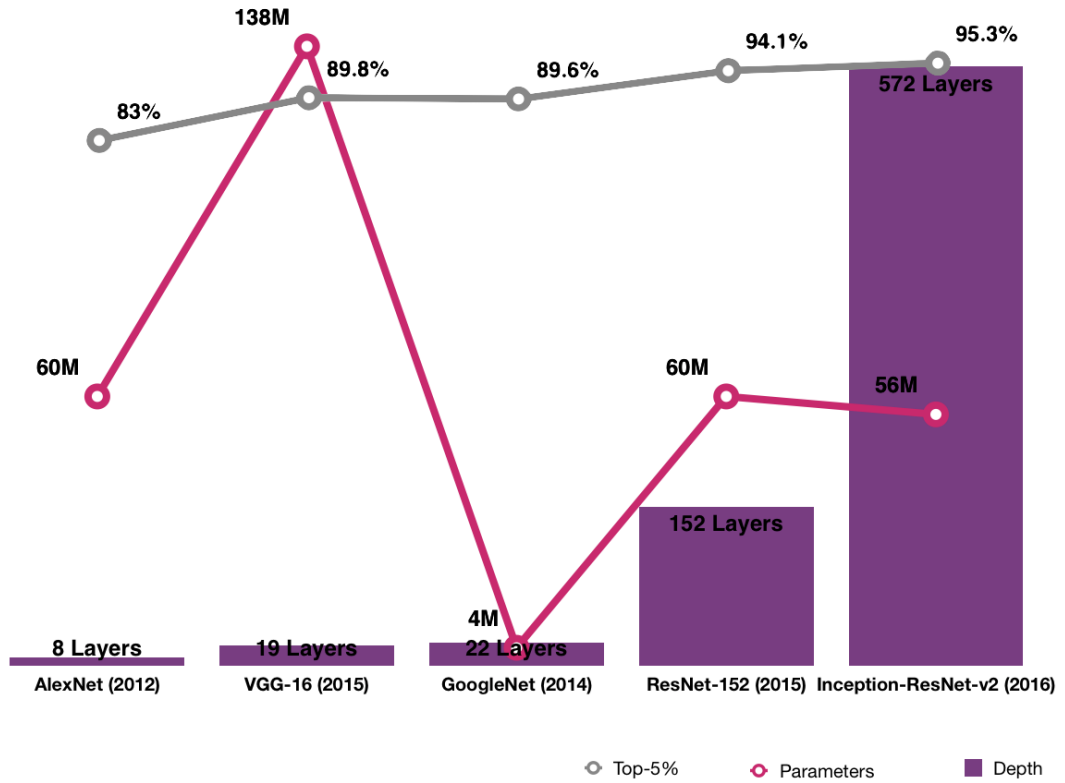
In contrast, the last layers detect higher-level features, and the outputted activation map is abstract and looks less or nothing like the related input. The concept of the residual network is to introduce a shortcut connection between the residual block's input and output, as shown in Figure 2.10, to allow the uncontrolled flow of the gradient. For adding the two layers, their dimensions should have the same size; thus, a zero-padding technique is used to adjust the output size. The mathematical representation of the residual block is the output of the block is  $y = F(x) + x$ , where  $x$  is the input to the block, and  $F(x) = (W_i * \sigma(W_{i-1} * x + b_{i-1}) + b_i)$ , in which  $W_{i-1}$ ,  $W_i$ ,  $b_{i-1}$  and  $b_i$  are the weights and biases for the first and second convolutional layers within the residual block, and  $\sigma$  denotes the ReLU activation function.

The proposed residual network won first place in the ILSVRC 2015 competition.



**Figure 2.10** A residual building block in ResNet architecture

Szegedy et al. (2017) proposed an architecture known as “Inception-ResNet-v2” that utilises the inception and the residual networks, as illustrated in Figure 2.12. The model scored the best performance in 2018. Figure 2.11 shows the evolution of different CNN architectures and how introducing the previously mentioned techniques has increased the performance and decreased the computation burden.



**Figure 2.11** A comparison between different famous CNN architectures

This figure compares the number of layers and parameters and the performance of AlexNet (Krizhevsky et al., 2012), VGG-16, GoogleNet, ResNet-152 and Inception-ResNet-v2 networks. It shows that 19 layers of conventional CNN (VGG-16) occupy the highest memory space. By introducing the residual and inception networks, the memory usage was dramatically reduced while the performance was increased.

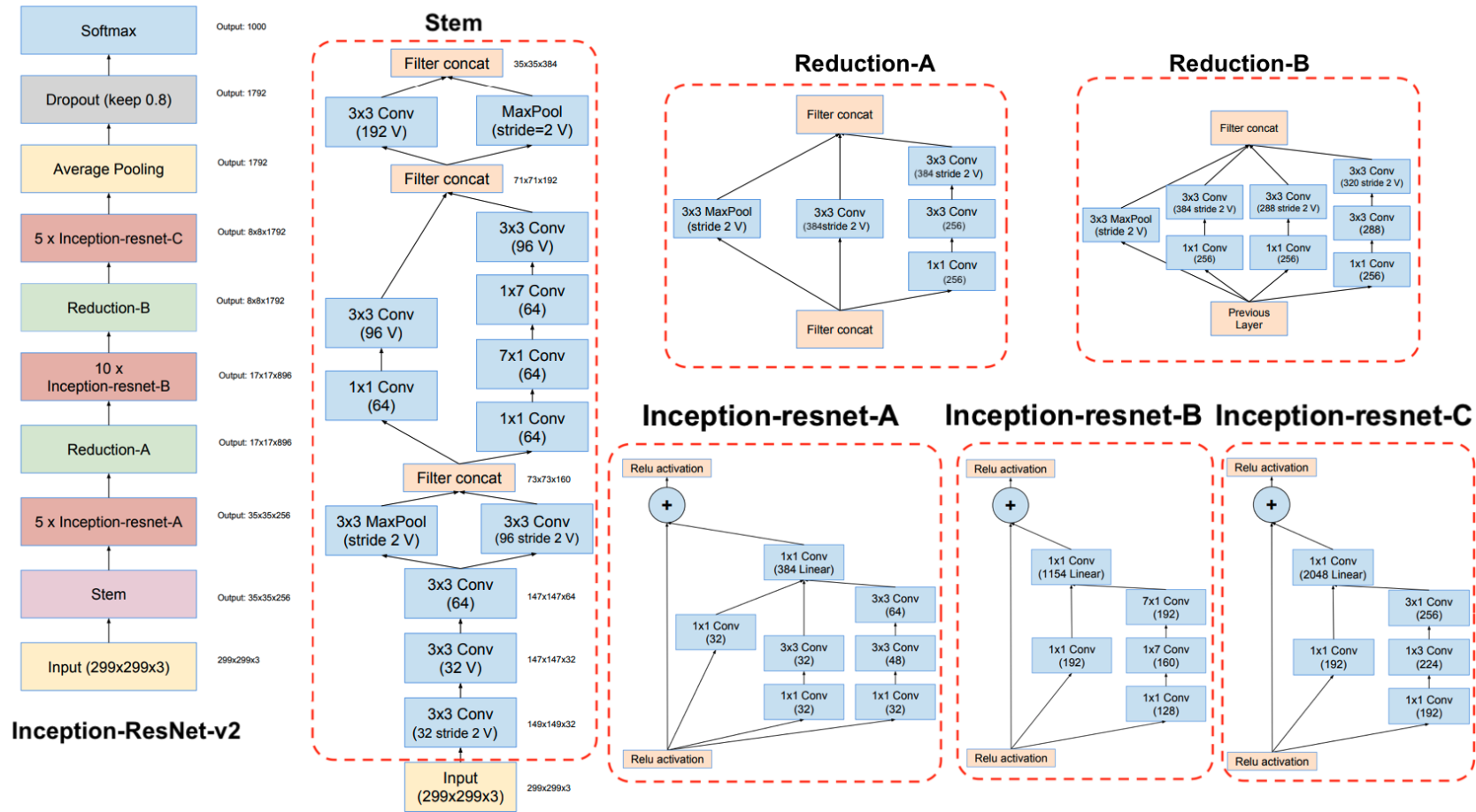


Figure 2.12 Illustration for the Inception-ResNet-V2 architecture (Szegedy et al., 2017)

## 2.6 Learning approaches of deep-learning architectures

This section aims to provide background information about the different learning schemes applied in histopathology for the deep-learning. That includes supervised, unsupervised, weakly supervised, and transfer learning. Also, it discusses some related works in the related field and defines a potential area where this research could investigate.

### 2.6.1 Supervised learning

In tasks where a training set  $T$  has  $n$  number of samples  $x_i$  with their labels  $y_i$  ( $T = \{(x_1, y_1), \dots, (x_i, y_i)\}_{i=1}^n$ ). The goal is to learn a function  $f: x \rightarrow y$  that during the inference time can predict the label  $y$  given the unseen test sample  $x$ . Figure 2.14 shows the training dataset nature of the supervised tasks. One of the first uses of CNN in histological images following the supervised learning was proposed by Malon et al. (2008) for the mitotic count in breast cancer, epithelial layers detection in the stomach, and signet ring cells detection. In histological image analysing, according to Srinidhi et al. (2019), supervised learning involves three tasks: classification, regression and segmentation models. The classification models can be trained to locally classify patches within the whole slide to detect objects, such as nuclei, mitosis or glands, or identify diseased regions. One limitation of the supervised classification at the local level is the unavailability of the ground-truth labels for the patches within the whole slide due to the high cost of the annotation process. Therefore, applications of such models are limited to problems where exhaustively labelled data is available, such as the provided dataset by International Conference on Pattern Recognition 2012 (ICPR 2012) (Roux et al., 2013), CAMELYON (Litjens et al., 2018) and Breast Cancer Histology Images (BACH) (Aresta et al., 2019) challenges; otherwise, the weakly supervised learning is applicable (will be discussed in section 2.6.3). Usually, in real-world histologically tasks, the local-level classification utilises MIL.

(Cireşan et al., 2013) and (Wang et al., 2014) are related work in the local-level classification, which trained different architectures of CNNs in a supervised manner to detect mitosis in breast cancer using the ICPR 2012 dataset. The first work is the winning approach in the ICPR 2012 competition that used deep CNN with max-pooling layers to extract high-level features

from H&E stained breast tissues to detect mitosis. After learning the features, they were fed into the Softmax classifier to predict the presence of a mitotic nucleus in each region. Wang et al. (2014) aimed to increase the robustness of the previous system by mixing hand-designed features with CNN learned features.

Moreover, the classification models can be trained to classify the whole slide based on globally patches classes. Wang et al. (2016) proposed a system to identify cancer metastases from WSIs of breast sentinel lymph nodes. They participated in the “CAMELYON” Grand Challenge 2016 competition and won the competition. Their system is composed of a deep convolutional neural network trained using millions of batches following the supervised learning approach. The trained deep CNN is locally classified patches into cancer or normal. Then the detected tumour patches were aggregated to produce the tumour probability heatmaps. Then, geometrical and morphological features were extracted from the heatmaps to be used later by a predictor (random forest classifier) for the slide-based classification and tumour localisation. The performance accuracy of the proposed deep-learning predictor has become near the human level.

On the other hand, regression models are used in detecting objects or localising them. It regresses the probability of each pixel in the image being the centre of an object (Srinidhi et al., 2019). Sirinukunwattana et al. (2016) proposed a regression model to detect nuclei using a spatially constrained CNN that calculates the probability of a pixel being the centre of a nucleus. Then, a spatial constraint is applied to the pixels with high probabilities to locate them in the nearest centre of nuclei. A neighbouring ensemble predictor works alongside CNN to increase the accuracy of the classifier. Also, Chen et al. (2016) proposed a deep regression network composed of a downsampling part that extracts the high-level information from the virtual stained slide and an upsampling part that generates a heatmap for the slide with the scores of each pixel. That model was used with the ICPR 2012 dataset. They utilised transfer learning to overcome the high cost of needing a huge labelled dataset as most of the works in the supervised learning combine that approach with the transfer learning technique (will be discussed in section 2.6.4).

Supervised learning involves segmentation models used to separate different layers of tissue as a prerequisite for extracting useful features. For instance, Galal and Sanchez-Freire (2018) proposed a Candy Cane system that consists of a fully convolutional “Densenet” architecture to segment the whole slides in the “BACH” dataset. That model is based on the auto-encoder architecture with downsampling and upsampling parts and skipped connections between them to save the low-level features. The proposed model was applied to the provided dataset by the “BACH” challenge to generate pixel-wise labels for each virtual slide. Each pixel was predicted as normal tissue, benign lesion, in situ carcinoma or invasive carcinoma.

Furthermore, deep CNN was used to recognise epithelial and stromal compartments to detect breast cancer (Xu et al., 2016). The proposed system generates small patches from the histological images, using a superpixel method to atomic segment regions and the square window method to resize the patches into fixed-size patches. Then, it uses the extracted patches as input to the deep CNN to learn features. Table 2.6 summarises some of the related works in supervised learning.

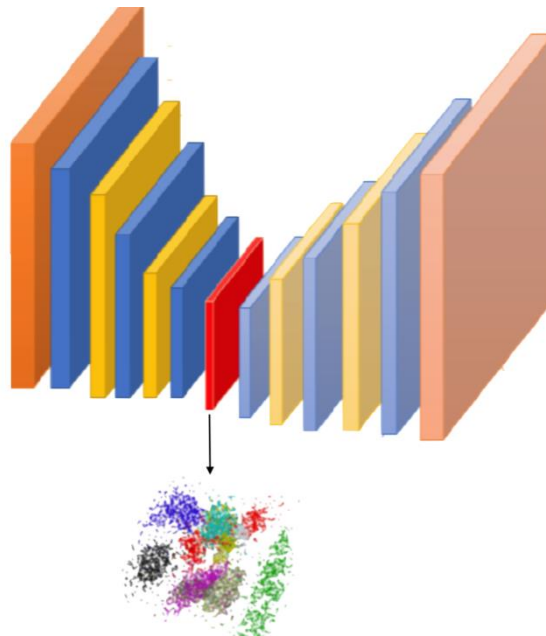
**Table 2.6** A summary of the related works that employ supervised learning in the field of histology

Paper	Application	Task	Approach
(Cireşan et al., 2013)	Mitosis detection in breast cancer	Local-level classification	Pixel-based CNN
(Wang et al., 2014)	Mitosis detection in breast cancer	Local-level classification	CNN and handcrafted features
(Wang et al., 2016)	Detection of breast cancer metastases	Local-level and global-level classification	CNN based patch classifier and random forest classifier
(Sirinukunwattana et al., 2016)	Nuclei detection in colon cancer	Regression	spatially constrained CNN and a neighbouring ensemble predictor
(Chen et al., 2016)	Mitosis detection	Regression	Auto-encoder scheme of a downsampled path and upsampled path
(Galal and Sanchez-Freire, 2018)	Pixel-wise classification for breast cancer.	Segmentation	DenseNet following the auto encode scheme
(Xu et al., 2016)	Breast and colorectal cancers segmentation	Segmentation	deep CNN



## 2.6.2 Unsupervised learning

Unsupervised learning aims to find the underlying representation of the data in the absence of ground-truth labels, which might be ambiguous due to the possible infinite representation (Srinidhi et al., 2019). Employing an algorithm, such as auto-encoders, to reduce the mapping dimensionality of the dataset will solve that issue, as illustrated in Figure 2.13. Auto-encoders are the most famous and successful unsupervised deep-learning algorithms. In computer vision, the auto-encoder architecture contains encoder and decoder networks. The encoder part aims to build an encoding function to extract the distinct features from pixel intensities. In contrast, the decoder part can remodel the original pixel intensities using the learnt features (Deng and Yu, 2014).



**Figure 2.13** An overview of the unsupervised convolutional auto-encoder model

An optimal learned stacked auto-encoder should be able to find underlying representations (the layer in red) for the data in a way that samples from the same class cluster in the same group.

Moreover, they are a powerful tool in image segmentation. Table 2.7 shows many works employed in tissue and nuclei segmentation that need dimension reduction for the dataset followed by aggregation. Zhang et al. (2015) employed a layer of sparse auto-encoder combined with the Softmax classifier to extract high-level representation and detect basal-cell carcinoma

cancer in the histology field. Also, layers of auto-encoders were employed to construct a nuclei detector that takes high-resolution histological images of breast tissues (Xu et al., 2015). A sparse auto-encoder was used by Hou et al. (2019) to detect and segment nuclei in breast cancer.

**Table 2.7** A summary of the unsupervised related works in the field of histology

Paper	Application	Approach
(Zhang et al., 2015)	Basal-cell carcinoma cancer detection	Sparse auto-encoder
(Xu et al., 2015)	Nuclei detection in breast cancer	Sparse auto-encoders
(Hou et al., 2019)	Breast cancer nuclei detection and segmentation	Sparse auto-encoders
(Hu et al., 2018)	Bone cancer cell-level classification and counting and nuclei segmentation	GAN
(Quiros et al., 2019)	H&E stained image generation and feature extraction	GAN

Generative Adversarial Networks (GAN) are needed in more complicated unsupervised tasks such as classifications. Goodfellow et al. (2014) introduced GAN to the deep-learning community. GAN is an unsupervised approach that aims to learn the underlying structure. Its structure relies on two components: generator and discriminator modules that run in an adversarial way. The generator part takes random numbers, such as noise from Gaussian distribution, and uses a stacked deconvolution network to build a non-real image. In contrast, the discriminator is a CNN trained using both the generated images (fake images) and real images to classify the test images into real or fake.

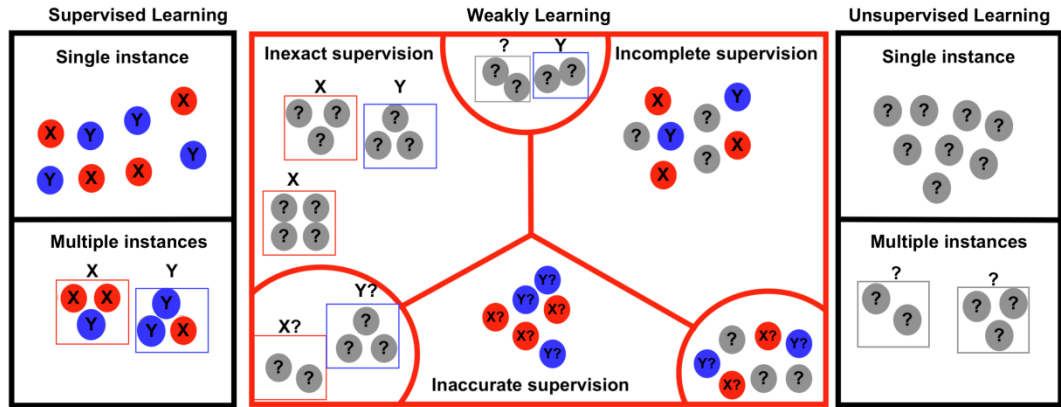
The training phase aims to learn a generator that produces real-like images and learn a discriminator that can distinguish between real and fake images. When the model can discriminate the real from fake, it indicates that the mutual information was maximised. As a result, the latent representation can be used for the classification task. Hu et al. (2018) proposed a GAN-based model for cell-level classification, counting, and nuclei segmentation. Quiros et al. (2019) adopted a GAN-based model to generate high-resolution H&E stained images and extract features that translate tissue feature transformations.

Although some attempts were made to apply the unsupervised deep-learning in histopathology, this approach is not common in that field and needs to be investigated more, as such approach will reduce the annotation burden on the pathologists and cut the high cost of such a process

### **2.6.3 Weakly supervised learning**

As discussed in section 2.6.1, the supervised learning approach is the best to train predictive models when a huge, accurately labelled dataset is available. Most of the real-world tasks suffer from insufficient labels, such as tasks in the medical field or tasks that a limited sensitivity data collectors collected the data. Labelling such datasets can be tedious and extremely expensive because expert annotators should be involved in the labelling process. Those tasks pave the way for weakly supervised learning, which is recommended in those cases. According to Zhou (2018), weakly supervised learning can be categorised into three main categories based on the levels of the fed information: incomplete supervision, inaccurate supervision and inexact supervision, as illustrated in Figure 2.14. Incomplete supervision is the case where part of the data is labelled, and the labels of the remaining data are not given. The most common form of that category is the semi-supervised classification. Inaccurate supervision is the case of tasks with labels that may suffer from error. In this category, the data labels might not be the ground truth. An example of that category is the crowd annotations, where a group of non-experts people provides labels in an attempt to reduce the cost of the annotation resulting in unsure ground truth for the data.

Another example is the candidate labelling system, where more than one label is assigned to an instance, and the task of a classifier is to find the correct label from the provided set. The last category is inexact supervision, where labels of the task are provided in a way that is not as precise as aimed. For instance, the coarse-grained labels are provided in a task but not the aimed fine-grained labels. Multiple instance learning (MIL) is a famous scenario of inexact supervision as the label of a bag is known while the labels of the instances in the contained bag are not. This problem was first addressed by Keeler et al. (1991), while the MIL term was first coined by Dietterich et al. (1997). Also, it is the most common problem in the field of histological images annotations.



**Figure 2.14** Illustrations for the scenarios of supervised, unsupervised and the different weakly supervised tasks

In this figure, circles denote instances, while squares denote bags of instances. “X” and “Y” are well-known labels, “?” is an unknown label and “X?” and “Y?” are noisy labels.

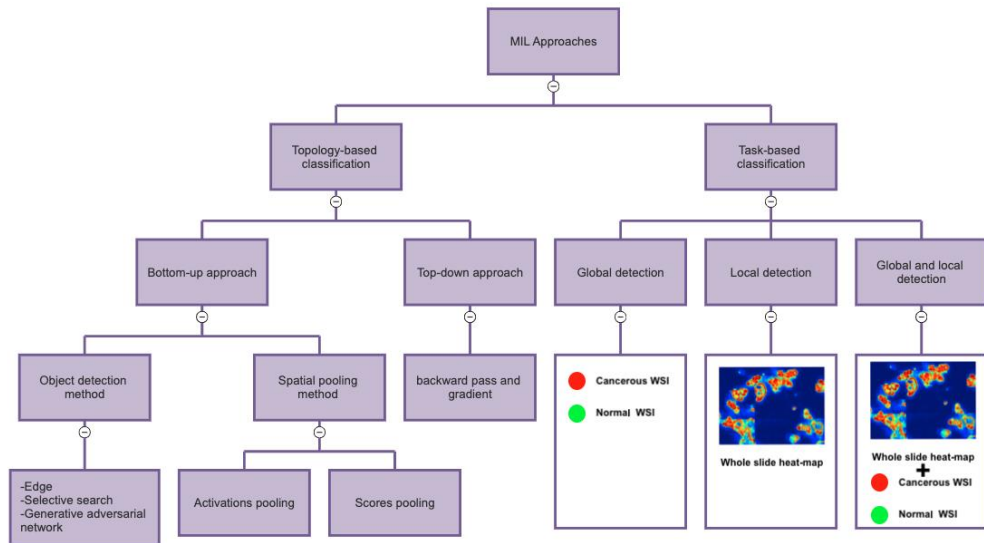
MIL is described as a task ability to learn a function ( $f$ ) from the training data  $D = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$  where any  $i^{th}, i \in \{1, \dots, n\}$  training data ( $X_i$ ) consists of a different number ( $m$ ) of instances  $X_i = \{x_{i,1}, \dots, x_{i,m}\}$ , to predict the label ( $Y$ ) for input test data ( $X$ ) where  $Y \in \{0, 1\}$ . The label  $Y_i$  for the training bag is known; however, the set of labels  $\{y_{i,1}, \dots, y_{i,m}\}$  are unknown.  $Y$  is negative when all the labels of its instances are negative and positive otherwise. Mathematically, the function is described as  $f: X \mapsto Y$  where:

$$Y = \begin{cases} 0, & \text{if } \forall y_i \text{ where } i \in \{1, \dots, m\}, y_i = 0 \\ 1, & \text{otherwise} \end{cases}$$

Multiple medical applications of the MIL approach were reviewed by Quellec et al. (2017). Also, several works investigated the usage of MIL for the H&E stained virtual glasses. This section focuses on the papers that adopted deep-learning models trained using H&E stained histological images with the annotation-level labels to perform a specific task such as classification, localisation, or segmentation of the whole virtual slide. In this thesis, the most relevant weakly-supervised approach is the inexact supervision and, more precisely, MIL scenario due to the nature of the histological images. The exhausted pixel-level annotations for these images are costly and rarely available. Instead, coarsely-annotated image-level labels are affordable. Many kinds of research for weakly supervised classification, localisation and segmentation using deep-learning approaches were proposed to alleviate that issue.

A survey was conducted by Rony et al. (2019) to apply weakly supervised learning in histological images classification and localisation. They found the research grouped into two primary approaches depending on the following topology: (1) bottom-up approach and (2) top-down approach. The difference between the approaches is the direction of passing the information, as the first approach uses forward passing while backward passing is used for the second one. All the related works in this literature follow the first approach. Furthermore, they recategorised the first approach into two methods: (1) the spatial pooling method that relies on the spatial pooling of the activations or the scores, and (2) the object detection method. Most of the deep, weakly supervised learning research in the histopathology field focuses on the spatial pooling method.

Moreover, Cheplygina et al. (2019) categorised MIL approaches into three categories based on the targeted task. The first category is the global detection task that aims to identify a pattern at the image level, such as learning a model to predict the image-level label. The second one is the local detection task, which aims to identify a pattern at the patch-level or the pixel-level, for instance, learning a model to predict the label of each patch or pixel in the bag and generate a visual heatmap for the histological image. The final category is the global and local detection task which aims to accomplish the tasks of the previous two categories, such as learning a model to detect if the histological image has cancer and locate it. Figure 2.15 illustrates the different categorisation schemes of MIL approaches.



**Figure 2.15** A tree diagram for the different categorisation schemes of MIL approaches

Following the categorising system by Rony et al. (2019), most works have a bottom-up topology. They were implemented based on the spatial pooling of the activation, as the authors assumed the spatial pooling could focus on the relevant patches. For instance, Xu et al. (2014a) proposed a MIL-Boost algorithm based on deep-learning. They used a CNN to extract feature representations from patches. Then, a generalised mean pooling function was applied to those representations to predict the image-level label. For classifying cancer in colon histopathology, the model scored 96% accuracy, whereas using the supervised approach on the fine-grained labels achieved 95% accuracy.

Altschuler and Wu (2010) argued that in microscopy images, classifying cellular phenotypes is challenging because of the cellular population heterogeneity and the presence of artefacts and neutral instances, which might be found in both negative and positive cases. Kraus et al. (2016) claimed that the available MIL pooling functions could not overcome this issue. To solve this situation, they introduced an alternative pooling function to the deep-learning MIL approach robust to outliers, the Noisy-AND pooling function. The Noisy-AND pooling function accommodated the heterogeneity of cellular phenotypes by learning a certain threshold for each class. Their proposed solution outperformed many methods in classifying and segmenting microscopy images with populations of cells.

Similar to the work by Kraus et al. (2016) is the work of Das et al. (2018). They introduced a new pooling layer to a deep-learning MIL framework, known as multiple instances pooling, to maximise the useful aggregated features locally about the patches and learn the global image classification. The multiple instances pooling consists of two max-pooling layers, one for the patch level and the other for the image level. That eliminates the need to explicitly annotate patches and be content with knowing the bag's label. The model achieved 0.89, 0.89, 0.88 and 0.87 accuracies at 40X, 100X, 200X and 400X magnifications respectively. Even though the model outperformed other methods in its accuracies at most magnifications, its performance at the lowest available magnification (40X) was not the best compared to other works. In practice, researchers tend to implement CAD systems with the lowest computational costs to guarantee the feasibility of their application.

Paschali et al. (2019) applied a spatial global average pooling on the generated activations (the feature vector) generated by a deep neural network framework at the image level. Then, the fine-grained logit heatmaps for the activations of the models were utilised to draw the decision by feeding them into a fully connected linear classifier that infers the logic values. A Softmax function converted those values to probabilities. Then the model was quantitatively and qualitatively evaluated. It showed better results than the famous attention-based approach proposed by Ilse et al. (2018). Their model is an attention-based approach based on the pooling of scores to overcome the interpretability of the classifier issue. When a classifier uses the ground-truth labels for unlabelled patches to be the label of the container image, it imposes the issue on the classifier. In this case, the classifier discriminates against the most abnormal part and neglects other less yet important abnormalities. They integrated an attention mechanism by Bahdanau et al. (2014) with the MIL approach. Based on the attention mechanism, they formulated the representations of the bags using weighted average scores of the instances activations. Their approach achieved similar performance to the best available MIL approaches when publishing their paper. The limitation of this work is its application to binary classification. Therefore, applying it to a multi-class classification task will require changing the pooling and scoring module to match the number of ground-truth classes in the new task. Also, Cui et al. (2020) utilised MIL pooling of the produced scores of the instances by an attention layer to aggregate the score of the bag. The proposed model achieved 0.84 AUC in classifying the IDH1

mutation in the glioma images. A final fully connected layer was used to predict the image-level label.

Besides Ilse et al. (2018), Campanella et al. (2019) conducted encouraging work based on the spatial pooling of the scores as well. They proposed a two-stage approach, where the first stage focuses on training a classifier using patch-level data with MIL; then, a spatial pooling of the predicted scores is applied. As a result, rich semantic patch-level representations are generated. The second stage adopted a recurrent neural network (RNN) to integrate those representations to obtain a slide-level diagnosis. The proposed approach yielded 0.98 AUC or greater in detecting four kinds of cancers from the whole slides obtained from more than one medical centre without fine-grained labelling.

An example of the global detection task for the MIL approach is the work proposed by Hou et al. (2016). It consists of two stages and relies on the discriminative patches to classify histological images of various cancers using the public TCGA dataset (Tomczak et al., 2015). In the first stage, a CNN is trained on the discriminative patches, which were aggregated by a novel formulated Expectation-maximisation method that utilises the spatial information of the patches for smoothing to generate patch-level predictions. Then the second stage decides the image-level label by feeding histograms of the patches predictions into a logistic regression classifier. Although this approach has shown an effective performance in classifying glioma, it is not recommended because it has a high computational cost. In return, it slightly outperformed other simpler approaches. Table 2.8 summarises the discussed MIL approaches.



**Table 2.8** A summary of the MIL weakly supervised related works in the field of histology

Paper	Application	Approche
(Xu et al., 2014a)	colon cancer detection	A MIL-Boosting method based on MIL pooling of activations
(Paschali et al., 2019)	classification of breast and colon cancers	MIL pooling of activations model trained on local features and a fine-grained logit heatmap visualisation
(Cui et al., 2020)	classification of the IDH1 mutation in the glioma	MIL pooling of scores
(Ilse et al., 2018)	classification of breast and colon cancers	MIL pooling of weighted average based on attention
(Das et al., 2018)	classification of breast cancer	MIL pooling of activations
(Kraus et al., 2016)	simultaneous classification and segmentation of breast cancer	MIL pooling of activations
(Hou et al., 2016)	Classification of brain cancers (glioma subtypes)	Expectation-maximisation based MIL/ global detection
(Campanella et al., 2019)	Diagnose WSI for multiple cancers(prostate, breast and skin)	MIL pooling of scores

An interesting work by Sudharshan et al. (2019) compared the performance of a MIL deep-learning approach with a conventional approach that relies on single-instance classification using the public BreakHis dataset (Spanhol et al., 2015), which consists of 8K benign and malignant breast tumours histologic biopsies. The MIL convolutional neural network was proposed by Sun et al. (2016) for object recognition. This work assumed that a label of an object might not be preservable when the object is augmented, thus creating a bag that inherits the object's label. It contains different instances of the generated version of the augmented object. They combined a CNN with a particular MIL loss function that concerned the generated bags. They concluded from the comparison that the MIL approach achieved a comparable or better performance than the conventional approach.

Xu et al. (2014b) proposed a model based on conventional machine learning approaches. It embedded the concept of clustering into the setting of MIL to offer a solution capable of accomplishing three tasks simultaneously to segment and classify colon cancer. Those tasks were image-level positive or negative to cancer classification, tissue segmentation into cancerous and

non-cancerous tissue and clustering patches into different classes. The model took labelled tissue and sample patches with labels similar to the label of the tissue. Then, it employed conventional machine-learning feature extractors with different classifiers and integrated them into a MIL framework for patch-level clustering into different types of cancer. The image-level classification was inferred from the patch-level clustering.

All the proposed works in the medical field focused on the classification, localisation or segmentation of the diseased lesion using weakly labelled data. The weakly supervised learning has not been previously investigated in the literature for cleaning the weakly labelled data as far as we know. Then, use the cleansed version to train a deep network following the supervised approach.

#### **2.6.4 Transfer learning**

In machine learning, the concept of utilising knowledge acquired by previous tasks has become the interest of many researchers in developing solutions for related problems. This approach is known as transfer learning, and it imitates how past experiences psychologically affect humans and improve how they deal with upcoming tasks. Thus, for a new target task, one can train an existing model borrowed from a related domain, aka source domain, as the target instead of creating a model from scratch. This approach can reduce the need for large training data and improve training time.

Transfer learning is considered a powerful tool, especially when developing models in domains that struggle with insufficient data, whether caused by difficulty in acquisition or annotation (Tan et al., 2018). The limitation in data in these scenarios motivates keeping as much data as possible outside the training phase achievable by transfer learning. Nevertheless, how much of the transfer knowledge is to be applied will depend on the problem on hand; to what extent do the domains relate and how large is the available dataset.

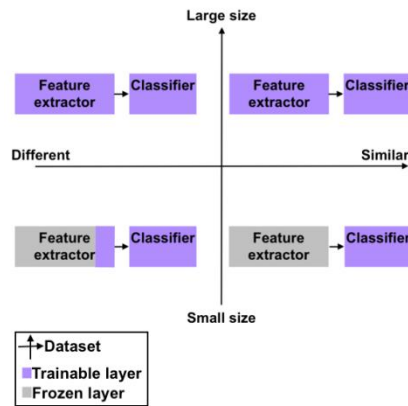
An example of transfer learning is image classification in deep-learning via CNN. CNN consists of many layers; the closer the layers to the input, the more general the features extracted by their neurones are to the data, whereas the closer to the output, the more specific they become. In this

case, a similar target task with a limited dataset can train only the last layers of the network on a subset of the training data. However, the less similar the source and task domains are, the more layers might need to be involved during fine-tuning.

Though great benefits could be attained with transfer learning, finding a suitable source domain might not be that straightforward. A negative transfer can occur if an unsuitable source domain is selected, resulting in degraded performance. Some measures that can assist the choice would be comparing the initial performance differences for the transferred model as is and a model built from scratch, the amount of time they take to learn the task and their final achieved performances after the training phase (Torrey and Shavlik, 2010). While similarities in source and target domains contribute positively to the learning process, in some cases selecting a different source domain can still produce encouraging results (Srinidhi et al., 2019). If the source model were to be extensively trained, its performance exceeds using averagely trained models from related domains when fine-tuned to a target task. That statement was supported by an experimental study (Kohl et al., 2018) that investigated the performance of classifying the BACH dataset using a pre-trained model on the non-domain related dataset ImageNet and the domain-related CAMELYON dataset. They found that the pre-trained model on “ImageNet” outperformed the performance of adapting the other model, as the pre-trained model on ImageNet yielded 94% accuracy whereas the other model scored 76.75% accuracy.

With deep-learning models, the most common transfer strategy is to reuse the pre-trained feature extractors from the source domain model, add the suitable classifier and fine-tune the new model. Different scenarios of fine-tuning the transferred model are closely attached to the target dataset size and its similarity to the source dataset. For instance, fine-tuning the whole model is the best scenario if a large dataset is available. In contrast, in the case of the availability of a small dataset, it is preferred to fine-tune some last layers depending on the similarities between the source and the target datasets. The number of trainable layers decreases with the increase of the similarities between the source and target datasets to be limited to the fully connected layer of the classifier in the high similarity (Yamashita et al., 2018). Figure 2.16 illustrates the different fine-tuning scenarios relying on the

nature of the target dataset. Table 2.9 introduces some of the research in histopathology that led that path and used published models.



**Figure 2.16** Fine-tuning scenarios for deep-transfer learning based on the available dataset

**Table 2.9** An overview of the best performing works that adapt the deep transfer learning approach in CAMELYON 16, CAMELYON 17 and BACH competitions

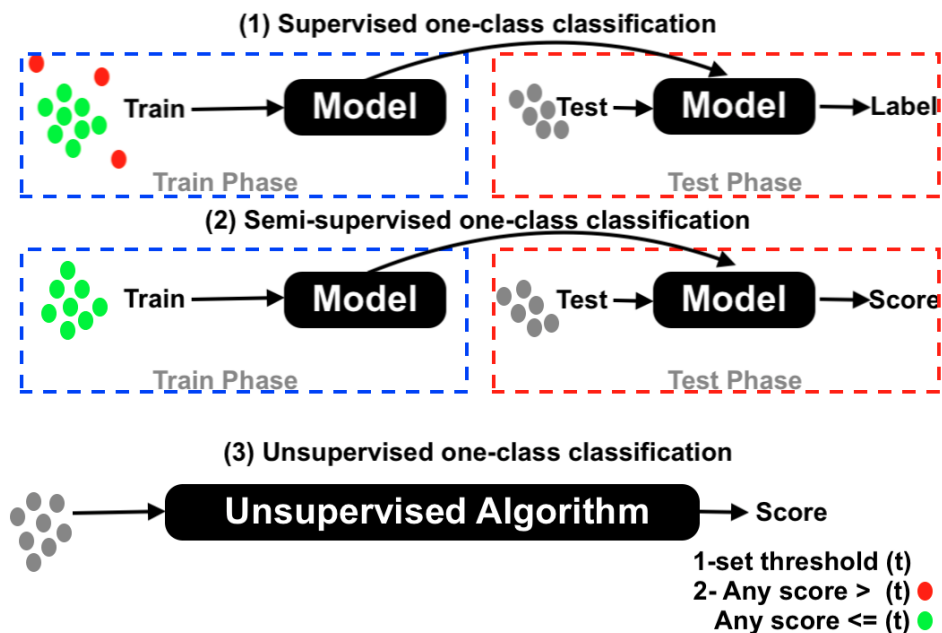
Paper	Application	Model	Dataset		Performance
			Source	Target	
(Wang et al., 2016)	Detect breast cancer	GoogleNet	ImageNet	CAMELYON 16	0.71 AUC localisation 0.93 AUC WSI classification
(Lee and Paeng, 2018)	Detect breast cancer and pN-stage classification	ResNet101	ImageNet	CAMELYON 17	0.985 AUC
(Chennamsetty et al., 2018)	Breast cancer classification	Resnet-101; Densenet-161	ImageNet	BACH	87% accuracy
(Kwok, 2018)	Breast cancer classification	Inception-Resnet-v2	ImageNet	BACH	87% accuracy
(Brancati et al., 2018)	Breast cancer classification	Resnet-34, Resnet-50, Resnet-101	ImageNet	BACH	86% accuracy
(Kohl et al., 2018)	Breast cancer classification	DenseNet-161	CAMELYON 17	BACH	83% accuracy

## 2.7 Deep one-class classification

The multi-class classification task is accomplished by training a model to discriminate between two, known as binary classification, or more objects. The number of samples from each participating category in the classification task should be balanced during such models' training. When this condition is not applicable, or the dataset contains samples from one category, one-class classification appears to suit the task more. One-class classification aims to make the model learn to identify objects from one class only, and any unidentified objects are considered anomaly or outlier. This type of classification is the best choice if the negative class is absent in the training dataset, not well defined, or poorly sampled. (Grubbs, 1969) provided definition of an outlier as “An outlying observation, or “outlier”, appears to deviate markedly from other members of the sample in which it occurs”. Detecting outliers was motivated by the need to clean data from outliers that hamper the performance of pattern recognition tasks, as they are known for their sensitivity to outliers in data. However, with the development of machine learning, researchers have changed their perspective toward outliers from noise in the data to an interesting incidence. That has increased the interest in outlier detection. As a result, the previous definition was modified to include that anomalies differ from normal because they have different features and rarely occur in the dataset (Goldstein and Uchida, 2016). It is important to mention that anomalies and outliers are synonyms; however, their names reflect how they are considered in the task as outliers indicate unwanted instances like noise in data, while anomalies are rare and unique instances such as detecting unusual behaviours in airports. Recently, the application of anomaly detection extended to fulfil tasks in intrusion detection, fraud detection, data leakage prevision, suspicious movements in which surveillance cameras capture them and most importantly, in medical applications such as patient monitoring (Goldstein and Uchida, 2016).

Goldstein and Uchida (2016) classified the one-class classification depending on the amount of available ground truth labels in the normal and abnormal cases during the training phase. One-class classification is grouped into three approaches: supervise one-class classification, unsupervised one-class classification and semi-supervised one-class classification. When the positive class is perfectly sampled while the negative class is poorly sampled, the supervised approach is applicable to train a model using the skewed dataset to generate labels for unseen new

data. In contrast, the unsupervised approach is the only choice when the ground-truth labels for both classes are absent and, consequently, there is no distinction between train and test datasets. The unsupervised approach does not train a model for the classification purpose. Instead, it applies the unsupervised algorithm, usually distances or densities based algorithms, to score the data using their intrinsic properties. Then, a threshold of the results scores is set to discriminate between the normal and abnormal. Finally, the semi-supervised approach, also known as novelty detection, is applicable when the training set only contains data with positive labels. Novelty detection is the process of identifying new unobserved cases by assigning novelty scores for them. A novelty detector model is trained using a set containing regular data only (inliers) to discriminate their intrinsic properties. And during the testing, it uses a decision threshold score with the generated novelty scores for the test data point to predict the class (inlier or outlier). Confidence scores are generated instead of labels using the unsupervised and semi-supervised approaches indicating the degree of abnormalities. Figure 2.17 illustrates the different scenarios of one-class classification learning.



**Figure 2.17** Different one-class classification approaches based on the availability of the ground-truth labels of the training dataset

Circles in green represent normal data, in red are abnormal data, and in grey are data with unknown labels.

Furthermore, according to Chalapathy and Chawla (2019), deep-learning with one-class classification can be categorised based on the objective of the training into two approaches: deep hybrid models and one-class neural networks. The first category uses a deep-learning approach, usually any type of auto-encoders, to learn features from the dataset unsupervised. Then, the learnt hidden representations are fed into a traditional anomaly detector, such as a one-class support vector machine (OC-SVM), to detect anomalies. Gutoski et al. (2017) proposed a convolutional auto-encoder to learn hidden representations from video frames captured by a surveillance camera. After that, the extracted hidden representations were fed into OC-SVM to detect anomalies. Another example of the hybrid approach is the introduced network by Oza and Patel (2018), as they froze all the convolutional layers in a pre-trained CNN and removed the Softmax regression. In parallel to the feature extractor, they added a zero-centred Gaussian noise generator and appended a noise to every extracted feature as the pseudo-negative class data. Then, they added a fully connected layer. The Softmax regression takes the representation along with the noise as input and sets the Softmax's output to two (normal or abnormal). During the training, they computed the loss and updated the fully connected layer only. They compared their work against out-of-date approaches (between 2000 to 2004) and one recent one-class neural network approach (Chalapathy et al., 2018); they found that their approach outperformed the old research and was slightly higher than the former recent work. From our point of view, we believe that their model outperformed the proposed model by Chalapathy et al. (2018) because they used a powerful feature extractor (AlexNet or VGG16), while the other model used the encoder part of one convolutional auto-encoder. Even though Oza and Patel (2018) claimed they proposed an end-to-end trainable approach, we believe it should be considered a hybrid approach because the one-class classifier objective did not influence the feature extractor (in this case, AlexNet or VGG16). Its loss was used to update the classifier fully connected layer only.

One major issue in using the hybrid approach concerned Chalapathy et al. (2018) and Ruff et al. (2018). They argued that the one-class classifier does not contribute to defining rich differential features that participate in detecting outliers because the trainable objective is not customised for the one-class classification task. In other words, the one-class classifier does not influence the hidden representation.

To overcome this limitation, they proposed two models, which will be discussed later, which belong to the second approach, the one-class neural networks. During training a model of the one-class neural networks approach, it updates the learnt features based on the loss of the one-class classifier. Hence, this approach combines the ability of deep-learning networks to learn intrinsic features concerning the one-class objective. For instance, Chalapathy et al. (2018) proposed a two-stages one-class neural network approach. A convolutional auto-encoder was trained with the training dataset until convergence in the first stage. After that, they transferred the learnt encoder part to the one class neural networks by adding a layer to flatten the compressed representation to feed it into a feed-forward network with one hidden layer with freezing the parameters in the transferred encoder. Finally, the feed-forward network is trained until convergence. The proposed approach managed to find a hyperplane that separates the inliers from the outliers using the CIFAR\_10 dataset (Li et al., 2017). It outperformed other existing deep hybrid anomaly, detection models.

Like that, Ruff et al. (2018) introduced another one-class neural network that aims to find a minimum hypersphere that separates normal data points from anomalies. Their models were trained in two stages. The first stage aims to reduce the dimensionality of the input data using deep auto-encoders. In contrast, in the second stage, the part of the stacked encoder is transferred and fine-tuned using the objective of support vector data description (SVDD) one-class classifier. At the beginning of the fine-tuning, the SVDD classifier selects and fixes a random data point, the compressed representation for input to the stacked encoders, to be the centre of the hypersphere. Then, it uses each data point in each training batch to compute the distance between that data point and the centre of the hypersphere. The distances for the data points in a training batch are accumulated and considered the loss to update the stacked encoders weights through the back-propagation. After fine-tuning the model until convergence, they introduced two methods for determining the outliers: soft-boundary deep SVDD and hard-boundary deep SVDD. In the soft boundary approach, they used some data points from the outlier class in a supervised manner to determine the boundary of the hypersphere. Whereas, in the hard-boundary approach, the boundary of the hypersphere is set by selecting a threshold ( $t$ ) that if a score for a data point ( $s_i$ ), which is the distance for the data point from the centre of the



hypersphere, is greater than it, then the data point is considered an outlier. This work is discussed in detail in Chapter 5. The proposed work outperformed other semi-supervised one-class classification using MNIST (LeCun, 1998) and CIFAR-10 datasets. To compare the previous two works, Chalapathy et al. (2018) compared their one-class neural network against the soft-boundary deep SVDD (Ruff et al., 2018) using AUC as a performance metric CIFAR\_10 as a dataset. They found that the deep SVDD outperformed their model in detecting a class against others.

GANs were of interest to researchers in solving the one-class problem. For example, Gu et al. (2018) adopted the GAN approach to accomplish an anomaly detection task. They proposed a novel corrupted generative adversarial network. In the competition training of the approach, the generator aims to produce fake images considered as outliers, while the discriminator is trained on the inlier images and the generated outliers. It aims to distinguish between the inliers and the outliers, and they tested their approaches on an image dataset and network intrusion dataset. The approach achieved a state-of-art performance using both datasets.

Akçay et al. (2018) introduced a novel deep anomaly detector based on GAN architecture, known as “GANomaly”. They trained an encoder-decoder-encoder sub-networks to solve the issue of the non-convex optimisation. They aimed to minimise the distance between the generated images and the latent space to capture the distribution of normal samples. The proposed model showed an efficient performance over the previous approaches.

Despite the success of the previous approaches in detecting novelties and anomalies, the objective of their one-class classifiers lacks either compactness or descriptiveness. A successful classifier should have these two features. The compactness indicates that the classifier compacts the data from the same class, and the descriptiveness indicates that the classifier can separate groups from different classes. In the discussed GAN related approaches, the proposed approaches aimed to learn the underlying representation of the inlier and outliers. The used classifier did not utilise any losses based on the distances between class-like instances. As a result, the learnt classifiers are descriptive yet not compact classifiers. Both Chalapathy et al. (2018) and Ruff et al. (2018) proposed models with objective functions

that focused on minimising the boundaries of a hyperplane or a hypersphere. The optimal founded hyperplane or a hypersphere should contain as many normal samples, and any sample located outside them is considered an outlier. These approaches neglected the descriptiveness due to the absence of the outlier data. Perera and Patel (2019) proposed a novel approach for novelty detection that fulfils the compactness and descriptiveness of the one-class classifier.

Their approach has two phases, the first phase trains two identical CNNs, known as the reference network and the target network, and they share the same set of weights, using two datasets (reference and target datasets). The target dataset is the one with normal samples only, while the reference dataset is a public fine-grained labelled dataset such as ImageNet. A simultaneous training for the two networks is performed by feeding the reference and target datasets into the reference network and the target network. Moreover, the reference network classifies its input using a Softmax regression classifier. The target network uses any nearest neighbour algorithm to minimise the distance between the representations of the training sets. The total losses for the two classifiers are computed and back-propagated to update the shared weights. In this way, the target network will guarantee the compactness of the learnt weights, while the reference network guarantees the descriptiveness. The second phase is the testing phase, where both classifiers are removed and substituted by the nearest neighbour classifier. A subset of the target dataset is driven to be used as the reference inlier points. When testing new unseen samples, the new classifier will compute their distance from the target dataset and generate anomalies scores for the test set. This work is discussed in detail in Chapter 5. Only related works in the semi-supervised one-class classification in light of deep-learning were discussed in this literature. They are summarised in Table 2.10 as there is vast literature in the one-class classification field.

**Table 2.10** A summary of the semi-supervised deep one-class classification related works

Paper	Approach
(Chalapathy et al., 2018)	Encoder and a fully connected layer
(Ruff et al., 2018)	Stacked encoders and SVDD
(Gu et al., 2018)	GAN
(Akçay et al., 2018)	GAN
(Perera and Patel, 2019)	Parallel transferred two CNN models and a joint loss (binary classifier for the reference dataset and one-class nearest neighbour for the target dataset)

## 2.8 Discussion and conclusion

Section 2.1 provided histopathology information that includes histological slides, the anatomy of a normal oesophagus, changes to the oesophagus in case of Barrett's oesophagus and dysplasia, biopsy extraction and the intermediate steps for preparing glass and virtual slides. In addition, it discussed detailed histopathology guidelines for diagnosing different grades of Barrett's related dysplasia at the cytological and architectural levels. Additionally, it highlighted the clinical challenge in grading Barrett's related dysplasia. Finally, it discussed stain normalisation to adjust the varieties in staining the samples.

Section 2.2 presented related works in diagnosing Barrett's related dysplasia using either a machine learning or deep-learning algorithm. Section 2.3 discussed the most common performance metrics in evaluating CAD systems. It is concluded that the proposed models in this thesis are better evaluated and compared using recall and specificity because recall indicates the percentage of the diseased cases that were predicted as abnormal. The specificity indicated the percentage of the normal cases predicted as normal. Moreover, KV is used to measure the agreement between the proposed CAD system with pathologists' diagnosis. Also, weighted KV is used to compare the performance of the proposed models with the related work

proposed by Adam (2015) in agreeing with the pathologists and measure the degree of agreement.

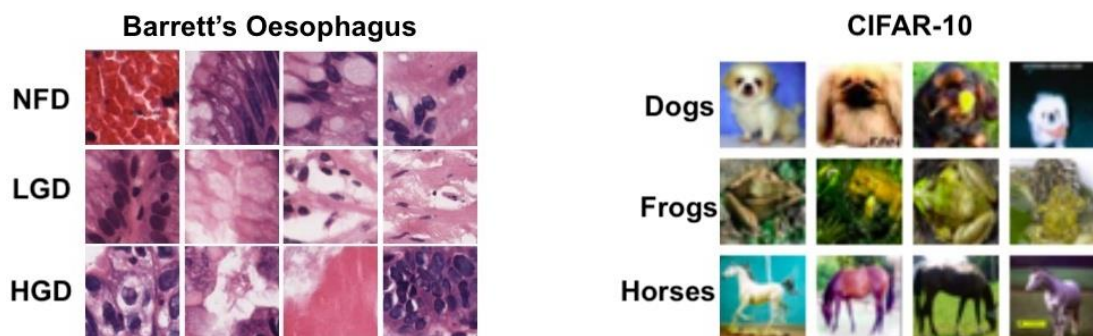
Section 2.4 overviewed the biological inspiration behind deep-learning. Section 2.5 explained deep-learning, particularly CNNs, as they are the most popular deep-learning architecture in the computer vision field. Also, it discussed the different architectures of CNNs and the different approaches for learning such networks. The different architectures of CNNs vary in length; however, at some point, the models' performances start to degrade due to gradient vanishing, and the computational costs dramatically increase. By introducing two concepts, residual networks and inception networks, the depth and width of networks were increased, and the performance was enhanced. This thesis employs the CNN architecture that adapts the two concepts (Inception-ResNet-V2) due to this model's success.

Then, section 2.6 explained the different techniques followed to learn those networks. The simplest way to learn a CNN is to follow the supervised learning approach. It uses a huge dataset that needs to be labelled by an expert, which is an impossible task in histological images considering their nature. Other researchers followed the unsupervised approach in their models, usually segmentation models. In the case of the availability of insufficient labelled data, researchers tend to adopt the weakly supervised approach.

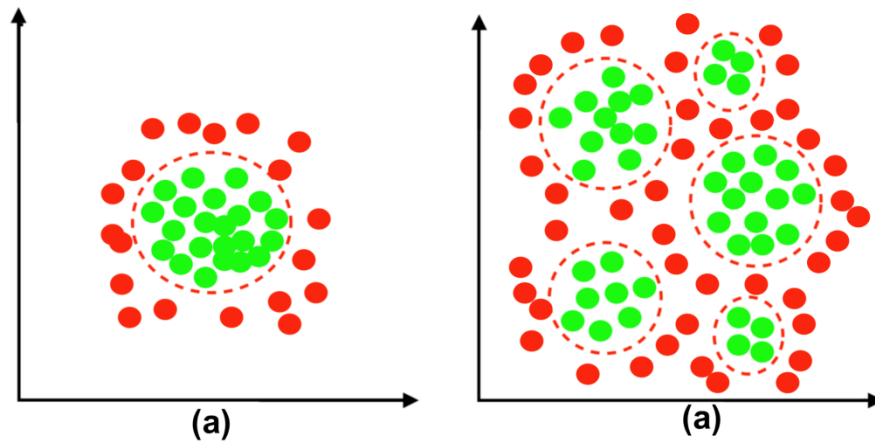
The literature has shown a growing interest in the bottom-up deep-learning MIL approaches using spatial pooling techniques in processing and analysing histological images. In contrast, there is less focus on object detection techniques. In addition, researchers have extensively explored the classification and the localisations of disease by examining the H&E stained biopsies, and some works showed an interest in lesion segmentation. However, to the best of our knowledge at the time of writing this chapter, studies lack deep-learning approaches that aim to either exclude negative for a diseased tissue from positive for a disease coarse-grained annotations or detect the discriminative patches should be analysed to diagnose whole virtual slides. That indicates the need to explore approaches based on object detection techniques to cleanse data or highlight key instances for further analysis.

Besides the supervised, unsupervised and weakly supervised learning, the transfer learning approach was adopted intensively by research alongside the previously mentioned techniques to overcome the limitation of the training models with a huge dataset size and reduce the training time. This thesis proposes a weakly supervised approach to fine-tune a transferred model to classify dysplasia in patches locally within Barrett's oesophagus tissue.

All the works in the literature on deep-learning and semi-supervised one-class classification were tested using small images with fine-grain labels and well-defined objects such as handwritten numbers, cats, dogs, and cars. To the best of our knowledge, deep semi-supervised one-class classifications have not been investigated in the histopathology field using the one-class neural networks approach yet. Images in histopathology are huge and usually follow coarse-grain labels, and each image contains a heterogeneous structure. For instance, in a whole virtual slide image for a biopsy from a patient with Barrett's oesophagus, there is a grid image that each cell in it might contain a nucleus, mitosis, goblet cell, basement membrane, or muscle tissue, or it might contain a mixture of the previous components, Figure 2.18 illustrates that point. Using the discussed approaches in section 2.7, with datasets that they evaluated their approaches on are expected to group the "cats" samples and separate the group from "cars", "planes", "frogs", and so on. However, the histological images are expected to group tissues from lamina propria in a cluster and group nucleus images in another. Additionally, it is expected to have normal nuclei in the cluster's centre while abnormal nuclei scatter around the nucleus cluster. Figure 2.19 illustrates that point of view.



**Figure 2.18** Samples from Barrett's oesophagus related dysplasia histological images dataset and the CIFAR-10 dataset



**Figure 2.19** The expected distribution of inliers and outliers for CIFAR-10 and Barrett's oesophagus related dysplasia histological images datasets

(a) Illustrates the expected population for the CIFAR-10 dataset, and (b) illustrates the expected population for Barrett's oesophagus related dysplasia histological images dataset. Green and red dots represent inliers and outliers, respectively.

Applying deep SVDD in this thesis raises the concern of cluttering the heterogeneous population of the histological images to minimise the compactness loss; even if it forces the model to update its weights set with values close to zero to reach zero loss, that concern will be investigated in Chapter 5. Also, considering applying the proposed approach by Perera and Patel (2019) will need to find a one-class classifier that suits the population of the heterogeneous structures of the histological images, such as Local Outlier Factor classifiers. More details about this classifier are provided in section 5.3.1.2.

The MIL problem assumes that the negative bags contain only negative ones while the positive ones might contain one or more positive ones without knowing which ones are positive. This case exactly matches Barrett's oesophagus related dysplasia coarse-grain annotations. It is known that NFD annotations are clear from dysplasia, while dysplastic annotations have a mix of non-dysplastic and dysplastic tissue due to the high cost of fine labelling them. Nevertheless, the novelty detection approach only allows training a model using positive data and tests it on unlabelled data to detect the novel class. We can consider the potential of employing novelty detection in solving the MIL problem. The questionable issue in the research dataset is "how to exclude the NFD instances from the dysplastic

annotations before making each instance within a bag inherit the label of the container bag?”. If that is achieved, then the cleansed data can be used as fine-grain labelled data to train a model in a supervised learning manner without worrying about fuzzing the model with the incorrectly imposed dysplastic labels on the actual NFD instances through the inheriting process. Moreover, to investigate its effectiveness as a dysplastic detector, the novelty detector module could be involved in the classification process, not limiting its role to the data-cleaning phase.

## Chapter 3. Histopathology Dataset of Barrett's Oesophagus

The Leeds Institute of the Molecular Medicine provides the used materials in this thesis in a spreadsheet file containing details recorded by two pathologists for their annotated regions that they relied on during diagnosing whole virtual slides. This chapter provides more information about data collection, selection, pre-processing, and dataset challenges.

In this chapter, the data collection work was accomplished by Treanor et al. (2009), and a huge part of the dataset filtration was carried out by Adam (2015). The dataset is the same dataset collected by Treanor et al. (2009), and it was used by the thesis of Adam (2015) in the first place. The overviewed description in this chapter is for completeness and not as a contribution. Moreover, for the WSIs pre-processing, the approach that was used by (Haggerty et al., 2014) was followed for preparing our dataset.

### 3.1 Research material

A consultant pathologist, Dr Darren Treanor, selected 148 H&E stained glass slides from the archives of the pathology department of Leeds General Infirmary<sup>1</sup>. These slides belong to 127 patients who were confirmed to have Barrett's oesophagus and were undergoing endoscopic surveillance. Multiple endoscopic biopsies (between 2 to 6) were taken from 21 patients out of the 127 at different stages of their dysplasia progression. There was evidence of columnar lined oesophagus on each glass slide for each patient, and each patient at least had one endoscopic biopsy with specialised intestinal metaplasia. The selected slides have biopsies that show the mucosal layer of the oesophagus, and most of them contain the epithelium lining and the lamina propria. As discussed in Chapter 2, dysplastic changes occur in those two regions, and dysplasia can be graded as low or high depending on its severity.

---

<sup>1</sup> This thesis falls under Dr Treanors Local Research Ethics Committee Approval (Leeds WestLREC 05/Q1205/220)



The pathologist reviewed the staining quality of the selected set and ensured it covered all the cases of different dysplasia grades from NFD to IMC following the Vienna classification (Schlemper et al., 2000). Amongst all the cases, the NFD grade represents the majority. After the glass slides selection, a unique random number was assigned to each slide to anonymise patients' identities. Then, the "Aperio T3" scanner was used to scan the selected glass slides, which were 144 glass slides, as the excluded four glass slides were either broken or had a thick cover slip, using a 40X objective lens to produce virtual slide images (some glass slides were scanned multiple times, for instance, the scanned virtual slides "10604.svs" and "10606.svs" belong to the same glass slide) of 0.23  $\mu\text{m}$  per pixel. The virtual slide images have a pyramid structure of around 40X, 10X, 2.5X, and 1.25X or 0.6X magnifications. They are saved as "SVS" files on a server at Leeds Institute of Molecular Medicine (they can be viewed online<sup>2</sup>). More details about the dataset can be found in (Treanor et al., 2009).

### **3.2 Ground truth of virtual slides**

The standard grading system for dysplasia in Barrett's oesophagus follows the Vienna classification, which contains six grades; however, the ground truth labels of the whole virtual slides were limited by Adam (2015) to the suggested three categories classification by Kerkhof et al. (2007) and Montgomery et al. (2001), due to the limited number of "probably positive for dysplasia" annotations. A domain pathologist approved the new classification that groups NFD and "probably negative for dysplasia" into NFD, "probably positive for dysplasia" and LGD into LGD, and HGD and IMC into HGD. Those classification systems can be reviewed in Table 2.1.

A hundred and forty-four of the collected glass slides and their scanned virtual slides were sent to two UK experts (gastrointestinal pathologists). They were asked to individually annotate regions of tissues where signs of dysplasia are prominent, based on the morphological appearances of the tissue, and they were asked to assign grades to them. In addition, they

---

<sup>2</sup> [http://129.11.191.7/Research\\_1/Darren/Barretts/](http://129.11.191.7/Research_1/Darren/Barretts/)

assigned grades to each whole virtual slide to be the grade of the highest grade of any contained annotation. At the end of the annotation phase, only 140 glass slides and their correspondence virtual slides with 18 extra copies of some of the virtual slides. The two experts (“Expert\_B” and “Expert\_E”) are United Kingdom national gastrointestinal pathologists with more than 24 years of experience in their field at the time of the dataset annotating process.

In general, 158 virtual slides are available for this research. Each whole virtual slide has two labels that the two experts assigned following the three category classification for 143 virtual slides. Moreover, the pathologists graded the remaining 15 whole slides used for testing the research approaches twice on different occasions under the microscope (whole glass slide) and the screen (whole virtual slide). As a result, four labels are available for each whole slide in the test set.

For all the virtual slides, the interobserver agreement between the two experts scored for three grades of dysplasia is 84.18%, with 0.689 KV and 0.787 weighted KV. The strength of the agreement for the virtual slides falls in the confidence interval of 0.61-0.80, and it is considered a substantial agreement. Table 3.1 shows the number of virtual slides in the train and test set with their statistics.

**Table 3.1** The ground truth labels for the whole virtual slides within different subsets of the data provided by "Expert\_B" and "Expert\_E"

Expert		E															
		Dataset slides before the first filtration phase				Dataset slides after the first filtration phase				Training set				Testing set			
		Negative	LGD	HGD	Total	Negative	LGD	HGD	Total	Negative	LGD	HGD	Total	Negative	LGD	HGD	Total
B	Negative	94	2	0	96	85	1	0	86	79	0	0	79	6	1	0	7
	LGD	18	10	3	31	16	8	3	27	14	7	3	24	2	1	0	3
	HGD	1	1	29	31	1	1	28	30	1	1	23	25	0	0	5	5
	Total	113	13	32	158	102	10	31	143	94	8	26	128	8	2	5	15
Interobserver agreements		84.18%				84.62%				85.16%				80%			
KV		0.689				0.7				0.7				0.674			
Weighted KV		0.787				0.797				0.796				0.789			

### 3.3 Ground truth of annotated regions

All the available annotated regions were extracted from 158 virtual slides, showing unhealthy tissue only. The pathologists annotated 433 regions in total, and out of these regions, 252, 73, and 108 regions belong to NFD, LGD, and HGD, respectively. The dimensions of those annotations vary, but generally, they have at least 250 pixels and at most 5000 pixels in each width and height. Table 3.2 shows the number of annotations for each grade by each pathologist. The pathologists surrounded these regions separately using rectangles at different magnifications, and IDs were assigned to each surrounding region. As a result, each pathologist might select the same or different regions, and there is a chance of partially overlapped or fully overlapped regions or regions within regions. The pathologists recorded information about their annotations, including the parent whole virtual slide ID, the annotated region ID, their ID, the upper-left corner coordinates, the width and height and the grade of dysplasia. This information was gathered and stored in a “CSV” file, and they are organised in a table with four columns. The first one contains the previewable “HTTP” link to the annotation in the following form:

[http://129.11.191.7/Research\\_1/Darren/Barretts/\[slide ID\].sys?\[x coordinate for the upper-left corner of annotation\]+\[y coordinate for the upper-left corner of annotation\]+\[ the width of the annotation\]+\[ the height of the annotation\]+\[zoom level for the annotation\]+\[quality\]](http://129.11.191.7/Research_1/Darren/Barretts/[slide ID].sys?[x coordinate for the upper-left corner of annotation]+[y coordinate for the upper-left corner of annotation]+[ the width of the annotation]+[ the height of the annotation]+[zoom level for the annotation]+[quality])

The second, third, and fourth columns contain the annotation grade, the annotation ID, and the annotator ID, respectively.

**Table 3.2** The ground truth labels for the annotations within different subsets of the data provided by "Expert\_B" and "Expert\_E"

Expert	Before the filtration of phase one				After the filtration of phase two				Training annotations				Test annotations			
	Negative	LGD	HGD	Total	Negative	LGD	HGD	Total	Negative	LGD	HGD	Total	Negative	LGD	HGD	Total
B	103	53	57	213	92	46	55	193	84	39	44	167	8	7	11	26
E	149	20	51	220	132	17	48	197	117	14	40	171	15	3	8	26
Total	252	73	108	433	224	63	103	390	201	53	84	338	23	10	19	52

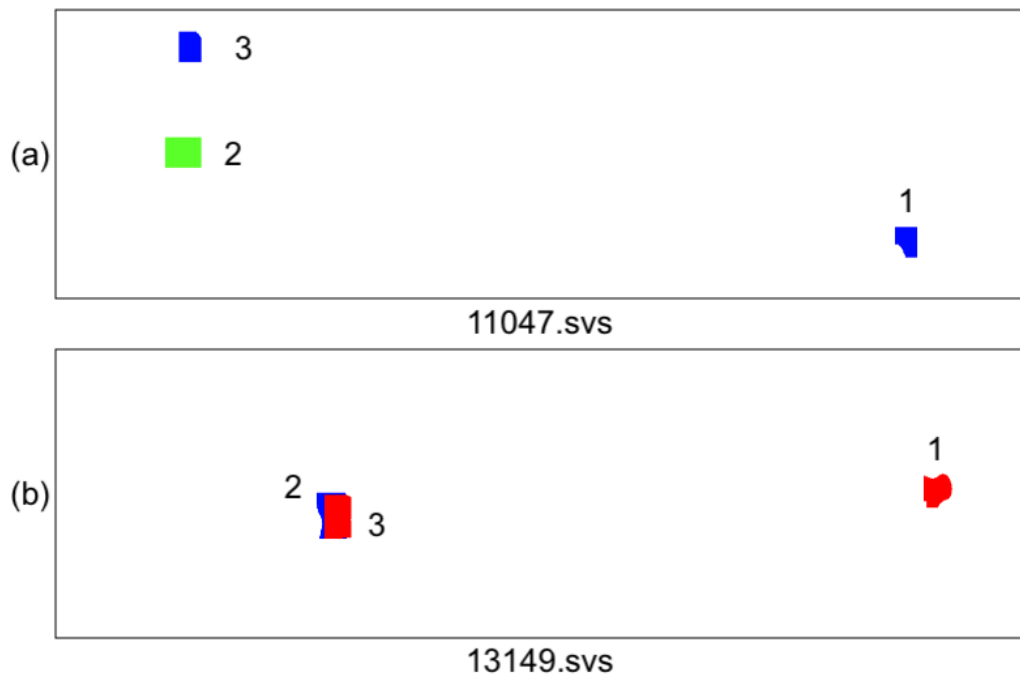
### **3.4 Virtual slides and annotated regions selections**

We rechecked all the 158 virtual slides images for quality, such as scanning resolution and tissue distortion, and each annotation was checked for technical and contextual quality. All the images and the annotated regions went through two phases of filtration. The first phase was filtering the whole virtual slides. As a result, 16 virtual slides were excluded. Eight of them have significant non-tissue artefacts, such as extra sliced wax smears covering the sliced tissue while preparing the glass slide. Two slides have abnormal staining. Three slides have blurred tissue due to issues during the scanning process. Finally, three slides include annotations graded differently by two experts or have large overlapped regions graded differently. In such cases, the exclusion criteria are referred to as misleading annotations. The interobserver agreement between the two pathologists increased to 0.7 KV and 0.797 weighted KV after removing the 16 virtual slides. Through the first phase, 33 annotations within the removed virtual slides were removed. After this phase, the virtual slides were divided into training, validation, and testing sets.

One-tenth of the remaining 143 slides were used to test the proposed approach. The test set was selected to be similar to the test set of a previous PhD research by Adam (2015), which was selected randomly and equally for each grade based on the glass slide ground truth to compare the outcomes of the two PhD. researches. Table 3.1 shows the number of available virtual slides before the filtration and after, the training set and the test set for each grade in the form of confusion matrixes for the diagnosis of the two pathologists, and their agreement calculations and scores. In this thesis, only annotations from the training and validation sets were involved in training all phases of the proposed model. The interobserver agreement for the training set is 0.796 weighted KV, and it is 0.789 weighted KV for the test set. More details are provided in Table 3.2.

The second filtering phase was only applied to the training and validation sets at the annotation-level. In this phase, annotations that attribute to grade a slide differently are accepted as long as they do not overlap with annotations of different grades. For instance, Figure 3.1 shows examples of accepted annotations (in (a)) and discarded annotations (region “2” and

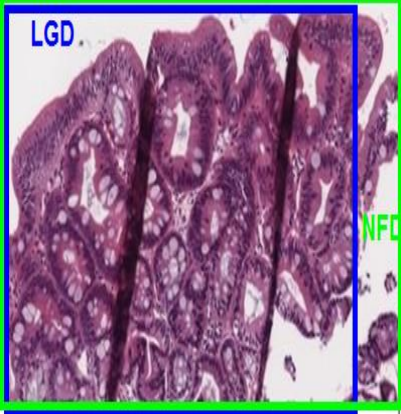
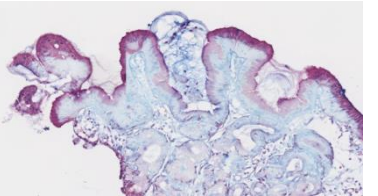
region “3” in (b)). In the first example, the region “2” was annotated by “Expert\_E” and led to NFD for the whole slide from the perspective of “Expert\_E”, while regions “1” and “3” were annotated by “Expert\_B”, and participated in grading the whole slide as LGD. All the regions were accepted because they are not considered misleading labels. However, in the second example, the overlapped regions were rejected as they will add noise to the dataset and confuse the model learning once they are included. After the second phase, ten annotated regions from the 128 slides were removed. Details about the filtered annotation are provided in Table 3.3.



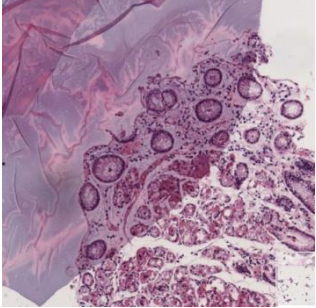
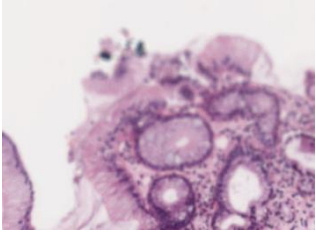
**Figure 3.1** Examples for accepted and rejected annotations

All the annotations in (a) are accepted, while in (b), regions “2” and “3” are rejected. Both green, blue and red represents NFD, LGD and HGD annotations.

**Table 3.3** Details about the filtered annotations through phases one and two

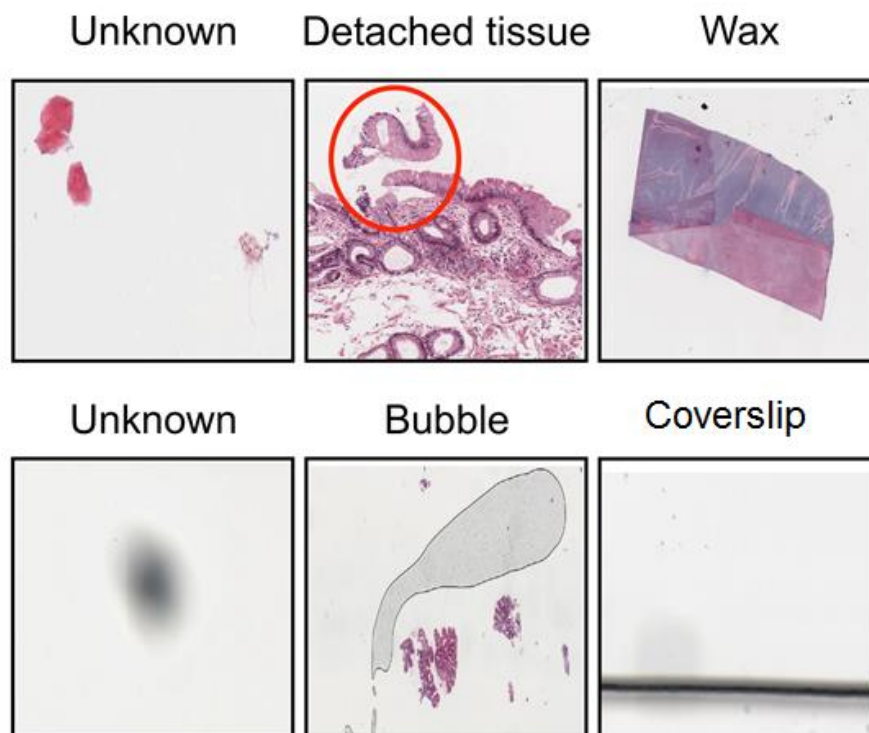
Reason	Number of regions		(Slide ID, Region ID, Grade, Expert ID)		Example
	Phase one	Phase two	Phase one	Phase two	
Misleading annotation	7	10	[(10596,32,NFD,E), (10596,33,NFD,E), (10596,29,LGD,B)], [(11002,214,NFD,B), (11002,215,LGD,E), [(13362,631,NFD,E), (13362,630,LGD,B)]	[(11051, 375,LGD,B), (11051, 376,NFD,E)], [(11051, 374,LGD,B), (11051, 377,NFD,E), (11051, 378,NFD,E)], [(13149, 533,LGD,B), (13149, 535,HGD,B), (13149, 537,HGD,E)], [(11046,347,LGD,B), (11046, 352,HGD,E)]	
Abnormal staining	2	0	(13221,578,LGD,E), (13358,617,NFD,B)	None	



Slide quality (overlapped tissue)	20	0	(10923,148,LGD,B), (10923,149,LGD,E), (10923,150,NFD,E), (10836,104,NFD,B), (10836,105,NFD,E), (10845,106,NFD,B), (10845,107,NFD,E), (10845,108,NFD,E), (10986,171,NFD,B), (10986,172,NFD,E), (11041,328,NFD,B), (11041,329,NFD,B), (11045,342,NFD,B), (11045,343,NFD,E), (11045,344,NFD,E), (11071,421,NFD,B), (11071,422,NFD,E), (11071,423,NFD,E), (13364,639,NFD,B), (13364,640,NFD,E)	None	
Blurred tissue	4	0	(10997,192,HGD,E), (10997,191,HGD,B), (11010,243,NFD,B), (11010,244,NFD,E)	None	

### 3.5 Tissue segmentation and Noise Reduction

The first phase in detecting and grading dysplasia is to detect and segment tissue regions by excluding white space background and noise (artefacts) such as wax, bubble, detached small tissue, and shadow. Examples of artefacts are shown in Figure 3.2. In whole slide virtual slides, tissue to WSI ratio usually has a minor amount compared to the background ratio. For example, Figure 3.3 shows one of the most populated slides of the test set with an 89% non-informative background. Systems that aim to analyse the whole virtual slides at high magnifications must segment regions that most likely contain only tissue to reduce the computation time and cost. In this research, the used method follows the method of tissue segmentation (Haggerty et al., 2014), and it is almost similar to the approach that was employed by the winning approach at the "CAMELYON16" challenge (Wang et al., 2016).



**Figure 3.2** Histological artefacts examples



**Figure 3.3** Tissue detection

Tissue detection visualisation for one of the testing set whole virtual slides. Boundaries in black show the detected tissue boundaries.

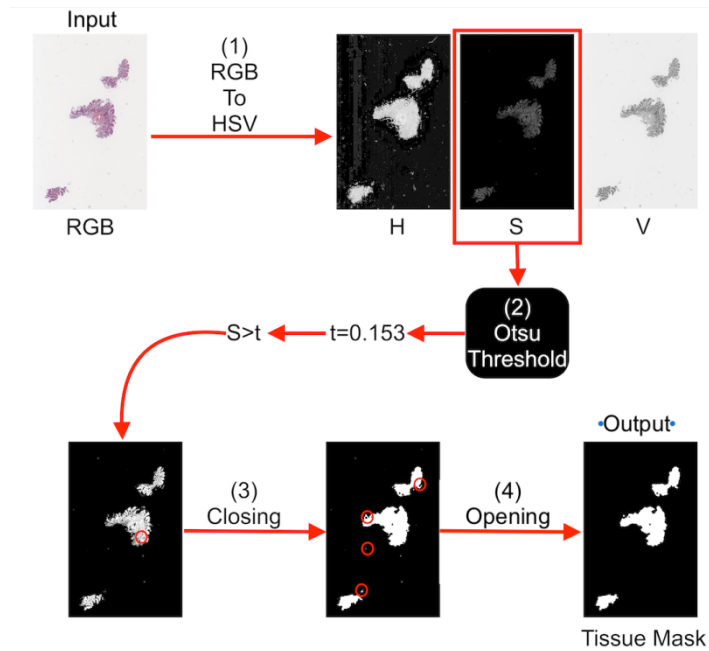
In general, all the proposed models in this thesis handle the WSIs using pyramidal approaches where the top of the pyramid takes the inputted whole slide at the lowest available magnification (0.6, 1.25, or 2.5) to detect the foreground and reduce noise following (Haggerty et al., 2014). Figure 3.4 illustrates their approach to segmenting tissue and reducing noise. In brief, the tissue segmentation (background elimination) method converts the RGB thumbnail image into an HSV image. It then applies Otsu's threshold (Otsu and cybernetics, 1979) to the saturation channel of HSV (Hue, Saturation, and Value). Algorithm 3.1 provides a summary of the background elimination. The foreground segmentation is followed by the morphological closing and opening methods to remove salt (a very small detected tissue) and pepper (a very small background in the tissue) noise.

**Algorithm 3.1** Background elimination

---

```
Data: Whole Slide image (WSI)  
Result: width, height and the top left corner coordinate of each detected foreground  
          bounding box ( $x_{coordinate}$ ,  $y_{coordinate}$ )  
 $image \leftarrow getThumbnailMagnification(WSI)$  ;  
 $h, s, v \leftarrow RGBtoHSV(image)$  ;  
 $threshold \leftarrow OtsuThreshold(s)$  ;  
 $mask \leftarrow (s > threshold)$  ;  
 $x_{coordinates}, y_{coordinates}, widths, heights \leftarrow findContours(mask)$  ;
```

---



**Figure 3.4** The applied tissue detection and segmentation methods on a whole virtual slide from the train set.

Although the previous step eliminates most of the artefacts, such as shadows and bubbles, two kinds of artefacts, wax smears and detached tissue, are still present in the detected foreground. Therefore, the detected foreground is processed at 5X magnification to eliminate these artefacts by applying two rules, and it returns detected foregrounds that are most likely to be biopsies. The conducted experiments concluded the first rule by us on Barrett's dataset, foreground bounding boxes that contain biopsies have widths and heights each more than or equal to 200 pixels (at 5X magnification), and the total length of their width and height is at least 800 pixels. By applying this rule, small and detached tissues were excluded. The second rule relied on the conducted experiments (Adam, 2015). She found that the histogram of the greyscale of each bounding box has a total number of frequencies of the bins between 190 and 210 that more than the overall mean frequencies of the histogram is considered a wax smear. The two rules are combined and illustrated by Algorithm 3.2.

### Algorithm 3.2 Tissue detection

---

**Data:** WSI,  $x_{coordinates}$ ,  $y_{coordinates}$ ,  $widths$ ,  $heights$   
**Result:** width, height and the top left corner coordinate of each detected tissue bounding box ( $x_{coordinate}$ ,  $y_{coordinate}$ )

```

factor ←  $\frac{5XMag}{thumbnailMag}$ ;
X ←  $x_{coordinates} \times factor$ ;
Y ←  $y_{coordinates} \times factor$ ;
W ←  $widths \times factor$ ;
H ←  $heights \times factor$ ;
for each (foregroundBoundingBox) x, y, w, h in X, Y, W, H do
    if w ≥ 200 & h ≥ 200 & w + h ≥ 800 then
        greyImage ← RGBtoGrey(foregroundBoundingBox);
        bins, frequencies ← histogram(greyImage);
        if ( $\sum_{bin=190}^{210} frequency_{bin} \leq mean(frequencies)$ ) then
             $x_{coordinates}, y_{coordinates}, widths, heights \leftarrow x, y, w, h$ ;
        end
    end
end
end

```

---

“OpenSlide-Python” and “Scikit” were employed in tissue detection and segmentation. “OpenSlide-Python” is a python interface for a “C” library, which was implemented to retrieve whole virtual slides. “Scikit” (Van der Walt et al., 2014) is an open-source “Python” image processing library that contains useful implemented algorithms and utilities such as “filters”, “color”, “disk”, “opening”, and “closing” modules. All the pre-processing methods were applied on the down-sampled whole virtual slide, approximately at 2.5X magnification, to accelerate the generation of tissue masks. Detecting the background does not need any cytological features that should be processed at higher magnifications. Then a tissue mask binary image is saved at the lowest magnification, usually at 0.6X. When analysing the whole slide at different magnifications, the tissue mask image can be mapped into higher magnification by Equation 3.1 and Equation 3.2.

#### Equation 3.1

$$X_t = X_c \times \frac{M_t}{M_c}$$

#### Equation 3.2

$$Y_t = Y_c \times \frac{M_t}{M_c}$$

Where  $(X_c, Y_c)$  is the coordinate at the current magnification  $M_c$ , and  $(X_t, Y_t)$  is the coordinate of the target magnification  $M_t$ .

### 3.6 Annotation masks generation

As mentioned in section 3.3, two pathologists highlighted the regions that helped them in deciding the grade of each whole virtual slide, and they assigned one of the six grades to them. Their annotations were collected, and they were re-assigned to one of the three grades. The provided annotations for this research are in the form of tabular data, as described in section 3.3, and they are rectangular annotations. In contrast to the curved annotations are the rectangular annotations. Their nature implies the existence of background regions within the annotations, especially in the case of Barrett's oesophagus, as the most important diagnostic features found in the epithelial layer, the outer layer lining of the oesophageal lumen. Sampling the dataset based on the rectangular annotation will nominate patches within the background, or at least most of their areas are background. To prevent that, the rectangular annotations were converted into curved connotations using the following technique.

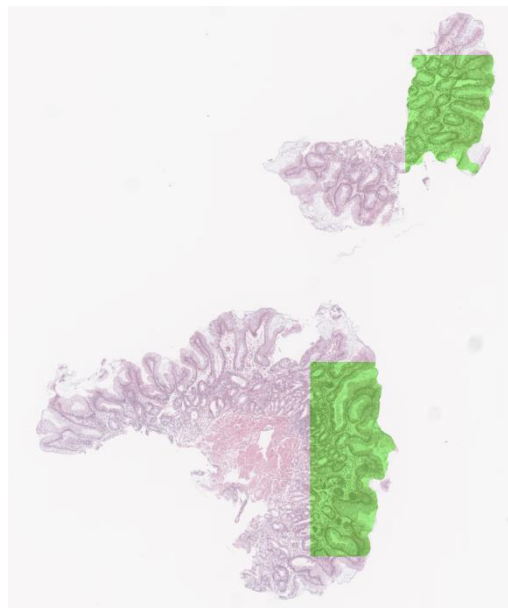
The first step is to find all the coordinates for all the annotation corners from the provided upper left corner coordinates and dimensions. Those coordinates are provided at 40X magnification, while the width and height represent the dimensions of the annotations at different scales. To rescale the dimensions, they are multiplied by the result of the target magnification divided by the current magnification. The provided zoom levels follow the "Aperio" scanner zoom level. Table 3.4 shows the "Aperio" zoom level with their correspondence standard magnifications.

**Table 3.4** Standard histological virtual slides magnifications and their associated Aperio levels

<b>Aperio levels</b>	1	2	4	8	16	32
<b>Standard magnifications</b>	40X	20X	10X	5X	2.5X	1.25X

All the contained annotations are gathered in an XML file for every virtual slide. Moreover, for each annotation, all the coordinates of its angles were calculated at 40X and recorded in the XML file along with the colour ID representing its grade. The coordinates were calculated as the following: if

the provided coordinate is  $(x,y)$  and the rescaled width and height are  $h$  and  $w$ , then the other coordinates are  $((x+h),y)$ ,  $((x+h),(y+w))$ , and  $(x,(y+w))$ . An example of a created XML file for the annotations in a WSI is provided in Appendix A.2. Then, RGB rectangular annotation was generated at magnification similar to the binary tissue mask, which is the lowest available magnification for each virtual slide. In addition, the background regions in the rectangular annotation were eliminated using the virtual slide binary tissue mask and the rectangular annotation to simplify the sampling process of the dataset. In addition, a new annotation mask was regenerated to show green, blue or red coloured foreground with curved annotation, which represents NFD, LGD and HGD, respectively, and white background, which can be either unimportant tissue or background. Figure 3.5 shows the result of a curved annotation mask for two biopsies within the training set.



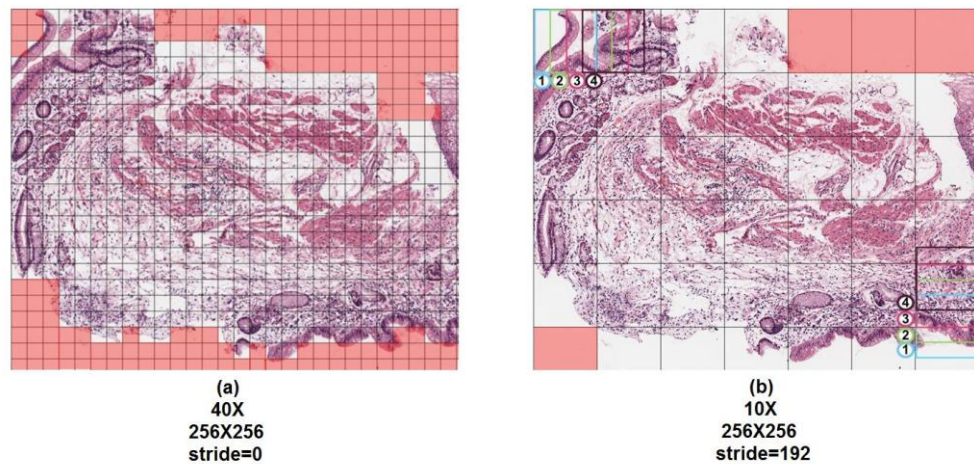
**Figure 3.5** Visualisation for the regenerated curved NFD annotation mask

### **3.7 Sampling patches from annotated regions**

The dataset used in this thesis contains whole virtual slides ranging from 1.5 GB to 48 GB. Due to their sizes, it is impossible to analyse these images or their annotations at once. Thus, patches were sampled at two magnifications from the provided annotations of the training set for further patch-based analysing. The first sampling phase was conducted at 40X magnification to analyse the WSIs at 40X magnification aiming to capture the cytological



changes. Each accepted annotation within virtual slides belonging to the training set was split via grid into non-overlapping image patches of  $256 \times 256$  pixels. Patches containing 50% of unannotated tissue or background were discarded, as shown in Figure 3.6 (a). The size of  $256 \times 256$  pixels at 40X magnification was chosen as it is commonly used in the “CAMELYON16” competition, and it gave the best results. Also, it was observed that this size captures the best cytological features, and it can be fed into most deep-learning architectures. In addition, the non-overlapping technique was chosen due to the large number of sampled patches at that magnification. Also, the aim is to use them in texture analysis for dysplastic tissue in Barrett’s oesophagus histopathology, which does not rely on the aggregation of other features. For instance, the nucleus's size, colour, shape, and presence of goblet cells are used for this type of analysis and not how the nucleus is arranged in the epithelial layer. More than 1.28 million patches were sampled at 40X magnification. Details about the total patches of each category are provided in Table 3.5.



**Figure 3.6** Patches sampling at two different magnifications

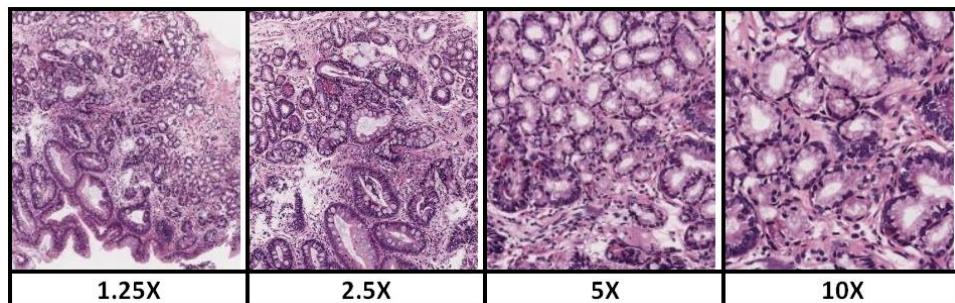
Figure (a) shows sampling non-overlapped  $256 \times 256$  patches at 40X magnification. Figure (b) shows sampling three-fourths overlapped  $256 \times 256$  patches at 10X magnification.



**Table 3.5** The number of extracted patches from annotations at 40X and 10X

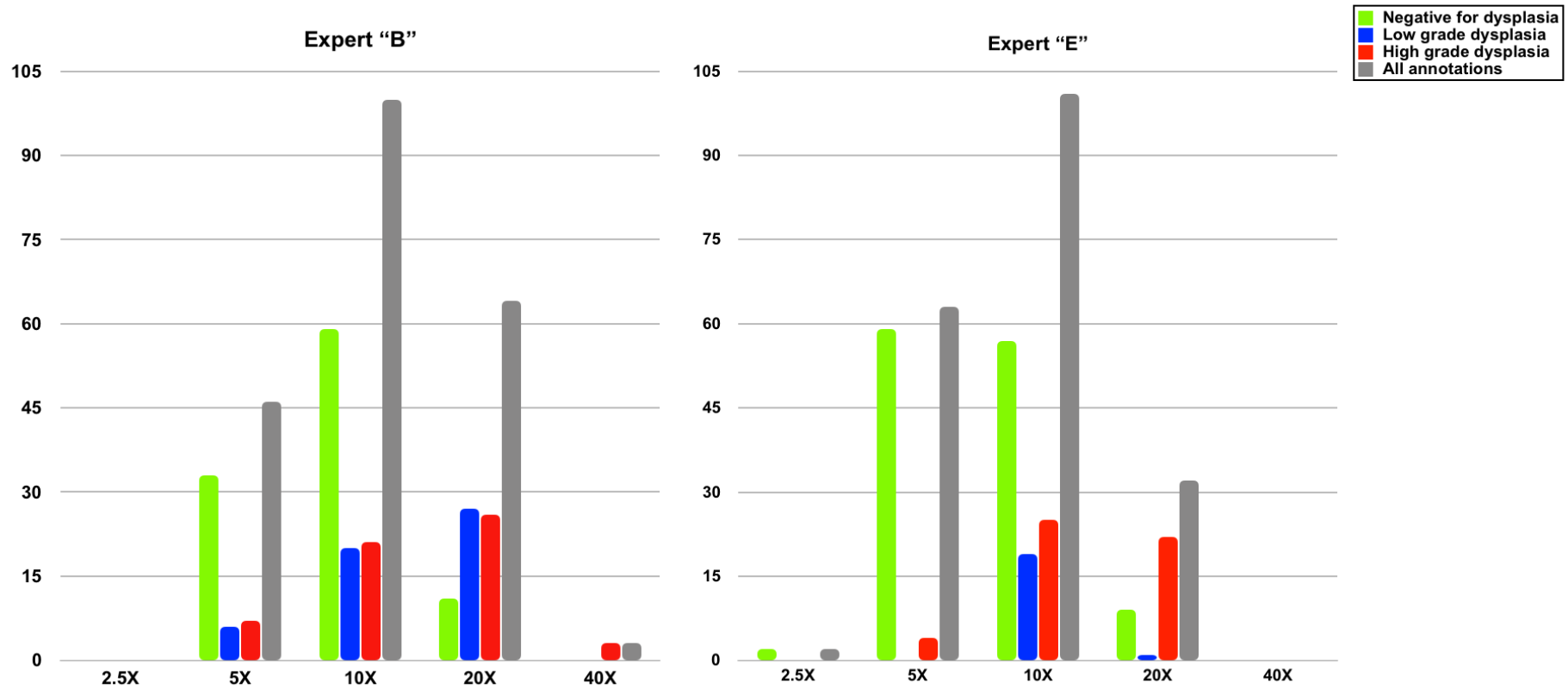
	40X	10X		
	Training set	Train	Validate	Total
NFD	979,561	273,784	48,233	322,017
LGD	176,840	47,852	16,607	64,459
HGD	132,870	39,603	26,139	65,742
Total	1,289,271	361,239	90,979	452,218

The second sampling phase was conducted around 10X magnification. Statistically, the pathologists used 10X magnification intensively to diagnose the annotated regions. Figure 3.8 shows column charts that count each pathologist's number of annotations at every magnification. Based on our observation, that magnification was the lowest-level magnification out of the available magnifications (1.25X, 2.5X, 5X, 20X, 10X and 40X) that captures the architectural changes while preserving the quality of the cytological features. Figure 3.7 shows the different annotations at different magnifications (1.25X, 2.5X, 5X and 10X) from the same tissue that belongs to image “11006.svs”. At 10X magnification, sampling patches with the size of 256 × 256 pixels capture nucleus shapes and arrangements and the glands shapes and arrangements in the lamina propria. While at 1.25 and 2.5X magnification shows the shape of crypts and glands and their arrangements only. Besides, sampling at 1.25X, 2.5X and 5X magnifications led to insufficient data to train a deep-learning architecture.



**Figure 3.7** Demonstration of sampling 256x256 patches from the same tissue at four different low magnifications

This figure investigates the best magnification that captures the best architectural arrangements and preserves some of the cytological features



**Figure 3.8** Column charts for the number of annotated regions at the available magnifications

The column charts show each expert's number of annotated regions for each dysplasia. Both pathologists increase the magnifications when annotations have a higher dysplasia degree. "Expert\_B" tended to have higher magnification in his annotation than "Expert\_E".

Each annotation was sliced into 256×256 pixels patches that overlapped with a stride of 64 pixels in the second phase. That means the next upper-left corner of the new patch will start at 64 pixels on the right or below the upper-left corner of the previous one, as illustrated in Figure 3.6 (b). Patches were rejected once they contained more than 90% background, unannotated tissue, artefact, or white background. The percentage of 90% was chosen to guarantee the extraction of tissue from the surface of the epithelial layer, which prominently contributes to grading dysplasia in Barrett's Oesophagus. As a result, more than 452K patches were sampled for training deep-learning architecture that detects regions of interest and analyses the WSIs at 10X magnification. The distribution among the different grades is provided in Table 3.5. Also, details about the sub-datasets used for each experiment are provided in their chapters (Chapter 4 and Chapter 5). Figure 3.9 shows random samples from NFD, LGD, and HGD at 40X and 10X magnifications.

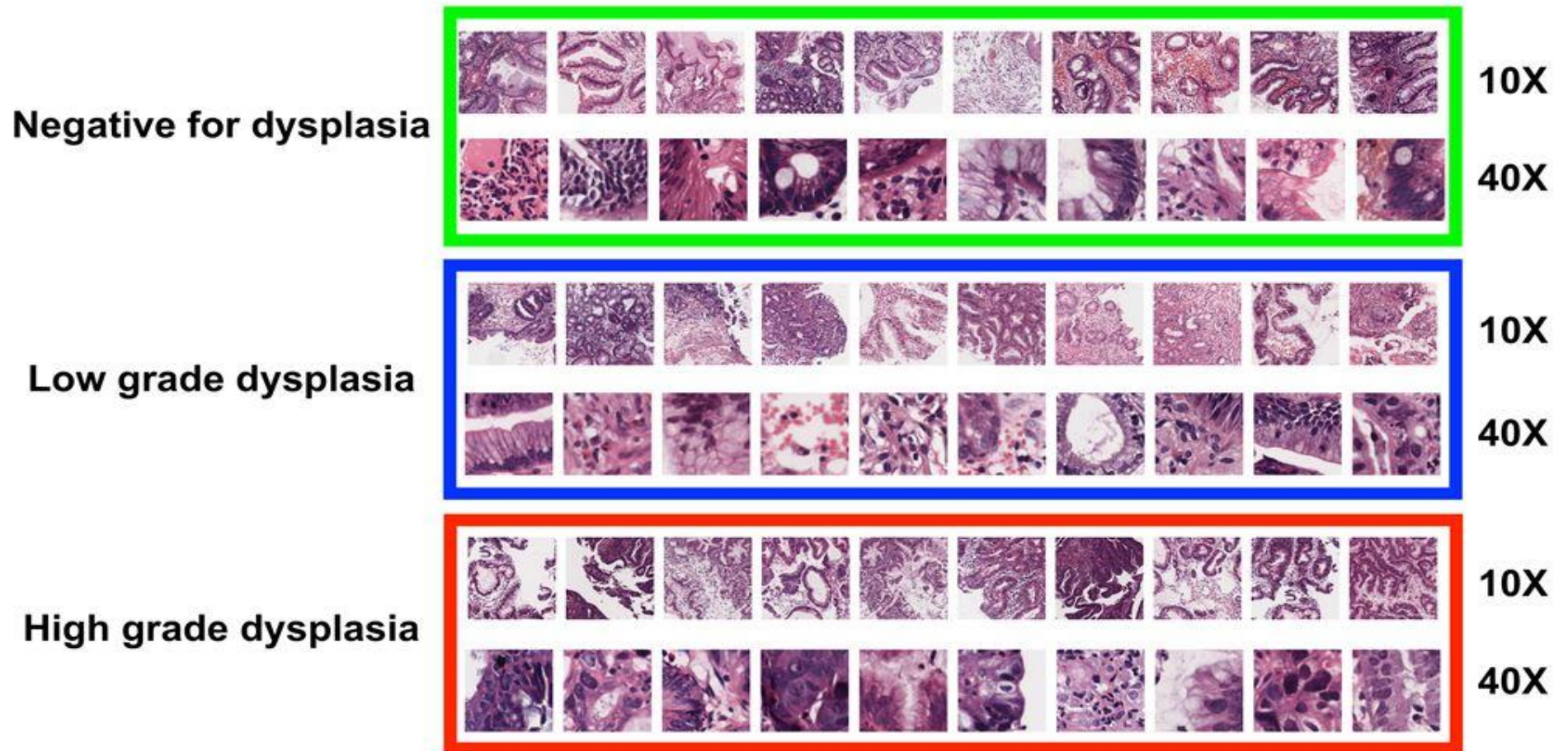


Figure 3.9 Samples for the extracted patches at different magnifications from different grades of dysplasia

### **3.8 Sampled patches pre-processing**

Input pre-processing is crucial in transfer learning because each model pre-trains to a specific large-scale dataset with special input requirements. For instance, all the experiments that were conducted in this research follow transfer learning by using a pre-trained Keras “Inception-ResNet-v2” model on the “ImageNet” dataset (Szegedy et al., 2017) (Deng et al., 2009). Thus, input to the transferred model includes training, validation, and testing stages that should follow the same input transformation method, which was used as data pre-processing before training the original Keras model before feeding them into the transferred model. Keras's “Inception-ResNet-v2” requires the input images to be sample-wise range normalised to a fixed range between -1 and 1 instead of 0 to 255 to speed up the training. Besides the range normalisation, some random methods for image pre-processing were adopted to reduce overfitting and make the model invariant to translation, rotation and flip. Those parameters are a random rotation by 20 degrees, a random 20% total of each width and total height, and a random horizontal flip. The type of augmentation was selected carefully to reduce overfitting and not affect the classification task. The augmentation choices and settings were similar to the pre-trained Keras “Inception-ResNet-v2” input pre-processing configurations, except that the image enlargement choice was excluded. Because enlarging the images is not an option for the thesis dataset, it might not learn the feature of nuclear enlargement in the case of HGD, which might result in misclassifying NFD with HGD. In future work, a random full range of rotation should be considered as a mean of augmentation; because it is expected to increase the performance of the trained models.

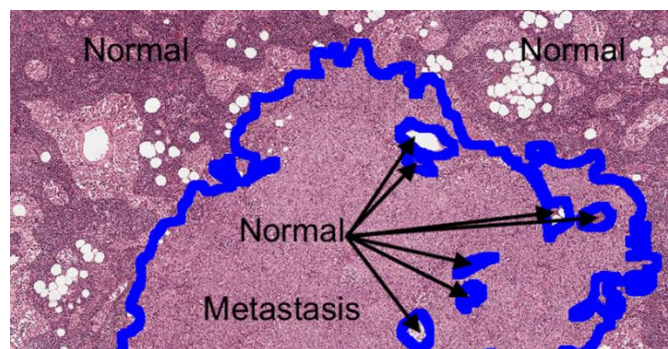
### **3.9 Dataset challenges**

The nature of dysplasia development in the oesophagus is vague, as was described in detail in Chapter 2. The continuous spectrum form of abnormalities that the dysplastic oesophagus tissue follows throughout their development leads to an inability to define clear criteria to diagnose the dysplastic tissue at the stages between NFD and LGD and the later and HGD. This ambiguous situation results in disagreement between pathologists when they diagnose tissue at those stages and disagreement between pathologists themselves while diagnosing the same sample on



different occasions. The used dataset has 84.18% agreement between “Expert\_B” and “Expert\_E” with 0.689 KV, which reaches a good yet not excellent level. Also, the test set has intraobserver agreements of 0.8 and 0.7 KV for “Expert\_B” and “Expert\_E”, respectively, when they diagnosed the same sample using the microscope for the glass slide and screen for its virtual slide.

Moreover, the provided annotations are imprecise, and that claim can be proved correct by three related facts. The first fact is the domain experts' technique to annotate regions of interest, which was arbitrary and not exhaustive. They surrounded the region of interest with rectangles. They assigned a general label to each rectangular region without localising the dysplastic tissue, which can be seen in the overlapped annotation from different grades. In other words, regions labelled as high grade might have regions of LGD and NFD. Still, the prominent features belong to HGD that could be recognised by the pathologists' brains after years of experience but might be hard to be achieved by machine learning algorithms. During the annotation, the pathologists did not localise each grade of dysplasia within their annotations as the pathologists of the “CAMELYON16” dataset did. Figure 3.10 shows an example of a region in the “CAMELYON16” dataset that was exhaustively annotated.



**Figure 3.10** An annotated region from the "CAMELYON16" dataset

The blue borders separate the normal tissue from the metastasis (cancerous) tissue.

Even if the pathologists accurately annotated the dataset, it would suffer from imprecision. That is due to the complexity of the dysplastic abnormalities in Barrett's oesophagus and the definitions of different dysplasia grades that have toleration to some degree of the presence of

higher-grade abnormalities in lower grade tissue, refer to section 2.1, and that is the second supported fact. Finally, the fact that most of the annotated regions were examined at lower magnifications implies that some of the regions at a higher magnification do not necessarily belong to the degree of dysplasia assigned to the annotated region. For example, when a pathologist assigns an NFD label to a region at 10X magnification, there is a chance that one of the contained nuclei has multiple nucleoli, which is a dysplastic feature, and that is observable at higher magnification such as 40X. Figure 3.8 shows column charts with the number of annotations at the available magnifications used in the examination by “Expert\_B” and “Expert\_E”. It shows that 10X magnification is the most used magnification for NFD grade and LGD grade, while in grading HGD, the 20X magnification joins 10X in importance. These statistics suggest that the pathologists tried to balance the architectural and the cytological features. They zoomed in whenever they were urged to examine more cytological features, such as the number of nucleoli in the nuclei.

### **3.10 Summary**

Two domain experts provided two types of ground truth labels individually: labels for whole virtual slides and labels for annotated regions within the whole virtual slides. The dataset has gone through two stages of filtration, one on the whole virtual slides level and the other on the annotation-level. As a result, 128 slides were used to train the models with 338 annotations, and a separate set of 15 slides was used to test them.

As earlier discussed, whole virtual slides are huge and analysing them after applying tissue detection. Segmentation reduces the computational time and cost by allowing different methods to analyse only the region where the probability of dysplasia is more likely to occur. For the test set of whole virtual slides, the average size of the images is 22.20 GB. After applying tissue detection and segmentation, the nominated regions were shrunk to 1.51 GB on average. In other words, the applied approach successfully removed 93% of every whole virtual slide. More detailed information about the test set images and their sizes before and after tissue detections are provided (Appendix A.1). Moreover, the implemented tissue detection method eliminated all the noises that appear in Figure 3.2, except large

pieces of wax and significant detach tissue. In general, eliminating the remaining noises is discussed in section 6.2.

After tissue detection, the proposed sampling technique has sampled patches with different sizes and amounts and at different magnifications. Besides, it has the option of extracting overlapped patches. More than 450K and 550K samples were extracted at 10X and 40 X magnifications.

In the end, a comprehensive explanation of the challenges faced using Barrett's oesophagus related dysplasia dataset. Some of them are directly connected to the nature of the tackled issue, and some are related to the annotation method that the domain experts followed.



## **Chapter 4. Regions of Interest Detection and High-level Analysis and Classification**

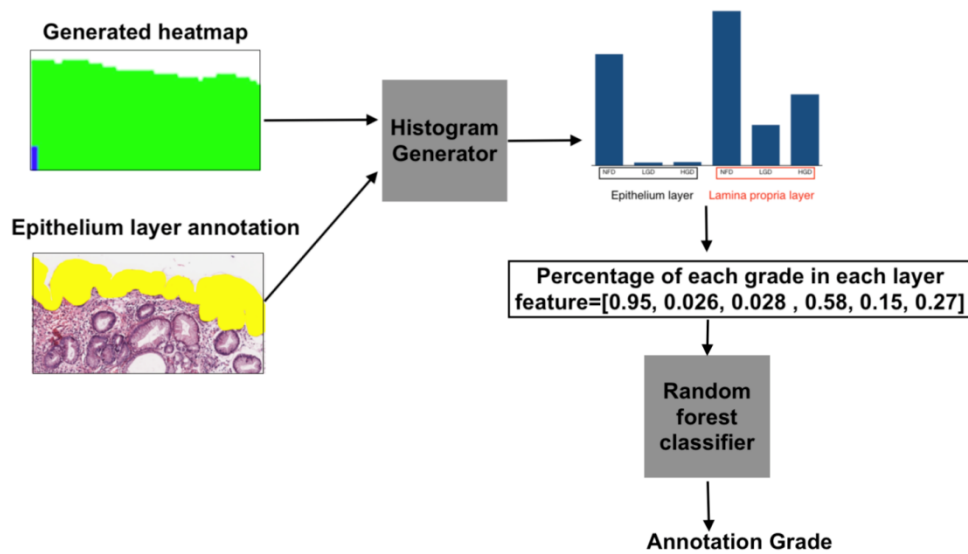
### **4.1 Introduction**

Deep convolutional neural networks have achieved outstanding results that attract many researchers in computer vision. Due to the complex form and huge data in histopathology, deep-learning approaches are closely attached to it. Amongst them, supervised approaches are the most common, which require accurate annotations. However, this is not the case in most of the histological images. That is attributed to the nature of the histological image, which follows the coarsely grained labelling. Thus, most of the available datasets, including the used dataset in this thesis, are considered a MIL problem (see section 2.5.3), a form of weakly supervised learning.

Moreover, training a supervised deep-learning model from scratch requires a vast labelled dataset, which is costly, especially in the medical field. Therefore, adapting transfer learning appears to be a promising solution. This chapter presents an approach to classifying and deciding the grade of dysplasia in Barrett's oesophagus by adapting transfer learning using a high-level analysis, at 10X magnification, of the histological images. Equally important is implementing a region of interest detector, as employing a CAD system in the histopathology field is feasible when a reliable algorithm for detecting crucial areas of interest is used. This step is derived from the need to reduce the cost and time of applying sophisticated algorithms for processing and analysing WSIs while keeping the accurate capture of all the critical regions.

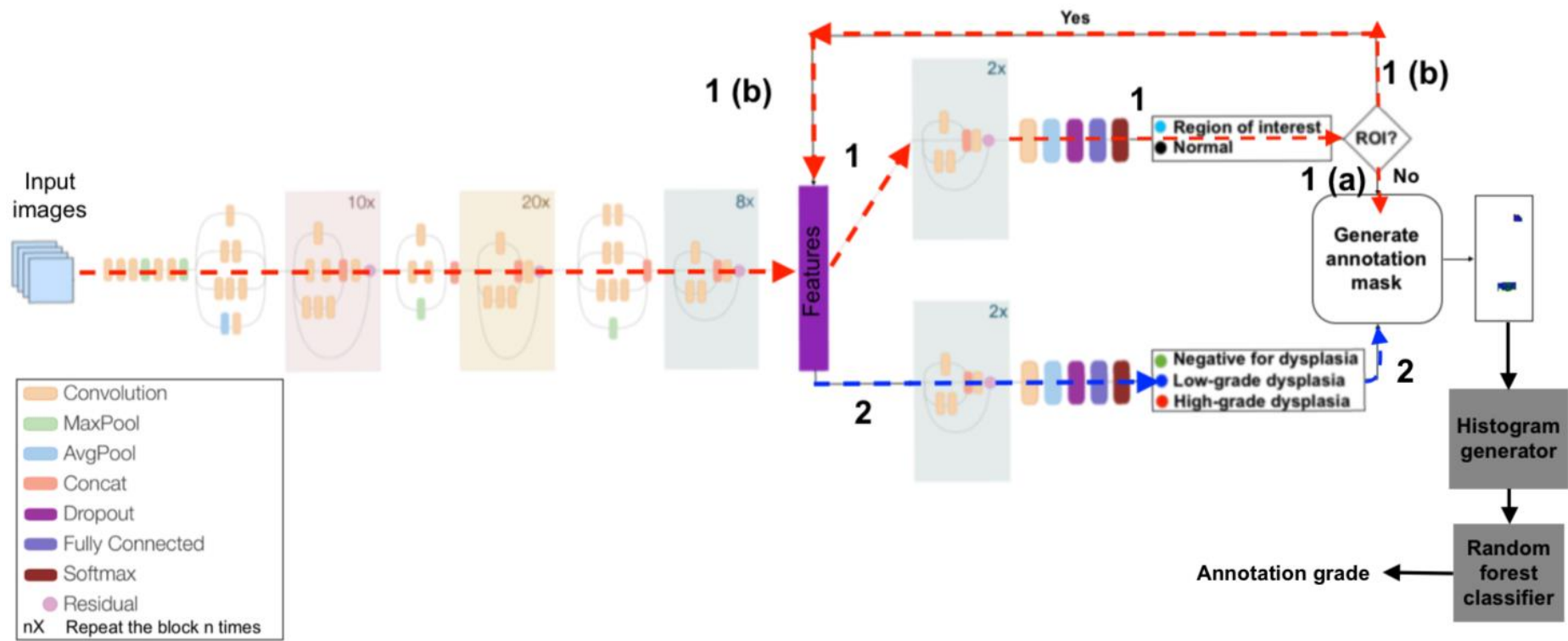
The previous chapter demonstrates the WSI preparation and patch sampling to be analysed by the proposed models in the following two chapters. In this chapter, the MIL approach was adopted to train the deep-learning model, as it employs the reported grade of an annotation to assume the label of its sampled patches. Every sampled patch was passed to convolutional layers to obtain low-level features. Then, these extracted features were fed into two further subnetworks of convolutional layers. One path determines whether the input patch belongs to a region of interest. The other extracts important

high-level features from it and classify them based on the three-tier classification discussed in Chapter 2. Then, the patch-level predictions of the annotation were aggregated to assemble a heatmap for its corresponding annotation. The heatmap is a thumbnail of the fed image (an annotation or WSI) containing four colours, white, green, blue and red, that indicate the neglected background or tissue, NFD, LGD and HGD regions, respectively. After that, an annotation-level histogram was used to train a random forest classifier to infer the annotation grade. The built histogram represents the distribution of each grade (NFD, LGD, and HGD) in each layer (the epithelium and the lamina propria layers) separately. That was achievable with the assistance of epithelial layer masks, That was achievable with the assistance of epithelial layer masks, which the researcher annotated manually. Figure 4.1 shows an overview of the annotation-level inference submodel.



**Figure 4.1** An overview of the annotation grade inference submodel

The proposed model is overviewed in Figure 4.2. The input dataset represents the sampled patches at 10X magnification, which was observed to be the lowest available magnification that could capture the tissue arrangements within the sampled patches. This chapter presents a weakly supervised deep-learning model to analyse and classify annotated regions at a high level (10X magnification). Also, it includes methods and experiments that were conducted to detect regions of interest and extract features from rectangular annotations.



**Figure 4.2** The proposed model for regions of interest detection and dysplasia classification

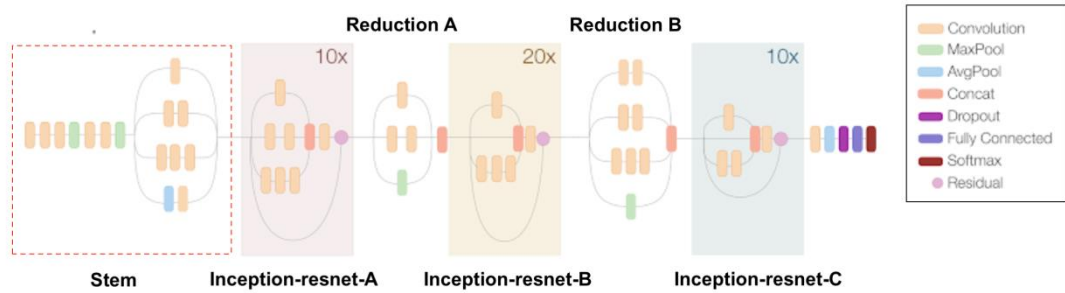
This chapter's main contribution to research is introducing a novel MIL weakly supervised deep-learning architecture for a single network that detects regions of interest and grades them sequentially using images at 10X magnification. The model combined two trained models following a fork-style to work cooperatively. Another contribution is a solution to infer the annotation-level grade. The inference technique considers the pathology guidelines to diagnose Barrett's oesophagus dysplasia. The degree of cytological and architectural changes in the epithelial and lamina propria layers varies in a dysplasia grade.

## 4.2 An overview of the proposed model

In this chapter, the proposed model has two phases. The first phase has two connected networks that work sequentially to detect regions of interest and classify dysplasia in the detected regions. The two networks adapted the Keras pre-trained "Inception-ResNet-v2"<sup>1</sup>. The Keras version of that model has the same concept as the proposed model by Szegedy et al. (2017); refer to section 2.5.2 for information about the inception and residual networks. However, it added five more blocks from each of "Inception-resnet-A" and "Inception-resnet-C" and ten more blocks from "Inception-resnet-B". Figure 4.3 illustrates the structure of the Keras version, and a detailed summary of the architecture with the input and output size for each layer is provided in Appendix B.1. Also, both parts had the same structure and were trained in the same manner, as will be discussed in sections 4.3 and 4.4, yet using different datasets (see section 4.5). The first part was trained to detect regions of interest, and its conducted experiment and results will be discussed in section 4.6.1. The second part was trained to classify grades of dysplasia in Barrett's oesophagus based on the image analysis at 10X magnification. Then, the two parts were combined in a way, as illustrated in Figure 4.2, to work sequentially. After detecting and classifying each sampled patch in the annotation, a heatmap for the annotated region was generated to be fed into the second phase of the model.

---

<sup>1</sup> [https://github.com/keras-team/keras-applications/blob/master/keras\\_applications/inception\\_resnet\\_v2.py](https://github.com/keras-team/keras-applications/blob/master/keras_applications/inception_resnet_v2.py)



**Figure 4.3** An overview of the architecture of Keras Inception-ResNet-v2

Clinically, one of the crucial guidelines that pathologists follow to decide the grade of dysplasia is that an annotation is highly influenced by the grade of the dominating grade on the epithelial layer within it, and the architectural features for the glands in the lamina propria should be considered but to a lower degree. This rule was inspired in inferring the grade of the annotations in this thesis, as the contribution of each grade in each layer results in different diagnoses. Thus, the second phase of the model was designed to decide the annotation grade using labels from each pixel within the generated heatmap based on its location. This part takes two inputs, the generated heatmap from the first phase and its corresponding epithelial layer mask (see section 4.5). It calculates the frequencies of each grade in each layer (the epithelium and the lamina propria layers) to generate a histogram with six bins (NFD, LGD and HGD in the epithelial layer and NFD, LGD and HGD in the lamina propria layer). After that, percentages of the frequencies against the total number of the labelled pixels in their corresponding layer produce a feature vector for the annotation. In the annotation inference phase, the annotation feature vectors were used to train a supervised learning classifier (random forest in our case) to predict the classes of the test annotations. Figure 4.1 illustrates the second phase.

### 4.3 Methodology

The proposed model consists of two main parts. One part detects regions of interest as patches and classifies them into one of the three grades sequentially. The other part is to determine the class of the annotations using their generated heatmaps from the first part of the model.

### 4.3.1 Regions of interest detection and dysplastic classification

This part comprises the first 38 residual inception blocks of the Inception-ResNet-v2 network with two identical branches. Each consists of the last two residual inception blocks, a convolutional layer, an average-pooling layer, a dropout layer, a fully connected layer, and a Softmax classifier. The model accepts 256x256x3 sized images, and the 38th residual inception block produces a 6x6x256 sized feature representation that is cached before it is fed into the branch of the region of interest network. If the region of interest detector flags the image as of interest, its cached representation goes through the dysplasia classifier; otherwise, it is discarded. One trivial solution is to add a class “Normal” to the three grades of dysplasia that contains all the examples of uninteresting regions, such as healthy oesophagus tissue, noise, or the oesophagus’s layers that are not contributing to diagnosing dysplasia. However, this solution is not applicable due to the imbalance in the dataset, as the “Normal” class has a large variety of examples. The proposed model solved this issue by training a network to discriminate the interesting regions and another to classify them.

For training the model, a deep transfer learning fine-tuned feature extraction strategy was utilised for training two “Inception-ResNet-v2” networks before combining them. In addition, two Softmax classifiers were used to classify the extracted features, a classifier to discriminate “Normal region” from “Region of interest” and another to classify NFD vs LGD vs HGD. A Softmax regression function is a general form of the supervised logistic regression function. It supports the direct classification of multiple mutual exclusive class problems. The Softmax regression classifies a given instance by calculating a score for each class, known as the logit layer, that is a vector of a size equal to the number of the problem classes, and then it applies the normalised exponential function to the calculated scores to estimate the probability of each class (Géron, 2017). The Softmax function takes the extracted feature representation of an instance in the dataset, with a size of 768 in all the experiments of this thesis, as an input. It then trains an additional fully connected layer to reduce the cost function using an Adam optimiser.

The form of the Softmax regression hypothesis is defined in Equation 4.1. Given a training instance  $(x_i, y_i)$ , where  $x_i$  is the input image to the model

and  $y_i \in \{1, 2, \dots, k\}$  is the corresponding label, and  $k \in \mathbb{R}$  is the number of classes in the classification problem. Also, given  $h_i$  is the output feature representation for the input  $x_i$  from the learned model and  $h_i$  is equal

to  $\begin{bmatrix} h_{i1} \\ \cdot \\ \cdot \\ h_{ic} \end{bmatrix}$ , where  $c$  is the size of the feature representation, then the logits layer

$l$  is calculated as the following equation:

$$l_i = \begin{bmatrix} h_i \cdot W_1 + b \\ h_i \cdot W_2 + b \\ h_i \cdot W_3 + b \\ \vdots \\ h_i \cdot W_k + b \end{bmatrix}$$

The Softmax regression function, see Equation 4.1, basically is a set of  $k$  linear classifiers  $P(y_i = j|x_i, \theta)$  resulting in non-zero probabilities to each element to prevent calculating  $\log(0)$  in some of the loss functions, where  $\theta$  is the classifier set of parameters  $(W, b)$ .

**Equation 4.1**

$$Softmax(x_i) = \begin{bmatrix} P(y_i = 1|x_i, \theta) \\ P(y_i = 2|x_i, \theta) \\ P(y_i = 3|x_i, \theta) \\ \vdots \\ P(y_i = k|x_i, \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^k e^{h_i \cdot W_j + b}} \begin{bmatrix} e^{h_i \cdot W_1 + b} \\ e^{h_i \cdot W_2 + b} \\ e^{h_i \cdot W_3 + b} \\ \vdots \\ e^{h_i \cdot W_k + b} \end{bmatrix}$$

In this chapter, all the conducted experiments utilised the categorical cross-entropy loss, also called Softmax loss, as the loss function. Categorical cross-entropy loss is used with single-label instances in multiple-class categorisation problems, where only one label prediction is accepted. The formula of the categorical cross-entropy loss ( $CCE$ ) of a classifier  $f$  is provided in Equation 4.2, given the number of the trained dataset  $n$  and  $y_{ij}$  which corresponds to the  $j$ th element of the one-hot encoded label  $y_i$ .

**Equation 4.2**

$$CCE(f) = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k y_{ij} \log\left(\frac{e^{h_i \cdot W_j + b}}{\sum_{o=1}^k e^{h_i \cdot W_o + b}}\right)$$

### **4.3.2 Annotation-level and slide-level inference**

This part of the model consists of two other parts. The first one is to build a histogram for the generated heatmap from the previous submodel by counting the number of pixels within the epithelial layer and the lamina propria layers with the guidance of the epithelial layer annotation mask (see section 4.5). Then the percentage of pixels for each grade in each layer is calculated to form a feature vector with six elements.

The second part is composed of a random forest classifier trained to classify each feature vector representing an annotation. Breiman (2001) proposed the random forest classifier, one of the ensemble algorithms that relies on voting for the class based on the predictions of different independent and random decision trees. Each decision tree is grown on a different random subset of the training set to vote for a class, and the average of the voted classes is computed to decide the annotation grade.

For annotation inference in Chapter 4, Chapter 5 and Chapter 6, the proposed annotation inference system was employed to determine the final annotation grade. Additionally, the grade of each WSI in this chapter and Chapter 5 was decided based on the highest grade of any annotation within it once it has a proper size, as discussed in section 4.6.2.

## **4.4 Experimental design**

The proposed model for detecting regions of interest and classifying dysplasia in Barrett's oesophagus based on the analysis of annotations at 10X magnification has an architecture that involves different main parameters that need to be carefully chosen, such as the number of filters and their sizes. Since the model adopted transfer learning for the "Inception-ResNet-v2" network, the architectural structure for the deep neural network, the number of convolutional layers, the number and sizes of kernels in each layer, their strides and padding techniques, and the activation function used after each layer were not considered in setting the parameters of the thesis' experiments; because they were predefined during the design of the "Inception-ResNet-v2" network and its training using the public "ImageNet" dataset. The choice of the pre-trained model on the "ImageNet" dataset was



motivated by the fact that most robust neural networks have many parameters adding to the burden of training them with an extensive dataset, which can be very computationally intensive if they were trained from scratch. Even though the “ImageNet” dataset is not related to the medical field, it was found by Kohl et al. (2018) that using a model pre-trained on the “ImageNet” dataset to classify histological images for breast cancer outperformed the performance of the model when it was pre-trained on the CAMELYON dataset, which is a problem-related dataset, as both datasets are H&E stained histological images. Thus, the availability of models pre-trained on the “ImageNet” dataset is considered a good start for fine-tuning them depending on the current task.

For the experiments on regions of interest detection and features classification, two “Inception-ResNet-v2” networks were fine-tuned separately on two different datasets. The first network used 256x256x3 sized images from the “region of interest” dataset. The second network used the extracted patches from Barrett’s oesophagus dataset at 10X magnification (see section 3.7 for further details). The data augmentation feature, provided by “Keras”, was used with the input images of the networks to make the model invariant to translation and rotation. The final fully-connected layers of size 1000 neurons in the pre-trained networks were truncated in both experiments. Fully connected layers replaced them with several neurons equal to the number of classes in each task (two and three neurons in the first and second networks, respectively). For each network, the fine-tuning strategy was started by freezing all the networks layers except the fully-connected layer; then, the newly appended layer was trained for 2000 iterations with a minimal learning rate ( $1e^{-5}$ ). After that, the last two residual inception blocks, starting from “conv2d\_197”(refer to Appendix B.1), and their following further layers were set to be trainable and trained until convergence. With fine-tuning, a small learning rate ( $1e^{-5}$ ) was used to avoid distorting the good quality learnt weights. Also, an Adam optimiser was used without weight decay and with 0.99  $\beta_1$  and 0.999  $\beta_2$ . A mini-batch size of 128 images was used. Binary cross-entropy (see section 5.3.1.1) and categorical cross-entropy (see section 4.3) were used as loss functions for the first and second networks. The fine-tuned models were trained on separate training sets and were evaluated on separate validation sets after iterating over all the instances in the training sets. In training the networks, learning rate schedulers and early-stopping methods were utilised to help in

enhance and monitor the fine-tuning process. “Keras” supports these two methods through the “ReduceLROnPlateau” and “EarlyStopping” callbacks. Early stopping is a regularisation technique used to prevent the trained model from overfitting. Overfitting occurs when the model starts to learn the statistical noise in the training set instead of learning the mapping. Thus, early stopping tends to observe the loss of the validation set while decreasing the loss of the training set. It stops the training when the validation loss starts to increase or stops its improvement for several predefined numbers of iterations.

The “ReduceLROnPlateau” callback was set to monitor the loss of the validation sets. It was set to reduce the learning rate by multiplying it by 0.05 after the monitored validation loss stopped its improvement for a predefined number of consecutive iterations, also known as epoch patience ( $E_p$ ). The epoch patience was set to 2 in the experiment. At the same time, the “EarlyStopping” callback monitors the validation set loss and stops the training when the validation loss starts to get worse or stops its improvement. In most cases, the first sign of a slight degradation in loss or no improvement might not be wise to stop training, as the model may face a plateau before reaching the optimum solution. A delay was added to the trigger, so the training will be stopped if there is no improvement or if the model keeps slightly worsening after three epochs. Usually, using an early stopping method accompanied by setting a large epochs number, the model will stop the training at the right time and save the last best performing model.

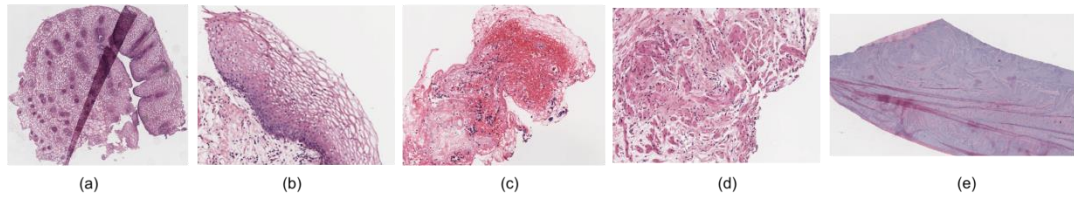
As a result of fine-tuning the two models until convergence, the two models had identical weight sets in the first 38 inception residual blocks and different weight sets starting from the last two blocks as they were fine-tuned. To avoid feeding an image twice into the same first part of the networks and to reduce the computational cost and time, the two networks shared the first 38 inception residual blocks. After that point, each network had its remaining layers and classifier. After the first part of the network, the resulting feature representation (the output from the “block8\_8\_ac” layer, refer to Appendix B.1) for an input image is saved temporarily. Then, it is fed into the region of interest network to determine whether it is a crucial region in diagnosing dysplasia or should be marked as background. If it is marked as an important region, then the saved feature representation is re-fed into the

network of dysplasia classification. A mask generation module gathers the results of the proposed model to draw an annotation mask.

Finally, the generated annotation masks of the training set were used to find histograms of each grade distribution in each layer (the epithelium and the lamina propria layers) using the manually annotated epithelial layer masks (see section 4.5). Then, feature vectors with the percentage of pixels belonging to each grade in each layer are stored to train a random forest classifier. The "Scikit" library provides the "RandomForestClassifier" class to train the random forest classifier. The "bootstrap" feature is set to "True" to use subsets of the training set to build different decision trees. The "bootstrap" feature is set to "True". The "n\_estimators" is another feature that needs to be set to specify the number of desired decision trees. The larger number leads to better performance but slower training. In this experiment, the number of decision trees was set to the default number of 100.

## **4.5 Datasets**

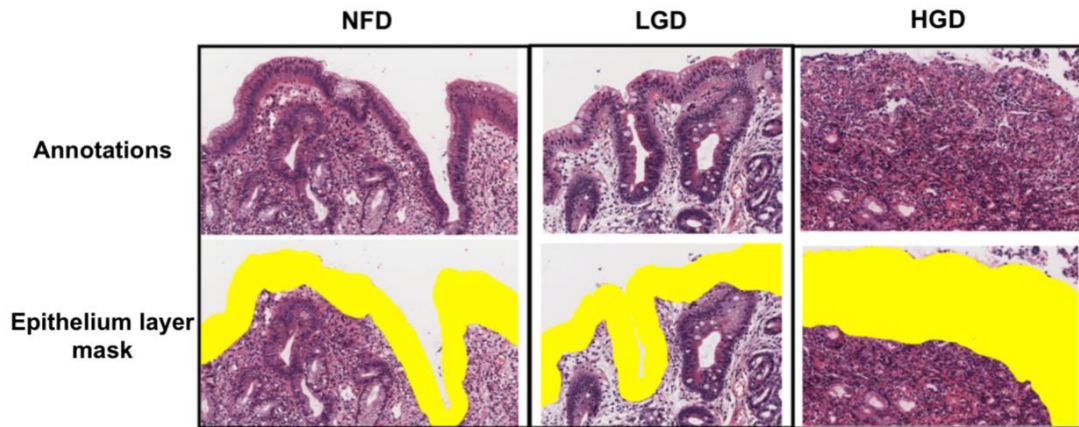
On the one hand, the first phase of the proposed model used two datasets derived from Barrett's oesophagus WSIs. The first dataset is for the regions of interest detection model, and it contains patches sampled from regions of interest and unimportant regions. The instances in the region of interest are the sampled patches at 10X magnification from the provided annotations, which the domain pathologists annotated. After reading Barrett's oesophagus literature, we understand the oesophagus's different layers and which layer contributes to diagnosing Barrett's oesophagus and the form of the dysplastic changes in the cells. We also clearly recognise the cells and how healthy oesophagus tissue looks histologically. Thus, as the pathologists did not provide annotations for unimportant regions, we annotated regions with healthy oesophagus tissue, noise and layers of the oesophagus that were not affected by the dysplastic changes, such as any layer deeper than lamina propria. Figure 4.4 shows samples of unimportant regions.



**Figure 4.4** Samples of the unimportant regions: (a) healthy oesophagus biopsy, (b) healthy epithelium, (c) Unknown but was never annotated by the pathologists, (d) muscular mucosa, and (e) wax

We annotated unimportant regions from 23 training WSIs; 20 belong to NFD, which has tissues for a healthy oesophagus, one LGD slide, and two HGD slides. Following the extraction method discussed in section 3.7, we sampled, at 10X magnification, 511,087 unimportant patches from 20 slides for training the model and 122,326 patches from three separate slides as a validation set. Also, the sampled patches at 10X magnification from the provided annotations were used as the region of interest class. More details about the data of that class are explained in section 3.7 and are provided in Table 3.5. The second dataset classifies dysplasia based on the extracted features at 10X magnification. The used dataset for this purpose is the same set that belongs to the region of interest class in the first model.

On the other hand, the second phase used the generated heatmaps from the first phase as a dataset to train the annotation classifier alongside the annotation masks for the epithelial layer. The researcher manually annotated the epithelial layer masks by roughly drawing borders that segment the epithelium lining apart from the lamina propria layer. Each epithelium annotation was conducted at a level similar to the level of provided annotation. Figure 4.5 shows samples of the epithelial layer masks for different grades.



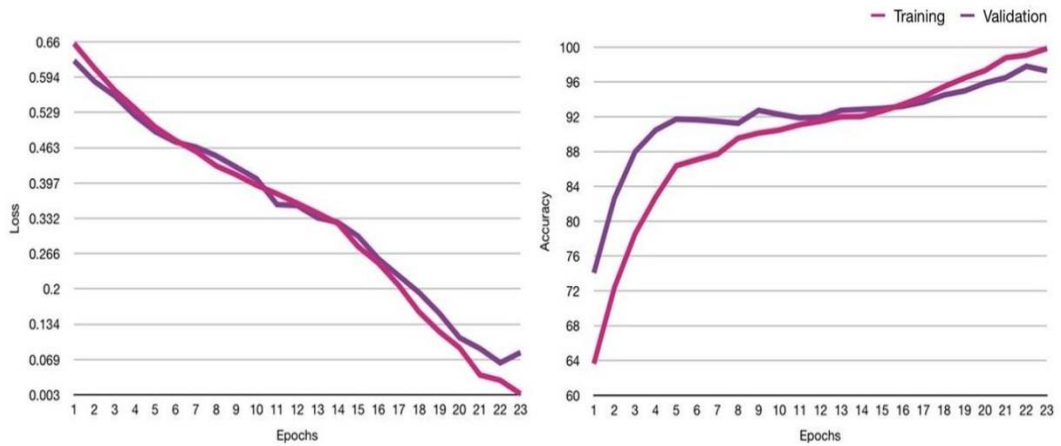
**Figure 4.5** Samples of the epithelial layer mask dataset

## 4.6 Experiments and results

The performance of the proposed system for detecting regions of interest and classifying dysplasia in annotations based on the analysis at 10X magnification is discussed in this section. Additionally, an individual performance measure for each component of the proposed system in this chapter is presented.

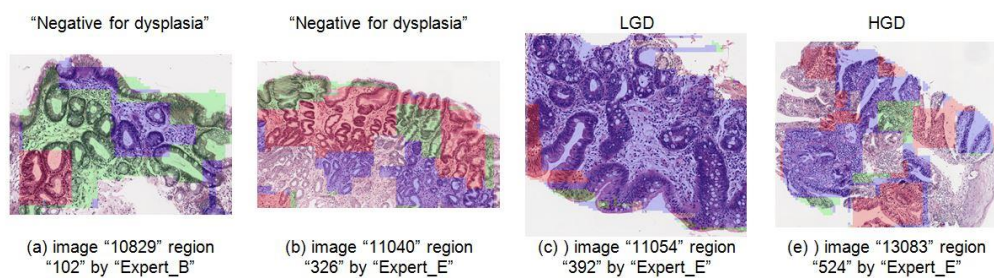
### 4.6.1 Regions of interest detection

872,326 were used for training regions of the interest detector network, and 213,305 extracted patches were used to validate the network. The model was trained to decrease both the training and validation losses. Once the training loss drops and the validation loss increases, that indicates model overfitting. At this point, the training was stopped, and the model was restored from the previous point. For this model, the training was stopped at epoch 22 (see Figure 4.6), where the model achieved the best validation loss at 0.0625 with 96.5% accuracy. The line charts for the training and the validation losses and accuracies progressions are shown in Figure 4.6.



**Figure 4.6** The losses and accuracies evolutions for the training (in pink) and the validation (in purple) sets during training regions of interest subnetwork

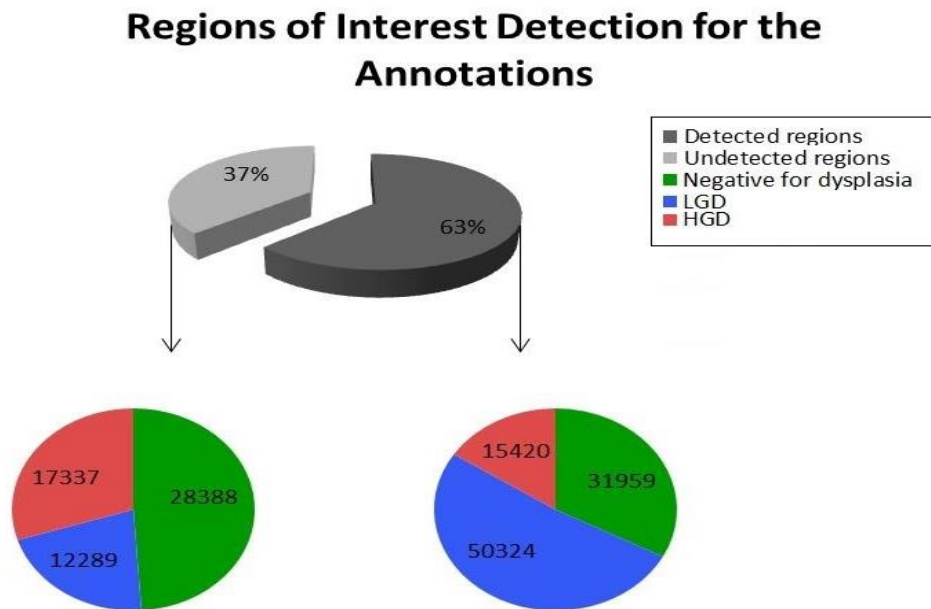
For evaluating the model's performance, it was tested on 155,717 extracted regions of interest patches from the 52 annotations of the test whole slides. The model successfully marked 97,703 as regions of interest and detected 40 annotations. Most of the detected regions belong to LGD and HGD annotations. Based on the provided test annotation, the model detected 18 out of 23 NFD annotations, 9 out of 10 LGD annotations, and 13 out of 19 HGD annotations. Figure 4.7 shows samples from the regions of interest detection results on different dysplasia grades.



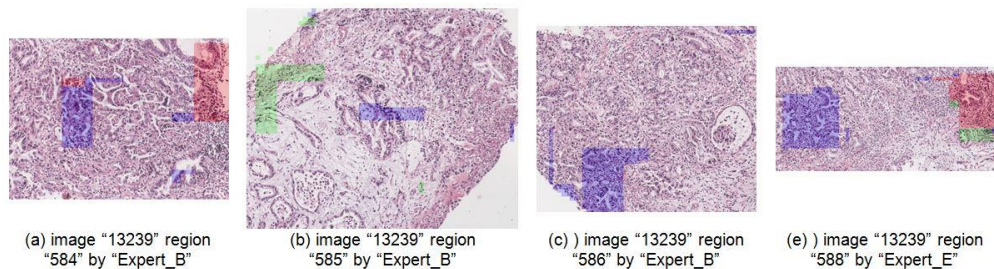
**Figure 4.7** Examples of the regions of interest detection results for some of the provided test annotations

The highlighted regions are the detected regions from the annotations. The prediction (green, blue and red represent NFD, LGD and HGD, respectively) is based on the network results in section 4.6.2. The label above each annotation represents the ground-truth label for that annotation.

Figure 4.8 shows pie charts with the patches number of the detected regions against the undetected regions for each grade. The LGD annotations had most of their regions detected, while at least 50% of the annotations of every HGD WSI were detected. Only one IMC WSI, the only IMC slide in the test set (see Figure 4.9), had 11% of its annotations detected from the tested slides. The evaluation of the region of interest detection model is based on the fact that every patch in the annotation is important, and it relies on the patch-level evaluation. An evaluation of the effect of adding the region of interest detector to Barrett's oesophagus related dysplasia grading system is discussed later in section 4.7.



**Figure 4.8** Pie charts show the percentage of detected regions within the provided test annotations and the number of detected patches within each grade



**Figure 4.9** The results of the region of interest detection model for IMC annotations from an IMC slide



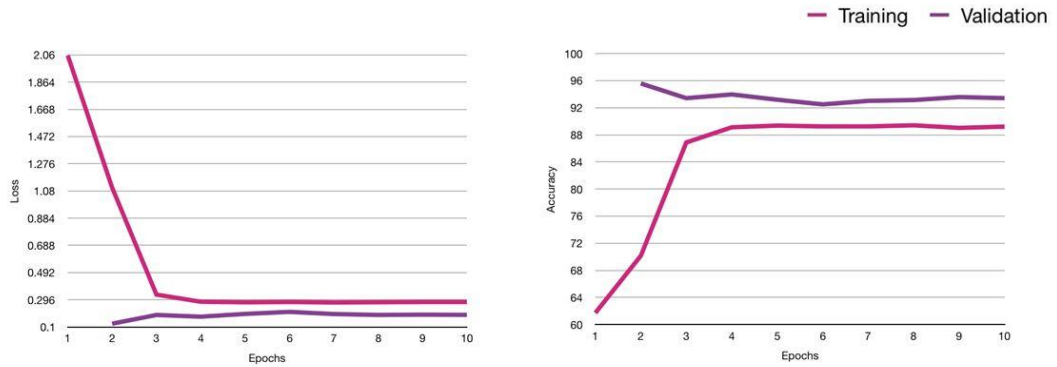
The highlighted regions are the detected regions from the annotations, and the prediction (green, blue and red represent NFD, LGD and HGD, respectively) is based on the results of the network in section 4.6.2

#### **4.6.2 Dysplastic high-level analysis and annotation grading**

A training set, an independent validation set and an independent test set, which were sampled from 112 training, 16 validation and 15 test WSIs, were used for training the model, optimising the model's parameters and selecting the best performing model and evaluating the final model's performance, respectively. Each validation and test set comprises 11% of the provided WSI. In addition, the extracted patches from the validation set comprise 20% of the total extracted patches from the training and validation sets. Each class of dysplasia in the validation set was represented by at least 10% of the total extracted patches from the training and validation sets of the corresponding class.

Like the first experiment, the model was trained while monitoring the loss of the validation set after each epoch (at the end of the training, the entire patches within the training set). The model was trained for ten epochs, and both the training and validation losses converged after the seventh epoch. The best validation and training losses pair was achieved at the tenth epoch with 0.187 error and 93.42% validation accuracy. It is important to monitor both training losses and the validation sets to prevent both from underfitting, where the validation loss is low. In contrast, the training loss is high, as observed at epoch 2, and overfitting, where the opposite case occurs. Figure 4.10 shows line charts for the progress of the training and validation losses and accuracies during the training of the feature extractor and classifier.





**Figure 4.10** The losses and accuracies evolutions for the training (in pink) and the validation (in purple) sets during the training of the subnetwork to classify Barrett's oesophagus related dysplasia based on the analysis at 10X magnification

At the first epoch, the Softmax layer was truncated, and all the layers of the model, except the Softmax layer, were frozen to train the new classifier. At this point, the model was not validated.

The test annotations were used to evaluate the model's performance at the level of patches based on the assumption that every patch within the annotated region has a label similar to the ground-truth label of the container annotation. Although relying on that assumption in evaluating the model may result in a biased evaluation, the patch-level assessment was performed to understand the general model better. An example of bias evaluation is provided in section 4.7 and Figure 4.14

From NFD annotations, 29,604 extracted patches were classified correctly, while 19,914 and 10,833 patches were misclassified as LGD and HGD, respectively. Also, from LGD and HGD annotations, 10,329 and 6,688 patches were misclassified as NFD, while 30,442 LGD and 15,839 HGD patches were classified correctly. From the LGD annotations, 21,837 patches were predicted as HGD and 10,231 patches and vice versa (the confusion matrix for patch-level is provided in Table 4.1). Based on the patch-level results, the performance measurements for the 10X magnification-based classification were calculated and summarised in Table 4.2. The model had a 0.49 score for each precision and recall, 0.48 F1-score, 0.74 specificity, and an overall 48% accuracy.

**Table 4.1** The confusion matrices for the model based on the analysis of the patches, annotations and slide levels at 10X magnification

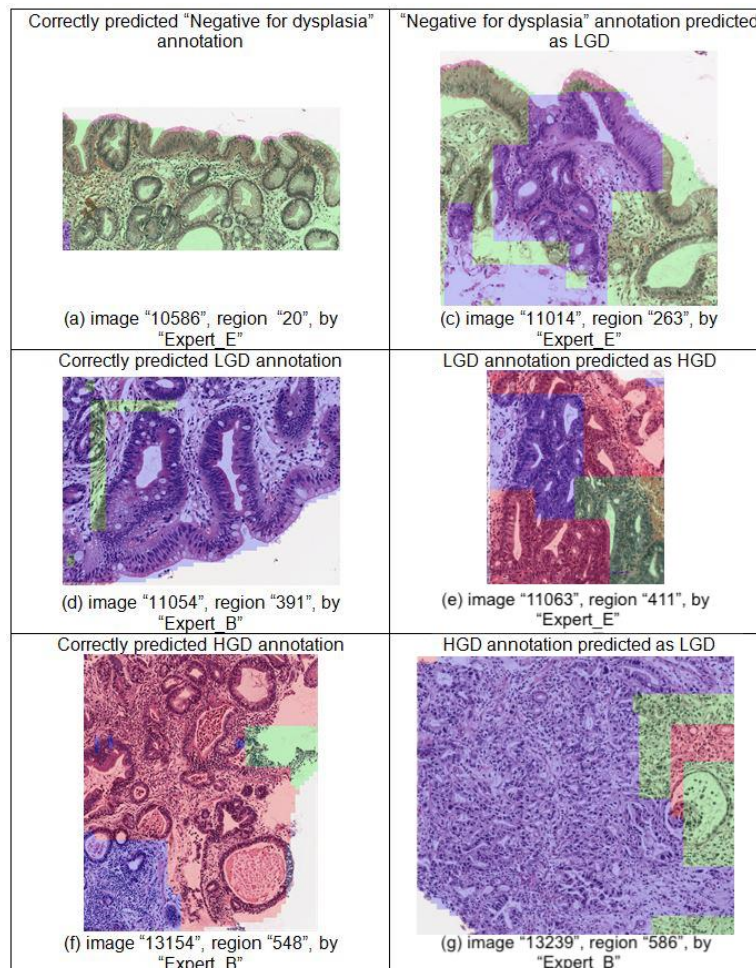
	Patch-wise			Annotation-wise			Slide-wise		
	NFD	LGD	HGD	NFD	LGD	HGD	NFD	LGD	HGD
NFD	29604	19914	10833	10	11	2	4	3	0
LGD	10329	30442	21837	3	5	2	0	1	2
HGD	6688	10231	15839	1	6	12	0	1	4

**Table 4.2** The performance measurements for the model based on the analysis of the patches, annotations and slide levels at 10X magnification (three-tier classification)

	NFD	LGD	HGD
<b>Patch-wise</b>			
<b>Precision</b>	0.63	0.50	0.33
<b>Recall</b>	0.49	0.49	0.48
<b>Specificity</b>	0.82	0.68	0.73
<b>F1-score</b>	0.55	0.49	0.39
<b>Accuracy</b>	69%	60%	68%
<b>Annotation-wise</b>			
<b>Precision</b>	0.71	0.23	0.75
<b>Recall</b>	0.43	0.50	0.63
<b>Specificity</b>	0.86	0.60	0.88
<b>F1-score</b>	0.54	0.31	0.69
<b>Accuracy</b>	67%	58%	79%
<b>Slide-wise</b>			
<b>Precision</b>	1.00	0.2	0.67
<b>Recall</b>	0.57	0.33	0.80
<b>Specificity</b>	1.00	0.67	0.80
<b>F1-score</b>	0.73	0.25	0.73
<b>Accuracy</b>	80%	60%	80%

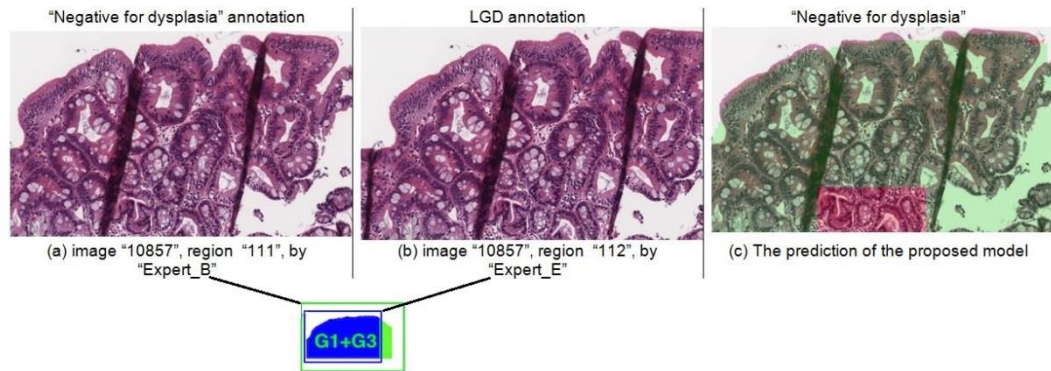
Moreover, the performance was measured based on the annotations and slides discussed in sections 4.1, 4.2 and 4.3.2. By feeding the test set into the annotation inference module, 10, 5 and 12 annotations were correctly classified as NFD, LGD and HGD, respectively. Figure 4.11 shows examples of correctly classified and misclassified annotations from each grade. Three

annotations from LGD each and one from HGD were misclassified as NFD. It is essential to highlight that one of the misclassified LGD annotations, NFD, has almost the same regions as another NFD annotation. Still, two experts diagnosed them differently (see Figure 4.12). Eleven and two NFD annotations were upgraded incorrectly to LGD and HGD, respectively, and two LGD annotations were upgraded to HGD. In comparison, six HGD annotations were downgraded to one level by the model. Based on the annotation-level predictions, the model's performance was enhanced compared to the patch level. The model scored an overall 0.56 precision, 0.52 recall, 0.78 specificity, and 0.51 F1-score, respectively, and the overall accuracy was decreased to 52%.



**Figure 4.11** Examples of the correctly classified regions and the misclassified regions by the analysis of the annotations at 10X magnification from each grade

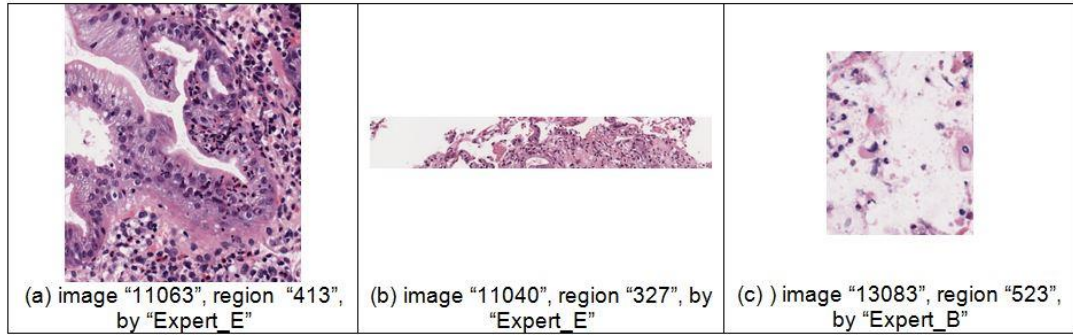
The green, blue and red stained regions represent NFD, LGD and HGD predictions. In the results visualisation, every square patch is stained using the previously mentioned colour code based on its predicted label. Also, the sampling process at 10X magnification involves overlapped patches, which explains the overlapped staining.



**Figure 4.12** Example of a region labelled differently by two pathologists

(a) and (b) are annotations provided by the pathologists, and (c) is the prediction mask generated by the proposed model, which agrees with "Expert\_B".

Secondly, as mentioned in section 4.3.2, the whole virtual slide follows the highest grade of any annotations as long as the annotation size is reasonably acceptable (see Figure 4.13 for small annotation samples). For instance, the region "413" (see Figure 4.13 (a)) has 274X233 pixels at 10X magnification, which is less than the size of any sampled patch; thus, the grade of this region should not influence the grade of the whole virtual slide. As a result, nine whole virtual slide images were correctly diagnosed, whereas six were misdiagnosed based on the analysis of the annotations at 10X magnification only. Three NFD slides were diagnosed as LGD. Two LGD slides were diagnosed with HGD. Finally, only one HGD was downgraded to LGD. The model achieved 0.62 precision, 0.57 recall, 0.82 specificity, 0.57 F1-score, and 60% accuracy on the whole slide level. Following the two-tier classification (dysplasia vs non-dysplasia), that model scored 0.86 precision, 0.79 for recall and specificity, 0.78 F1-score, and 78% accuracy. Table 4.2 summarises confusion matrices and all the calculated metrics for each grade (following the three-tier classification) at the patch, annotations, and slide levels.



**Figure 4.13** Examples for small annotations that their predictions do not influence the label of their container slides

### 4.6.3 The proposed model

The confusion matrices are provided in Table 4.3. The assembled network results at the level of annotations and slides as measured by precision, recall, specificity, F1-score and accuracy are provided in Table 4.4. Examples of the results are provided in Figure 4.7.

**Table 4.3** The confusion matrices for the proposed model at the annotation and slide levels (three-tier)

	Annotation-wise			Slide-wise		
	NFD	LGD	HGD	NFD	LGD	HGD
NFD	10	10	3	3	3	1
LGD	5	5	0	1	2	0
HGD	1	7	11	0	2	3

**Table 4.4** The performance measurements for the proposed model at the annotation and slide levels (three-tier)

	NFD	LGD	HGD
<b>Annotation-wise</b>			
<b>Precision</b>	0.63	0.23	0.79
<b>Recall</b>	0.43	0.50	0.58
<b>Specificity</b>	0.79	0.60	0.91
<b>F1-score</b>	0.51	0.31	0.67
<b>Accuracy</b>	63%	58%	79%
<b>Slide-wise</b>			
<b>Precision</b>	0.75	0.29	0.75
<b>Recall</b>	0.43	0.67	0.60
<b>Specificity</b>	0.88	0.58	0.90
<b>F1-score</b>	0.55	0.40	0.67
<b>Accuracy</b>	67%	60%	80%

## 4.7 Discussion

As discussed earlier, the proposed model is composed of two networks. One part is to detect the critical regions in 10X magnification images to be analysed by the second part. The region of interest detection model was evaluated based on the number of detected annotations only because the test annotations have only regions of interest and the examples of healthy tissue were unavailable; thus, the model was penalised if it left a patch within the annotation not detected. The region of interest detection model had the best performance with the LGD annotations and a moderate performance with NFD annotations. The undetected NFD annotations were not concerning, as detecting a non-dysplastic tissue is not essential in grading dysplasia. The absence of dysplasia in tissue leads to NFD grade.

The lowest performance was achieved for HGD annotations, more precisely, IMC, when the model was tested on the test set. The two pathologists diagnosed all the undetectable HGD regions as IMC, and those annotations affected the performance as only 11% of their patches were detected. The model did not perform well in detecting IMC, attributed to two reasons. The first reason is that IMC is the most challenging grade for the models to recognise and predict in this thesis. It could be due to the underrepresentation of that class in the training set. The second reason is

the nature of the dysplastic abnormalities in this grade, as they differ from the other lower grades (including “high-grade dysplasia” following the Vienna classification). In this case, the changes interfere with the mucosal layer, and the epithelial layer is no longer distinguishable (refer to Figure 4.9 for IMC annotations).

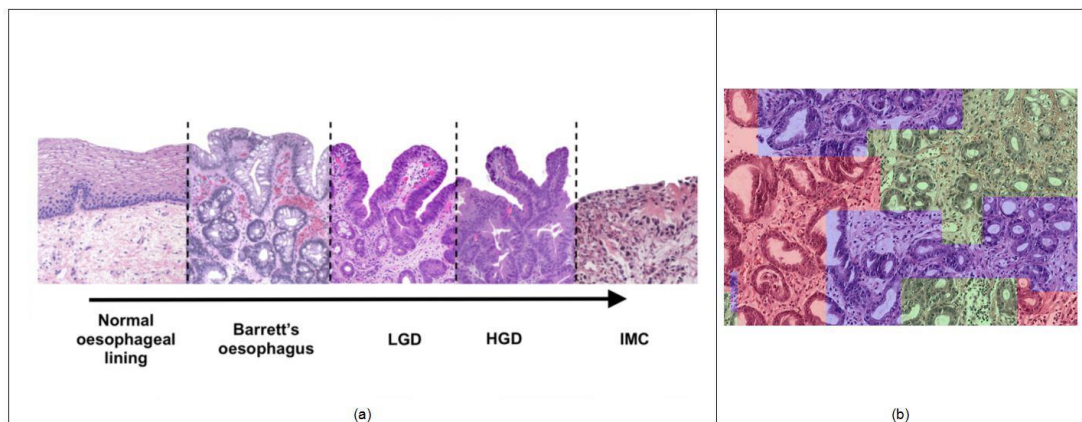
It is crucial to highlight that for the dysplastic high-level analysis model, the proposed model and the benchmark models, which will be discussed later in this section, all the provided annotations by the domain experts were used in training the models. Moreover, in evaluating the performances of those models, all the test annotations by both experts were used for assessment. However, at the test time, the performance of a model was measured by inferring the grade of the slides regarding the agreement with either of the experts, except when it is stated that the agreement with "Expert\_E" or "Expert\_B".

In addition, the performances of the presented models in this thesis were compared against the pathologists' diagnoses based on the virtual slides only, despite disagreements in the diagnosis of three test slides. On the one hand, the model's predictions of the provided annotations were considered correct if they matched the diagnosis of the expert who annotated the region. On the other hand, the slides' inferences were considered correct if they matched either expert. That unusual slide-level performance evaluation is introduced to the process of diagnosing the grade of dysplasia in Barrett's oesophagus by the fact that neither of the experts is absolutely correct or incorrect due to the absence of guidelines that draw well-defined boundaries between two grades of dysplasia and that increases the interobserver and intraobserver agreements. In this thesis, the model's predictions were considered expert opinions, and interobserver disagreements are accepted as long as they fall in the range of disagreement with real pathologists. Moreover, Cohen's Kappa was used to calculate the agreement between the models and the experts each time we evaluated the slide-level prediction of the models.

For the dysplastic high-level analysis model, the performance was not measured based on the patch level because the patches were assigned to labels based on the earlier mentioned assumption. Suppose the model



correctly marks patches in NFD annotation as LGD or/and HGD. In that case, the model is penalised for assigning labels to them conflicting with the ground-truth label for the annotation. For instance, in image “11040”, the annotated region “324” is assigned to NFD grade (see Figure 4.14), and every patch within the annotation was labelled as NFD. The learned model highlighted regions with blue and red (LGD and HGD, respectively) instead of the expected green highlight (NFD). Referring to Figure 4.14 (a), which illustrates the architectural changes along the continuous spectrum of the dysplastic changes, the glands arrangement for the highlighted red regions in Figure 4.14(b) match the form of a high level of dysplastic glands in Figure 4.14 (a) and similar to that the blue highlighted regions. However, in pathology, the highlighted red and blue regions should not influence the annotation grade as long as the dysplastic abnormalities did not occur in the epithelial layer. That explains the reason behind labelling the annotation with NFD. In general, the model had better performance at the slide-level than the performance at the patch level, which is attributed to the previously explained reason.



**Figure 4.14** Comparison between the grade of the glands arrangements in annotation (b) and the precancerous progression form in (a)

(a) is a figure that describes the precancerous progression in Barrett's oesophagus, (b) region “324” from image “11040” labelled as NFD.

After analysing the abnormalities at 10X magnification on the whole patches within the annotations without identifying the region of interest, it resulted in overall 0.52 sensitivity, 0.78 specificity, 0.51 F1-score, and 52% accuracy at the annotation-level with a “moderate” agreement with the domain experts (0.416 weighted kappa). Moreover, it achieved 0.57 sensitivity, 0.82



specificity, 0.57 F1-score, and 60% accuracy with 0.579 weighted Kappa (moderate agreement) at the slide-level.

Following the two-tier classification (see Table 4.5 for the confusion matrices and Table 4.6 for the performance results), the classifiers work better in discriminating features of dysplasia against NFD based on the classified patches, annotations and slides.

**Table 4.5** The confusion matrices for the model based on the analysis of the patch, annotation and slide levels at 10X magnification (two-tier classification)

	Patch-wise		Annotation-wise		Slide-wise	
	NFD	Dysplasia	NFD	Dysplasia	NFD	Dysplasia
NFD	29604	30747	10	13	4	3
Dysplasia	17017	78349	4	25	0	8

**Table 4.6** The performance measurements for the model based on the analysis of the patch, annotation and slide levels at 10X magnification (two-tier classification)

	Patch-level		Annotation-level		Slide-level	
	NFD	Dysplasia	NFD	Dysplasia	NFD	Dysplasia
Precision	0.63	0.71	0.71	0.66	1.00	0.73
Recall	0.49	0.82	0.43	0.86	0.57	1.00
F1-score	0.55	0.79	0.54	0.75	0.73	0.84
Accuracy	69%		67%		80%	

By combining the two networks and running them sequentially, the performance of the combined model was slightly diminished at both the annotation-level and the slide-level, as its performance scores are: 0.50 sensitivity, 0.77 specificity, 0.50 F1-score, 50% accuracy and 0.37 weighted kappa, which is in the range of fair agreement between the proposed model and the pathologists, at the annotation-level. Also, it had 0.57 sensitivity, 0.79 specificity, 0.54 F1-score, and 53% accuracy at the slide-level and did not change the interobserver agreement. The new addition to the high-level analysis based classifier enhanced the prediction of LGD annotations. Table 4.7 compares the performance metrics for the high-level analysis

based classifier before and after adding the region of interest detection module.

**Table 4.7** Performance measurements for the high-level analysis and classification solely against it coupled with regions of the interest detection model

	Analysis and classification at 10X magnification	
	NFD	The proposed model Dysplasia
	<b>Annotations-wise</b>	
<b>Precision</b>	0.56	0.55
<b>Recall</b>	0.52	0.50
<b>Specificity</b>	0.78	0.77
<b>F1-score</b>	0.51	0.50
<b>Accuracy</b>	52%	50%
<b>Agreement with “Expert_B”</b>	0.467	0.472
<b>Agreement with “Expert_E”</b>	0.156	0.089
	<b>Slide-wise</b>	
<b>Precision</b>	0.62	0.60
<b>Recall</b>	0.57	0.57
<b>Specificity</b>	0.82	0.79
<b>F1-score</b>	0.57	0.54
<b>Accuracy</b>	60%	53%
<b>Agreement with “Expert_B”</b>	0.291	0.416
<b>Agreement with “Expert_E”</b>	0.135	0.057
<b>“Expert_B” and “Expert_E” agreement</b>		0.674

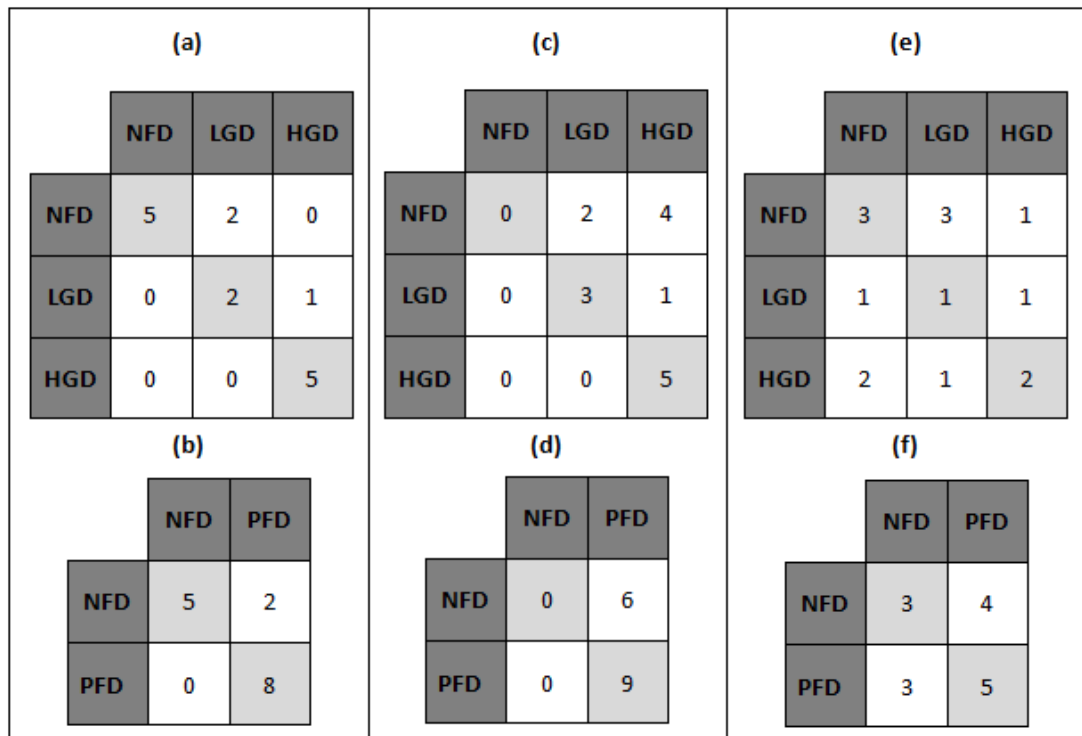
For evaluating the proposed model in this chapter, the model's performance was compared with two benchmarks. The comparison was carried out based on the three-tier and two-tier classifications against two benchmarks. The same train and test sets were used in training and evaluating the models. Table 4.8 compares the proposed model performance with the two related works assessed based on precision, recall, specificity, F1-score, and accuracy with 95% confidence intervals. Also, the results were rounded up to two decimals. It is important to point out that the strength of evidence for the comparison between the studies is weak due to the limitation of the size of the dataset; however, to control it, the performances of the compared models were fit to the same test data. The first benchmark is research that

was conducted by Adam (2015). It adopted a machine-learning approach to analyse and classify dysplasia (see section 2.2) using Barrett’s oesophagus histological images at a 10X magnification dataset discussed in Chapter 3. Her performance was evaluated by her using 15 independent WSIs (the test set) following the three-tier classification (NFD vs LGD vs HGD). In addition, the performance based on the two-tier classification (NFD and PFD) was calculated in this chapter. Based on the three-tier and two-tier classifications, the first benchmark had confusion matrices as provided in Figure 4.15 (a) and (b), respectively, using her suggested consensus grading system for grading slides. The suggested consensus grading system grades a slide as HGD if the number of HGD annotations represents 40% or more from the detected regions in the slide and as LGD if the LGD annotations represent 32% or more. To have a relevant comparison between the proposed model and hers, the suggested consensus grading system in this chapter was applied, where the slide follows the highest graded annotation within the slide. As a result, her model had confusion matrices, as shown in Figure 4.15 (c) and (d), following the three-tier and two-tier classification, respectively. The proposed model outperformed the benchmark model in both classification categories, with 0.54 (three-tier) and 0.64 (two-tier) F1-score against 0.22 and 0.38 recall. Also, the proposed model achieved a moderate agreement with the domain experts with 0.412 weighted KV and 0.587 KV, whereas the benchmark model achieved a fair agreement with 0.217 weighted KV and 0.250 KV. The benchmark model failed to predict NFD slides while showing high performance in recognising dysplasia.

**Table 4.8** The three-tier and two-tier classification for the proposed model against other works at the slide-level with 95% confidence intervals

	Three-tier classification			Two-tier classification		
	The proposed model	(Adam, 2015)	(Tomita et al., 2019)	The proposed model	(Adam, 2015)	(Tomita et al., 2019)
<b>Precision</b>	0.60 (+/-0.27)	0.37 (+/-0.27)	0.40 (+/-0.27)	0.69 (+/-0.26)	0.30 (+/-0.25)	0.53 (+/-0.28)
<b>Recall</b>	0.57 (+/-0.27)	0.58 (+/-0.27)	0.39 (+/-0.27)	0.65 (+/-0.26)	0.50 (+/-0.28)	0.53 (+/-0.28)
<b>Specificity</b>	0.79 (+/-0.22)	0.77 (+/-0.23)	0.70 (+/-0.25)	0.65 (+/-0.26)	0.50 (+/-0.28)	0.53 (+/-0.28)

<b>F1-score</b>	0.54 (+/-0.28)	0.22 (+/-0.23)	0.30 (+/-0.25)	0.64 (+/-0.27)	0.38 (+/-0.27)	0.52 (+/-0.28)
<b>Accuracy</b>	53% (+/-28)	53% (+/-28)	40% (+/-27)	67% (+/-26)	60% (+/-27)	53% (+/-28)
<b>Agreement with "Expert_B"</b>	0.416	0.250	0.088	0.587	0	0.054
<b>Agreement with "Expert_E"</b>	0.057	0.273	0.184	0.034	0	0.250



**Figure 4.15** The slide-level confusion matrices for the first and second benchmarks follow the three-tier and two-tier classifications

(a) The confusion matrix for the first benchmark follows the three-tier classification and the suggested consensus grading system by Adam (2015), (b) the confusion matrix for the first benchmark follows the two-tier classification and the suggested consensus grading system by Adam (2015), (c) the confusion matrix for the first benchmark follows the three-tier classification and this chapter's suggested consensus grading system, (d) the confusion matrix for the first benchmark follows the two-tier classification and this chapter's suggested consensus grading system, (e) the confusion matrix for the second benchmark follows the three-tier classification, and (f) the confusion matrix for the second benchmark follows the two-tier classification.

The other benchmark model is research that introduced a novel deep-learning weakly-supervised approach by Tomita et al. (2019). The approach is an attention-based model with two parts: a feature extractor part and an attention network to classify the extracted feature (see section 2.2). It was conducted using 44 WSIs for gastroesophageal junction mucosal biopsies. Their study was based on a four-tier classification (healthy vs NFD vs dysplasia vs “adenocarcinoma”). To regenerate the results using the dataset used in this thesis with their approach, we reimplemented their model except that their network in the feature extraction phase (ResNet-18) was substituted with “Inception-ResNet-v2” to unify the quality of the extracted features by the two compared models. The second benchmark had confusion matrices, as shown in Figure 4.15 (e) and (f), respectively, using both classification categories. The proposed model outperforms the performance of the deep-learning-based benchmark considering the precision, recall, F1-score and accuracy results. Also, the benchmark model slightly agrees with the experts with 0.126 weighted KV. However, it has a slightly higher agreement with “Expert\_E” than the proposed model, which has a poor agreement with the expert. Refer to the first and the third columns in the last three columns in Table 4.8.

In brief, both deep-learning-based approaches (the proposed model and the deep-learning-based benchmark) showed better performances than the conventional machine-learning-based approach. That result was expected because deep-learning approaches have outcompeted traditional machine learning when applied in different fields, including histopathology. Moreover, the proposed weakly supervised deep-learning model surpassed the weakly-supervised work introduced by Tomita et al. (2019) in analysing and classifying dysplastic features at 10X magnification. The previous result suggests that the introduced annotation inference system in this chapter performs better than the attention network to decide the grade of the bag of instances. Although the proposed model that analyses virtual slides at 10X magnification performs better than the two benchmark models, the analysis at that magnification did not provide a useful CAD system.

According to Treanor et al. (2009), the pathologic diagnosis of Barrett’s oesophagus related dysplasia suffers from an interobserver agreement that does not exceed moderate agreement amongst expert gastrointestinal tract pathologists. This section discusses the agreement between the learned

models in this chapter (dysplastic analysis and classification at 10X magnification model with and without regions of interest detection module) and the two domain experts “Expert\_B” and “Expert\_E”, who were senior pathologists specialising in gastrointestinal tract pathology at the time of the annotation process. Besides, this section compares those agreements with the reported agreements in (Treanor et al., 2009) between the experts and two trainees, “Trainee\_D” and “Trainee\_G”, who were, at the publishing time, three years experienced pathologists. The agreement levels are detailed in Table 4.9. For the test set at the level of WSIs diagnosis, the agreement between the experts is substantial (0.674 KV). At the same time, the proposed model manages to have a moderate agreement with “Expert\_B” and a slight agreement with “Expert\_E” at the slide-level and the annotations-level. In summary, the models always agree more with “Expert\_B”, and by employing the region of interest detection model, the agreement is enhanced.

**Table 4.9** The proposed model slide-level and annotation-level agreements with experts

		“Expert_B”	“Expert_E”
Slide-level	Analysis and classification at 10X magnification	0.291 (fair agreement)	0.135 (slight agreement)
	The proposed model	0.416 (moderate agreement)	0.057 (slight agreement)
	“Expert_B” / “Expert_E”	0.674 (substantial agreement)	
Annotation-level	Analysis and classification at 10X magnification	0.467 (moderate agreement)	0.156 (slight agreement)
	The proposed model	0.472 (moderate agreement)	0.089 (slight agreement)
	“Trainee_D” (Treanor et al., 2009)	0.17 (slight agreement)	0.27 (fair agreement)
	“Trainee_G” (Treanor et al., 2009)	0.29 (fair agreement)	0.46 (moderate agreement)

At the annotation-level, the proposed model was compared against the trainee pathologists who scored slight and fair agreements with “Expert\_B” and fair and moderate agreement with “Expert\_E” for grading 46 biopsies following the Vienna classification. At the same time, the proposed model scored the highest agreement with “Expert\_B” and the lowest agreement

with the other expert. The models scored the lowest agreements with the experts at the slide-level compared to the experts' agreement.

## 4.8 Conclusion

The pathologic diagnosis of Barrett's oesophagus related dysplasia relies on a combination of cytological and architectural abnormalities in biopsies. This chapter presented a study that investigated the analysis of histological WSIs at 10X magnification using a deep-learning approach and compared its performance against a conventional machine-learning approach. It demonstrated a novel weakly supervised deep-learning model with a fork style to detect regions of interest, which are patches within an annotation containing abnormal features, and classify them into one of the three grades of dysplasia sequentially. Then, it infers the grade of an annotation by classifying its corresponding feature, which represents the percentage of pixels that belong to each grade in each layer, using a random forest classifier. The proposed model follows a transfer deep-learning-based approach to train the two subnetworks using two separate datasets. Also, it follows the MIL approach as a form of weakly supervised learning to train the subnetwork that classifies the detected regions. Weakly supervised learning is considered a solution to the coarsely annotated WSIs. During training the MIL approach, each patch in an annotation test has the same label of the annotation to fine-tune a pre-trained "Inception-ResNet-v2" on the public dataset "ImageNet". After training the two subnetworks, they were connected by sharing all the frozen layers and having their individual finetuned layers.

A pre-trained "Inception-ResNet-v2" on the public dataset "ImageNet" was used in the transfer deep-learning training. The model was fine-tuned using 256X256 pixels images, sampled at 10X magnification from the provided 338 annotations. The model was tested on 52 annotations from the WSIs from a separate test set, and each expert annotated around 26 annotations. After testing the test annotations, prediction masks were generated. They were fed into the trained random forest classifier to determine the grade for the test annotations. The model achieved 52% and 60% accuracy at the annotation-level and the whole slide level, respectively.

The statistics presented in Table 4.2 suggest that the classification based on high-level analysis succeeds more in diagnosing HGD. At the same time, it upgrades the diagnosis in most cases of lower grades. The higher predictions for the annotations are attributed to the nature of grading dysplasia in Barrett's oesophagus. As in the lower grades of dysplasia, the glands might have abnormal atypia, while the top of crypts should not be affected and preserve their normal structure. Therefore, an additional model for detecting the region of interest is added to the model as a booster to filter unnecessary patches that might affect the grading. The new model combined the region of interest detector with the 10X magnification analysis based classifier to share the non-trainable layers and have their own fine-tuned parameters to decrease the computational cost. The latest addition to the model increased the model's performance in grading LGD. However, the model's performance decreased by 2% and 7% accuracies at the annotation and slide levels. The proposed model's diagnosis agrees moderately with the senior gastrointestinal tract pathologists.

Moreover, a comparative study on a weakly supervised deep-learning approach introduced by Tomita et al. (2019) to predict the grade of dysplasia in Barrett's oesophagus was conducted. Based on the provided small-size dataset, the proposed model scored better than their model.". Also, a comparison between the proposed model and (Adam, 2015), as the comparison focuses on comparing the performance of a deep-learning approach against a machine learning approach, shows that at a high-level analysis, the deep-learning approach yields better performance. Analysing the WSIs at low magnifications using the proposed approach saves time and computational cost and produces good results compared to the conventional machine-learning benchmark. However, the results could be enhanced by an approach that analyses the WSIs at higher magnifications.



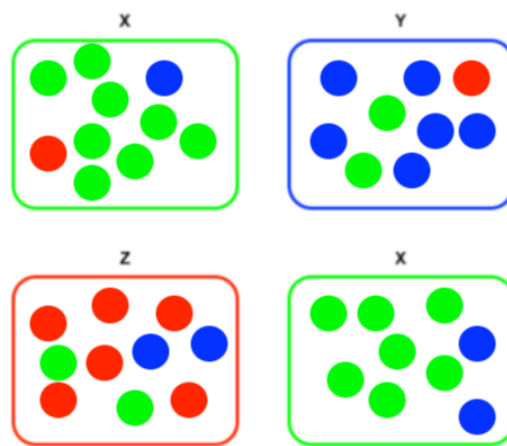
## Chapter 5. Histopathology Low-level Analysis and Classification

### 5.1 Introduction

The previous chapter discussed detecting regions of interest and analysing and classifying dysplasia in Barrett's oesophagus histological images at a high level, which is the analysis at 10X magnification. Conversely, this chapter will handle them at the lowest available level by detecting the potential dysplastic tissue and analysing the slides at 40X magnification. On the one hand, the region of interest detector in the previous chapter is a binary classification task that discriminates between Barrett's oesophagus tissue (dysplastic or non-dysplastic) and healthy oesophagus tissue. On the other hand, the potential dysplastic tissue detector is a deep-learning one-class anomaly detector trained on the NFD annotations only to detect the anomaly patches, specifically dysplastic tissues.

In the field of histological images, most of the studies rely on supervised and unsupervised deep-learning approaches to develop CAD systems. In contrast, the weakly supervised approach is rarely investigated. This chapter will handle the problem introduced by the annotation technique for Barrett's oesophagus dataset, as discussed earlier in Chapter 3, using a weakly supervised approach and a novelty detection approach. The annotation issue is introduced when pathologists assign labels to annotations that grade the comprehensive annotations. However, each tissue patch within each annotation does not necessarily have the same label as the container annotation. Especially when the annotation is retrieved with a high magnification such as 40X, which pathologists rarely is used in annotating the regions, the existence of non-dysplastic tissues within the dysplastic annotations has a high probability. For Barrett's oesophagus dataset, it is assumed that the NFD annotations are clear from abnormal tissues because the grade NFD is applied when atypia does not exist in the annotation. However, LGD and HGD annotations might contain a mixture of NFD and dysplasia patches. That challenge is MIL, as bags (annotations) of instances (patches) are labelled without patch-level labels. Figure 5.1 illustrates the related problem. One of the solutions that could be considered to handle the dataset's challenge is to work on the dataset cleaning and convert the learning process from weakly supervised to supervised. This solution was

applied by adopting a novelty detection deep-learning approach, a one-class classification, for filtering the non-dysplastic tissues contained in the dysplastic annotations and analysing the results of the approach. Moreover, it employs a supervised approach trained on the cleaned data, using the proposed novelty detector to eliminate the occurrence of non-dysplastic tissue from the dysplastic annotation for the classification task. Finally, it discusses the effect of combining the two approaches to boost the model's performance and solve the weakly supervised problem. An overview of the proposed model is discussed in section 5.2.

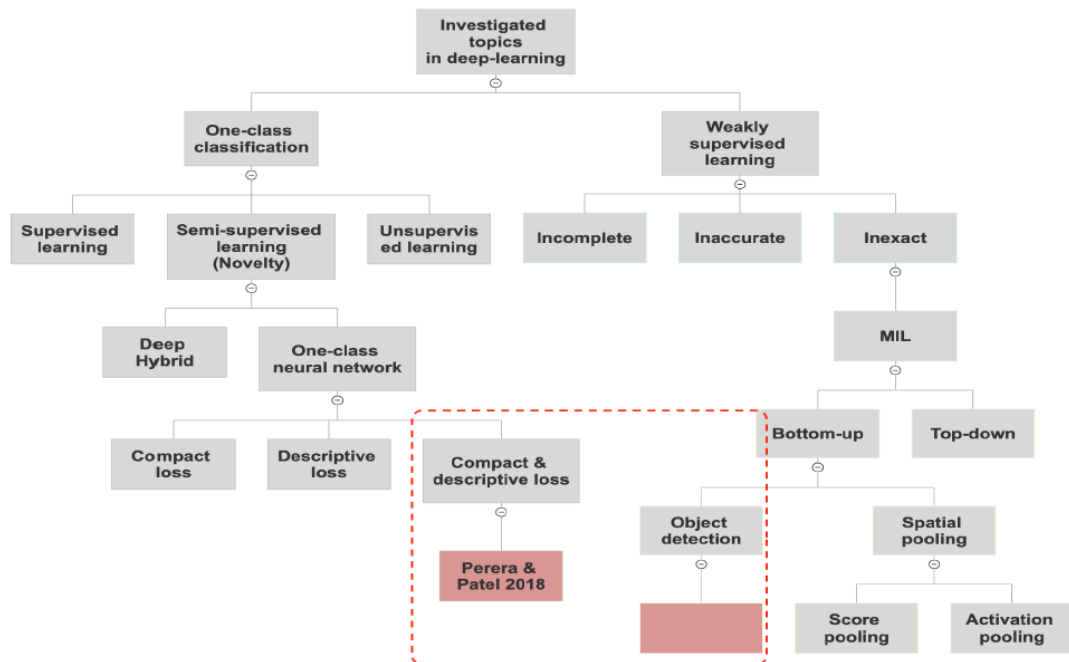


**Figure 5.1** Illustration for weakly supervised MIL problem

Scenario case for a MIL problem, where the rectangles are bags with known labels (X, Y and Z), and each bag has instances(circles) that belong to different classes (green, blue and red circles belong to classes X, Y and Z, respectively). In the standard MIL scenario, the negative bag is labelled negative when it is clear from positive instances. At the same time, the positive bag is labelled positive when one instance is positive in the bag. However, in many situations, especially in computer vision problems, it is difficult to ensure that the positive instances do not appear in the negative bags. Such as Barrett's oesophagus data issue, where pathologists provided labels for the annotations (coarse-grained labelling) without labels for the patches within the annotations (fine-grained labelling).

The novel research contribution in this chapter is investigating the bottom-up object detection approach, one of the MIL approaches, by employing a one-class neural network that is considered a semi-supervised one-class classification approach. When writing this thesis, that field has not been investigated before using a deep one-class classifier. Figure 5.2 shows the filled gap in the literature. The employed one-class neural network in this

chapter utilises a novel solution introduced by Perera and Patel (2019) to compute the compact and descriptive losses in learning the model. Their approach to learning their model was adopted to learn our model. A slight modification to their approach was made in the testing phase for the proposed model as the Local Outlier Factor (LOF) (Breunig et al., 2000) classifier was used instead of their nearest neighbour classifier for the reason that will be discussed in section 5.3.1.2.



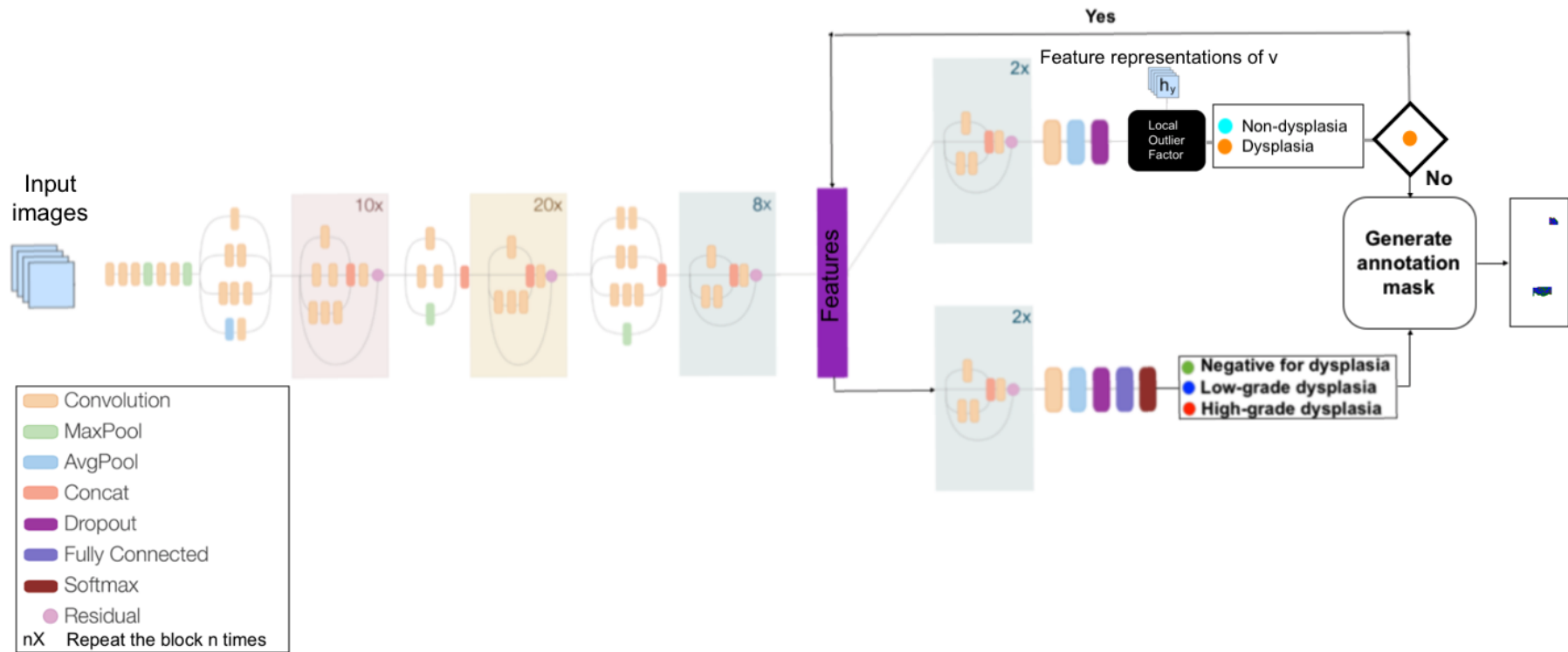
**Figure 5.2** A diagram for the unfilled gap in the literature

The red dotted line highlighted the unexplored field in the literature.

## 5.2 An overview

The proposed model comprises two subnetworks (refer to Figure 5.3). The first subnetwork is a one-class deep-learning model trained following the training technique of Perera and Patel (2019). The subnetwork is an “Inception-ResNet-v2” model trained twice using two different datasets, and two losses were computed. One dataset is the NFD patches, and the compact loss was computed. While the other is a problem-related dataset containing normal and tumour images, and the descriptiveness loss was calculated. The subnetwork parameters were updated using the total loss of the two losses. In the testing phase of the subnetwork, a LOF classifier was trained to detect dysplastic tissue. On top of the subnetwork's role in the

proposed model in detecting dysplastic tissues, it was employed to clean the training dataset to be used later in training the second subnetwork by detecting the dysplastic patches and eliminating the NFD patches within the dysplastic annotations. By removing the NFD patches from the PFD annotations, the second subnetwork was trained following the supervised learning approach to classify the detected patches into NFD, LGD or HGD and generates the annotations' heatmaps. Moreover, both subnetworks adopted transfer learning. In assembling the subnetworks, they were connected following the same assembling technique used in connecting the subnetworks in Chapter 4. A random forest classifier was trained for the annotation-level and slide-level grades inference, as is discussed in sections 4.1 and 4.3.2.



**Figure 5.3** The proposed model for dysplasia classification based on the analysis at 40X magnification

## 5.3 Methodology

The overall structure of the proposed model is illustrated in Figure 5.3. It is composed of two networks. The task of the first subnetwork (the top branch in Figure 5.3) is to detect potential dysplastic tissue from the sampled annotations using a deep one-class classifier, to be classified later by the second subnetwork (the bottom branch in Figure 5.3).

The first subnetwork is an existing novel work introduced by Perera and Patel (2019) for deep one-class classification. In this research, Their proposed model was adapted to solve the MIL problem in two ways:

- It was used as a data cleansing method. The need to implement a data cleanser was motivated by the need for fine-graining Barrett's oesophagus dataset to train a deep-learning network in a supervised manner to classify different grades of dysplasia. The network was employed to train the network in section 5.3.2 and build up the proposed CAD system in Chapter 6 (section 6.2.2).
- It was used as a performance booster during the testing phase by combining it with the dysplasia classification model, as is shown in Figure 5.3 and discussed in section 5.4.4.

The first subnetwork was trained in two stages. Once for feature extraction (see section 5.3.1.1 and section 5.4.1), and another for training an anomaly detector using the extracted feature by the first stage (see section 5.3.1.2 and section 5.4.2). Training the second subnetwork is discussed in sections 5.3.2 and section 5.4.3.

### 5.3.1 Potential dysplastic tissue detection

The used system of dysplasia classification in this research has one grade for non-dysplastic tissues and two others for PFD with two different degrees. When an annotation is graded as NFD by pathologists, the absence of abnormalities or the presence of some mild abnormalities that are not considered dysplasia is essential. However, in the case of positive dysplasia, abnormalities exist in some areas of the biopsy. As discussed in Chapter 3, the following annotation process was not precise, and the majority, mainly NFD cases, were extracted at 10X magnification. Only a few HGD regions

were annotated at 40X magnification. As a result, sampling patches from normal tissue within dysplastic regions for training the classifier might hamper the performance of the deep-learning neural network model. To combat this problem, the following hypothesis was formulated.

$H_0$ : the provided “positive for dysplasia” annotations only have dysplastic tissues.

$H_1$ : the provided “positive for dysplasia” annotations might have non-dysplastic tissues.

A deep novelty detection network was employed to filter the metaplasia tissues from dysplastic annotations to prove the hypothesis. The “Inception-ResNet-v2” model was fine-tuned using patches extracted from NFD annotations contained in NFD WSIs only. For this purpose, a one-class deep SVDD objective (Ruff et al., 2018) was utilised to minimise the loss. The loss for each mini-batch was set as the summation of the distance between every output feature representation of every input within that mini-batch from a centre, which was set as a random and fixed feature representation in the output space for the “Inception-ResNet-v2” network. The equation for the deep SVDD hard-boundary objective is provided in Equation 5.1, and it seeks a hypersphere with the minimum volume.

#### Equation 5.1

$$\text{Min}_W \frac{1}{n} \sum_{i=1}^n s(x_i)$$

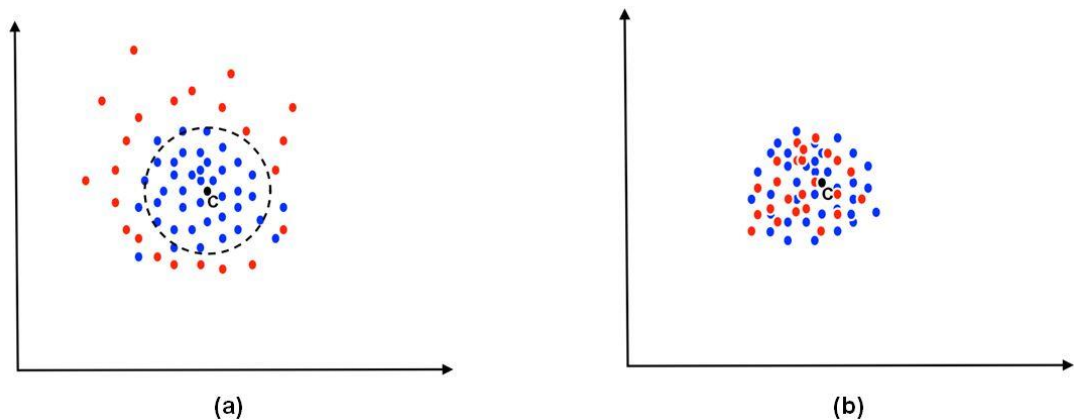
#### Equation 5.2

$$s(x_i) = \| f(x_i) - c \|^2, \text{ where } c \in \mathcal{F} \text{ and } x_i \in \mathcal{X}$$

Where  $W$  is the network parameters set,  $n$  is the number of inputs in the mini-batch,  $f(x_i)$  is the feature representation for the input  $x_i$  that belong to the input space  $\mathcal{X}$ ,  $\| \cdot \|^2$  is Euclidean distance, and  $c$  is the fixed centre that belongs to the output space  $\mathcal{F}$ . The network was trained until its parameters converged. Then, to use the trained model for the novelty detection task, a corresponding score  $s(x_i)$  (see Equation 5.2), The distance of the feature representation of the input from the centre of the hypersphere was associated with each instance in the one-class training set and a multiple-class test set. A threshold  $\delta$  was determined based on the 95%-quantile of

the training set score distribution to flag the data points that significantly vary from the other observed data points (outliers).

The model failed to flag outliers after testing the model on both dysplastic and non-dysplastic tissues. Most of the feature representations for the input from both groups were assigned to scores around one number. These results suggest that the model could not separate the feature space of dysplastic tissue and non-dysplastic tissue; thus, the learned representations of both groups were compact in the output space. An optimal hypersphere, such as Figure 5.4 (a), was searched; however, the hypersphere we found was like Figure 5.4 (b).



**Figure 5.4** Feature space obtained using deep SVDD features: (a) expected, (b) real

Blue dots represent normal instances (non-dysplastic patches), and red dots are abnormal instances (dysplastic patches).

A possible explanation for that result was predicted by Perera and Patel (2019), which is that a trivial solution was learned by the model, as the discrimination feature of the model was neglected by training the model on one-class only and not penalising the model for miss-classifying instances from another class. That results in feature representations compacting without being descriptive. For example, the worst possible solution scenario is to update the network weights to zeroes during fine-tuning the model to reach the best loss. In more detail, the standard multiple-class classification task aims to decrease the distance between instances within a class (intra-class variance) and increase the distance between instances within different classes (inter-class variance) to discriminate them, while in one-class



classification tasks, the availability of one class only limits the objective to compact instances within the one available class. Another class should be introduced to overcome this issue to force the network to be descriptive. They suggested two solutions. One possible solution is to fine-tune the model using a binary classifier with non-dysplastic tissue data as the first class and another public dataset, such as “ImageNet”, as the second class. Even though this solution will induce the discrimination ability, it is predicted to fail in discriminating dysplastic tissues from non-dysplastic tissues. The failure is attributed to the similarities between dysplastic and non-dysplastic tissue datasets, while there is a massive difference between dysplastic tissues and “ImageNet” instances. As a result, in testing such a model, the dysplastic tissue will be grouped with non-dysplastic tissue, proving that the suggested solution cannot learn compact and descriptive features.

Another promising solution that balances the compactness and the descriptiveness features in a one-class classifier was introduced by Perera and Patel (2019). Their idea focuses on updating the network weights based on a total loss composed of compactness loss and descriptiveness loss. Their solution is suitable to be adapted in this thesis to detect dysplastic tissues from NFD tissues. Also, a modification was made to the model mentioned above, including changing the compactness loss from deep one-class SVDD to batch-variance loss. The use of batch-variance loss to quantify the compact loss is motivated by the fact of the inverse proportion relationship of the distribution of the normal instances to their compactness. Usually, normal instances are expected to compact in a group, while abnormal instances suppose to scatter around the normal group.

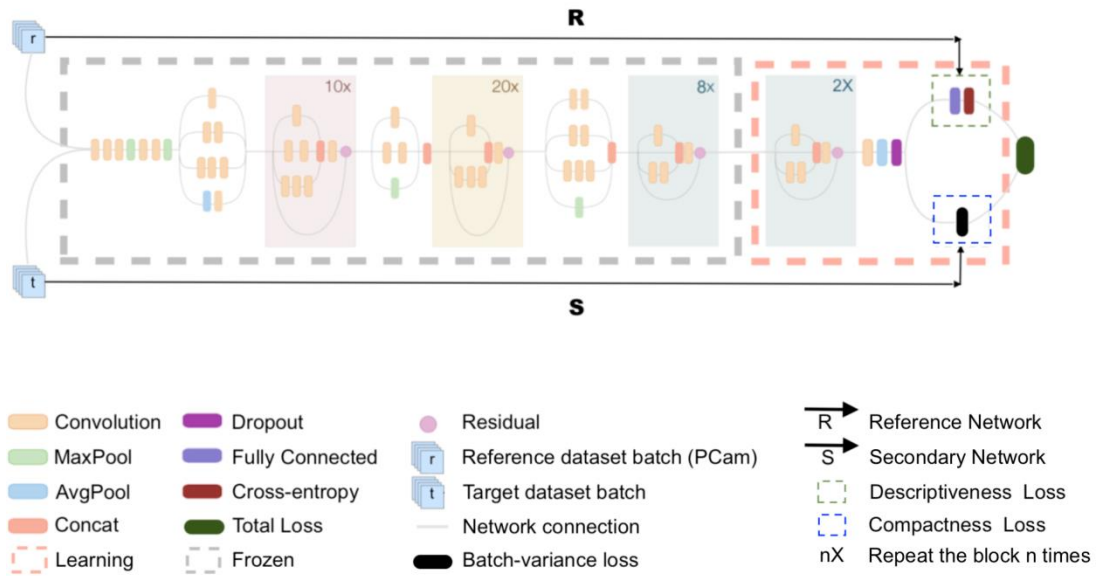
On the other hand, the descriptive loss is computed based on the network's ability to discriminate classes from an external labelled dataset using the cross-entropy loss; in this research, the Patch CAMELYON (PCam) public histological dataset was used as an additional dataset (reference dataset) to simultaneously fine-tune the model (Inception-ResNet-v2) alongside the target dataset (the driven NFD instances from Barrett's oesophagus dataset) to add the descriptiveness feature to the network. Both datasets went through the same network part of the feature extractor; nevertheless, their losses differ. For the PCam, a Softmax classifier classified the extracted features from its instances, and the cross-entropy function computed the loss (the descriptiveness loss). That loss contributes to updating the

network's parameters. After using this approach, the descriptiveness of the network is guaranteed once the learned features are capable of achieving consistent good performance in classifying the reference dataset. PCam is the most suitable reference dataset because it consists of sampled patches that were fine-grained labelled to be used with supervised learning. Also, the nature of the dataset close to Barrett's oesophagus dataset will make learning a network that can binary classify PCam instances and detect novelty in the other dataset feasible.

Figure 5.5 and Figure 5.10 show the architectures for training and testing the dysplastic tissue detector model. They are composed of a reference ( $R$ ) and a secondary ( $S$ ) networks which are identical networks that share the same set of weights ( $W$ ); however, they differ in the loss function. ( $l_c$ ) is the compact loss that is calculated based on the secondary network outputs, and ( $l_d$ ) is the descriptive loss that is calculated by the reference network and represent the binary classifier performance PCam (see section 5.5.2 for further description of the dataset).

### 5.3.1.1 Feature extraction

During the training, the last layer (Softmax classifier) from the pre-trained "Inception-ResNet-v2" model on the "ImageNet" dataset was truncated, and all its layers were frozen except the last two blocks (see Figure 5.5). The pre-trained model was set to accept a pair of batches simultaneously from both the non-dysplastic dataset and the "PCam" dataset, known as the target dataset ( $t$ ) and reference dataset ( $r$ ) respectively, as the model input. The target dataset contains NFD instances that the network target to learn their underlying representation. Then the model layers were shared by two networks, the secondary network and the reference network.



**Figure 5.5** Training framework for the potential dysplastic tissue detection model

On the one hand, the reference network accepts PCam batches with images resized to (256x256x3 pixels) to fit the pre-trained model. Also, on top of the pre-trained model, a new fully-connected layer was added that takes each produced feature representation vector of size 768 by the shared model as an input. It outputs a vector known as the logit layer of size 2. A Softmax activation function turns the logit vector into probabilities for each class label that sum to one. The form of the Softmax regression hypothesis is defined in Equation 4.1. Based on the produced probabilities, Softmax predicts whether the PCam patches belong to normal or cancer tissue, and the binary cross-entropy loss function (CE) of a classifier  $f$ , which is a special case of the categorical cross-entropy where the number of the task classes  $k$  is equal to two, and the available labels  $y_i \in \{0,1\}$ , and  $y'_i$  is the predicted label of the input  $x_i$ , calculates the descriptiveness loss of the misclassification, as shown in Equation 5.3.

**Equation 5.3**

$$l_D = CE(f) = -\frac{1}{n} \sum_{i=1}^n y_i \log(y'_i) + (1 - y_i) \log(1 - y'_i)$$

The secondary network accepts (256x256x3 pixels) batches from the non-dysplastic inputs, and then features are extracted by the convolutional layers of the network. Feature representations of size 768 are produced as the

output of the secondary network. Based on the produced feature representation, the compactness loss of the secondary network is calculated to contribute later to the total loss that will be back-propagated to update the model's parameters. As all the instances in the training dataset belong to the same class, then compactness loss ( $l_C$ ) is expected to evaluate the average intra-class variance of all the samples within a given batch. The average Euclidean distance function is used to formulate the compactness loss to quantify the compactness loss, as shown in Equation 5.4.

**Equation 5.4**

$$l_C = \frac{1}{nk} \sum_{i=1}^n s_i^T s_i$$

Assume that  $h_i$  is the corresponding produced feature representation for an input  $x_i$  By the feature extractor part from the secondary network (Inception-ResNet-v2 model). Also, assume that  $k$  is the size of  $h_i$ . For each input, it has a corresponding  $s_i$  is computed as  $s_i = h_i - m_i$ , where  $s_i$  is the distance between the  $i^{th}$  feature representation of the input and the remaining feature representations of the trained data. Mathematically, it is interpreted as the difference between the feature representation of the given sample and the mean of the remaining sample  $m_i$ , where ( $m_i = \frac{1}{n-1} \sum_{j \neq i} h_j$ ).

At the end of the forward pass, the two computed losses are accumulated to form the total loss, as shown in Equation 5.5. The total loss is used to back-propagate the error to update the shared parameters. By using that approach, the model was trained until convergence.

**Equation 5.5**

$$l_{total} = l_D + l_C$$

**5.3.1.2 Feature classification**

After training the previous model to extract intrinsic features that identify non-dysplastic tissues, a one-class classifier was trained to detect the occurrence of abnormal changes. Based on the distribution of the extracted features, LOF (Breunig et al., 2000), which is an unsupervised novelty detection algorithm, was chosen as a novelty detection method to detect the potential dysplastic tissue. The reason for choosing this is that most of the one-class classifiers do not work correctly when the feature representations of the normal class cluster form different densities. As discussed in Chapter

2, in the case of datasets that fluctuate in density, the best solution for novelty detection is to follow the local strategy instead of the global one. This situation exists in the histological structure of the oesophagus. For instance, features extracted from the epithelial layer cluster form a group, while features from the lamina propria cluster form another. To support that decision, different one-class classifiers were experimented with to help in the one-class classifier choice, and their results are presented in section 5.6.1.

LOF is an approach based on the nearest neighbour technique. It calculates and assigns a score known as an outlier factor to each data point (feature representation) that describes the degree to which the data point is an anomaly. A higher score indicates that the data point is more likely to be an outlier. LOF finds the density for each data point and compares it with the densities of its peers. The neighbourhood size ( $k$ ) parameter is needed to be set by the user. It represents the number of neighbours considered while calculating the outlier factor for each data point.

The first step of preparing a LOF classifier for novelty detection is to train it on the excitation maps ( $h_{signature}$ ) of a small non-dysplastic dataset known as a signature dataset ( $v$ ), which is drawn from the target dataset ( $v \subset t$ ) to be used as a matching template. Given a positive integer  $k$ , data points  $p_i, o \in h_{signature}$ , and  $N_k(p_i)$  a set of  $k$  neighbours to  $p_i$ , then  $distance_k$  for every data point  $p_i$  in the matching template is calculated using the Euclidean distance between  $p_i$  and its  $k^{th}$  nearest neighbour  $o$ , as shown in Equation 5.6.

**Equation 5.6**

$$distance_k(p_i) = distance(o, p_i) = \sqrt{\sum_{j=1}^n (o_j - p_{i,j})^2}$$

Then, the reachability distance, which is calculated for each point to estimate its local density, for  $p_i$  concerning  $o$  ( $reachdist_k(o \leftarrow p_i)$ ) was determined using Equation 5.7. If  $p_i$  within the neighbours of  $o$ , then the reachability distance for  $p_i$  will be equal to  $distance_k(o)$ . Otherwise, it will be the Euclidean distance between the two points.

**Equation 5.7**

$$reachdist_k(o \leftarrow p_i) = \max\{distance_k(o), distance(o, p_i)\}$$

After that, the reachability distance for  $p_i$  is used to calculate the local reachability density  $LRD_k(p_i)$  which is the inverse of the average reachability distance for  $p_i$  concerning each nearest neighbour within the  $k$  neighbour size (see Equation 5.8).

**Equation 5.8**

$$LRD_k(p_i) = \frac{1}{\left( \frac{\sum_{o \in N_k(p_i)} reachdist_k(o \leftarrow p_i)}{|N_k(p_i)|} \right)} = \frac{|N_k(p_i)|}{\sum_{o \in N_k(p_i)} reachdist_k(o \leftarrow p_i)}$$

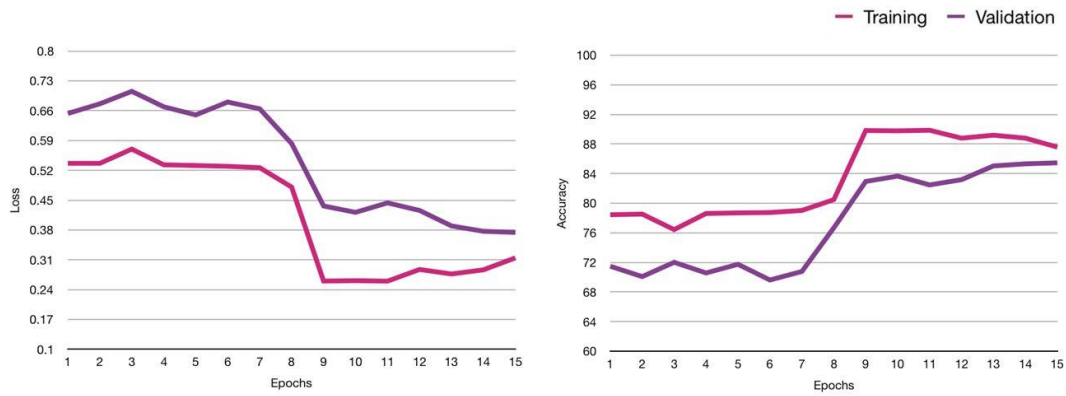
Finally, the LOF for point  $p_i$  ( $LOF_k(p_i)$ ) which will be used as an indicator of the degree of  $p_i$  is being an anomaly.  $LOF_k(p_i)$  is the mean of the ratios of the local reachability density of each data point within the  $N_k(p_i)$  to the local reachability density of  $p_i$ , and it is calculated using Equation 5.9.

**Equation 5.9**

$$LOF_k(p_i) = \frac{\sum_{o \in N_k(p_i)} \frac{LRD_k(o)}{LRD_k(p_i)}}{|N_k(p_i)|}$$

**5.3.2 Feature classification for the unfiltered patches**

After eliminating the non-dysplastic patches from the dysplastic annotations, the unfiltered patches were fed into another deep neural network to classify them as NFD, LGD, or HGD. It is important to emphasise that the classification should include the NFD category, as the first phase of the system focuses on excluding the non-dysplastic patches from the dysplastic annotation and does not recognise all the cases of the NFD class. For this purpose, a deep-learning transfer approach was adopted, similar to the architectural feature classification model in Chapter 4. The model was trained for 15 epochs, and the loss for the validation set was monitored to stop the training. Figure 5.6 shows the evolution of the losses and accuracies for the training and validation sets.



**Figure 5.6** The losses and accuracies evolutions for the training (in pink) and the validation (in purple) sets during training dysplasia low-level based classification subnetwork

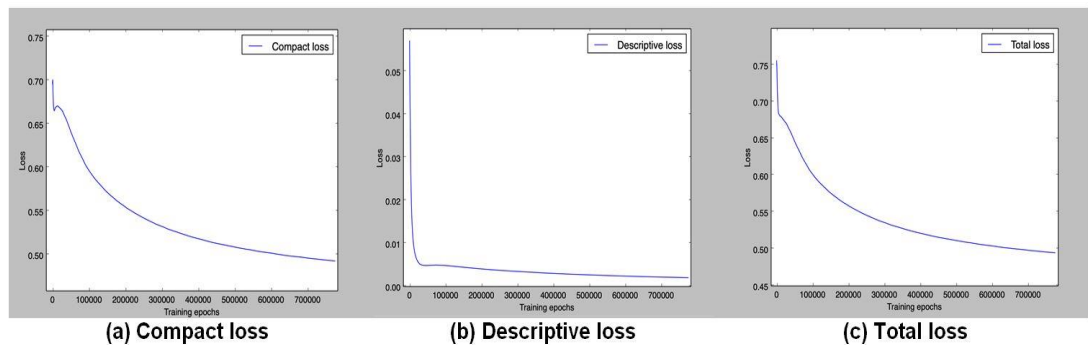
## 5.4 Experimental design

The proposed weakly-supervised model for diagnosing Barrett’s dysplasia at 40X magnification was implemented based on novelty detection for the dysplastic tissues. The detected tissue patches were further classified through a CNN-based network. Details of the training, testing, and parameter setting for the two components of the model are clarified in the following two sections.

### 5.4.1 Novelty detection training

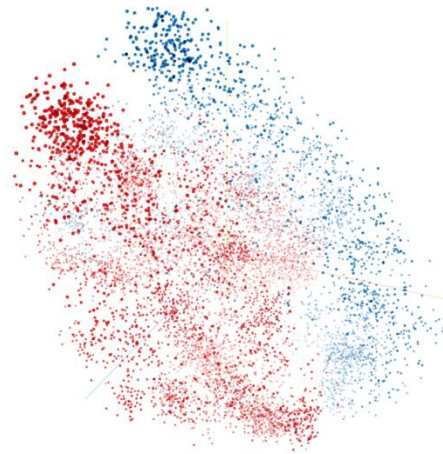
Two networks (secondary and reference networks) were coupled to train the dysplastic tissue detector model in this experiment. The secondary and reference networks share the same “Inception-ResNet-v2” model pre-trained on “ImageNet”. The transferred model’s parameters were set similarly to the settings of the models in Chapter 4. The entire model was initialised with the pre-trained model weights  $W_0$  except for the loss function and the classifier layers. For the reference network, the pre-trained model has the last layer with the size of 1000 neurones (one neurone for each class in “ImageNet”), a new layer was added to the model to fit the two-class dataset, and during the training phase all the layers of the model were frozen except the last layer, and the model was trained for 2000 iteration with 126 mini-batch sizes and a minimal learning rates ( $5e^{-5}$ ). Then, the last four blocks of the model were set to be trainable, and the two networks were trained with the same value of learning rate, using Adam optimiser and back-propagation to update

the trainable parameters based on the total loss. Choosing the number of training iterations is critical in training deep neural networks as a large or small number of iterations may lead to overfitting or under-fitting, respectively; thus, an arbitrarily large number of iterations was set, a learning rate scheduler and an early stopping method were employed to control the model performance on the validation set. The model was stopped at the 776K iteration. Figure 5.7 (a), (b), and (c) shows the compact validation loss from the secondary network, the descriptive validation loss from the reference network, and the total validation loss that was back-propagated to update the parameters of the networks. The learnt features from the trained secondary network were visualised using the t-SNE visualisation provided in Figure 5.8. The figure shows patches extracted from non-dysplastic annotations (in blue), and patches from dysplastic annotations are clustered in separate groups.



**Figure 5.7** Different loss functions during training the potential dysplastic tissue detection subnetwork. (a) Compactness loss, (b) descriptiveness loss, and (c) total loss for the validation set.





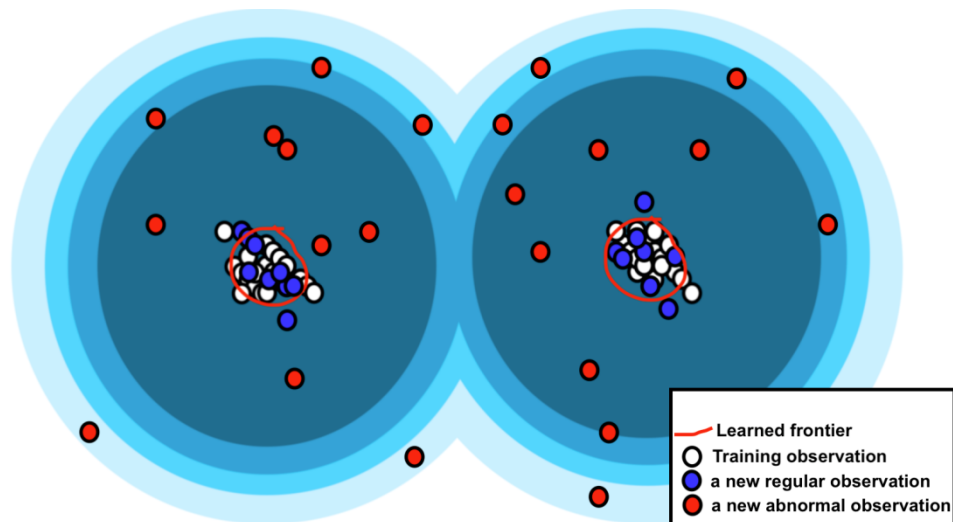
**Figure 5.8** The obtained feature space projection, using t-SNE visualisation, using the novelty detection method for the testing set patches. Non-dysplastic (in blue) and dysplastic patches (in red) are relatively separated.

#### 5.4.2 Novelty detection testing

The reference network and the compact loss were removed to prepare the model for testing. A subset from the training set was drawn to be used as a signature dataset; more details about the signature set are provided in section 5.5.3. The test preparation has two phases: template generation and one-class classifier training. The template generation phase involves obtaining the representation of the feature from the signature dataset by feeding them into the model of the trained secondary network. Each instance  $h_{v_i}$  in the templates has the size of 768 and is stored to be used in the second phase. Based on the generated templates, the LOF classifier was trained.

LOF was trained by calculating the outlier factor for each instance in the templates based on Euclidean distance and the number of neighbours ( $k$ ) using the equations in section 5.3.1.2. It is essential to be careful in selecting the number of neighbours ( $k$ ) as a large  $k$  will lose local outliers, because each point will try to reach a large number of neighbours even if they are real outliers to fulfil the number of reachable neighbours. That will result in a possible few large clusters. In contrast, a small  $k$  will have a more local focus and result in many small clusters, and with noisy data, the local focus will be huge. Generally, a rule of thumb is used to decide the value of ( $k$ ) is "the value of ( $k$ ) should be greater than the number of points that are expected to be contained in a cluster and smaller than the maximum number

of neighbour points that can potentially be local outliers”. In this experiment, different  $k$  values were used and it was found that the classifier starts degrading with  $k > 10$  and fails at  $k = 30$ , hence, the best neighbourhood size was selected to be 10. After computing all the outlier factors for the points in the template and given a test image  $y$ , the calculation of the outlier factor for the corresponding feature representation  $h_y$ , where  $h_y \in h_{test}$ , starts with finding the  $k$  – distance of the data point using the Euclidean distance between the data point and its  $k^{th}$  neighbour, then a series of calculations include finding the reachability distance, the local reachability density, and finally, the LOF for the point. Then the local reachability density for  $h_y$  is compared with the local reachability density for its  $k$  neighbours. If the local reachability density for  $h_y$  is lower, then it is considered as an outlier. More specifically, the LOF for  $h_y$  is the average ratio of the local reachability density of the neighbours of that point to its local reachability density. Therefore, if  $LOF(h_y) > 1$ , then that implies that the density of  $h_y$  is on average smaller than the density of its neighbours, and that indicates the image  $y$  is more likely to be an outlier (see Figure 5.9) because the distance from that point to the nearest point or cluster is longer than the points with  $LOF(h_y) < 1$ . The threshold for LOF can be set based on task, and it is not restricted to 1. LOF was trained on the signature dataset (see section 5.5.3). Then the matching phase, testing the test set, is accomplished by computing LOF for each test patch to decide whether the test patch is an outlier (a potential dysplastic tissue).



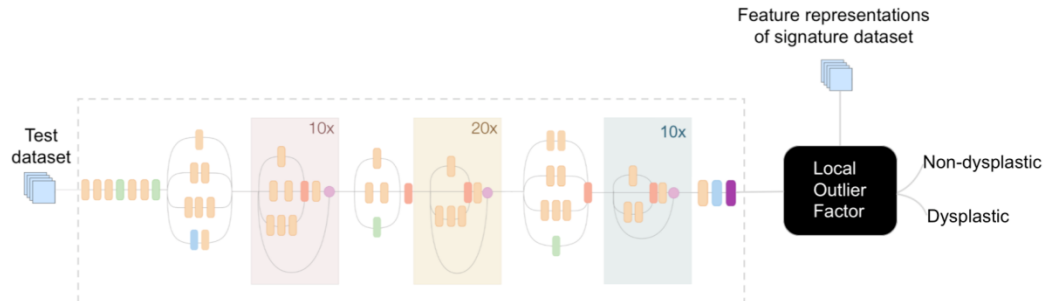
**Figure 5.9** Novelty detection with LOF

White circles belong to the training set, while blue and red circles belong to the test set. The ground truth label for the blue set is “normal”, while it is “abnormal” for the red set. Classifying the test set using LOF results in all the data points outside the red-drawn boundary are novel instances. All the points inside the bounded area have LOF score smaller than the ones outside the bounded area.

The “Scikit” library provides a “LocalOutlierFactor” class that can be trained as a novelty detector following the LOF algorithm. For creating an instance from the class, some attributes should be specified, such as the number of neighbours (“n\_neighbors”), the desired distance computation algorithm (“metric”), and a flag indicating whether the task is novelty detection (“novelty”). In the experiment, “n\_neighbors”, “metric”, and “novelty” were set to 10, “euclidean”, and “True”. After creating an instance, the method “fit” was used to train the model on the training set. Finally, the method “decision\_function” was used to calculate the shifted opposite of LOF for the test sets; thus, using this method, the larger LOF values for a sample imply that the sample is more likely to be an inlier. LOF for each sample represents a score for that sample. Based on the computed scores, the outliers were flagged based on a threshold  $\delta$  as shown in Equation 5.10, where 0 and 1 denote normal (non-dysplastic tissue) and abnormal (potential dysplastic tissue), respectively. The value of  $\delta$  was determined based on the 95%-quantile of distribution of the scores for the validation set. In this experiment,  $\delta$  was set to (-0.5). Figure 5.10 illustrates the final model ready to be used in testing new images.

**Equation 5.10**

$$Label(y) = \begin{cases} 0, & \text{if } LOF(y) > \delta \\ 1, & \text{if } LOF(y) \leq \delta \end{cases}$$



**Figure 5.10** Testing framework for the potential dysplastic tissue detection model

### 5.4.3 Dysplasia classification

As discussed earlier, the proposed model was trained using three networks; two were for training the potential dysplastic tissue detection module. The third network was a supervised network to classify the unfiltered patches. It was fine-tuned similarly to fine-tuning the architectural feature classification network in Chapter 4 and using the same parameters and techniques. The only difference is the input dataset. The used dataset in fine-tuning this network was the filtered dataset, using the trained novelty detector, sampled patches set at 40X magnification discussed in section 3.7. The output of this network is the class of the input image which is one of the three classes (NFD, LGD and HGD).

### 5.4.4 The proposed model assembly

After training and testing the earlier two models, the proposed model is ready to be assembled. The resulting feature representations from “block8\_8\_ac” (see Appendix B.1) are cached to be fed into the last two blocks of the possible dysplastic tissue detection. If the tested image is flagged as “non-dysplastic tissue”, then the images will be classified as NFD; otherwise, its temporarily saved feature representation will be fed into the last two blocks of the feature classification model to classify the input following the three-tier dysplasia classification. Finally, the decisions of the two assembled models are gathered to generate an annotation mask for the

annotations grades at 40X magnification. Figure 5.3 illustrates the proposed model.

## 5.5 Experiment datasets

This section briefly describes the datasets used in training and validating the proposed model. The potential dysplastic tissue detection model was trained using two kinds of datasets, as mentioned earlier, the target dataset and reference dataset, and a signature dataset was used to train the one-class classifier. The target and signature datasets were drawn from the sampled data provided in Table 3.5. At the same time, the low-level-based classification model was trained on the sampled patches from the training annotations of Barrett's oesophagus WSIs at 40X magnification.

### 5.5.1 Target dataset

The selected sampled patches from Barrett's oesophagus WSIs should meet the following criteria to be included in the training phase:

- The patch is accepted if it is sampled at 40X magnification
- The patch is accepted if it is sampled from the NFD annotation
- The patch is accepted if it is sampled from NFD virtual slide image

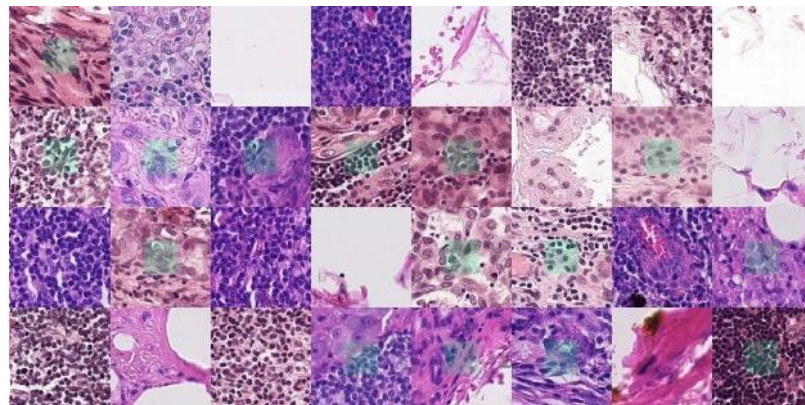
As a result, 929,561 out of 1,289,271 sampled patches extracted from 79 WSIs are nominated. 83% of the nominated sample patches, from 80% of the WSIs, were used to train the model, and the remaining patches were used for validating the model. The excluded 49 WSIs were used as an initial test before applying the model to the test set. Table 5.1 provides the number of patches used in training, validating and the initial testing of the model.

**Table 5.1** The number of patches within the training, validation, and testing sets in the target dataset

Training patches	Validation patches	Testing patches		
		NFD	LGD	HGD
775,530	154,031	50,000	151,472	105,202

### 5.5.2 Reference dataset

The “PCam” dataset is a histologic dataset sampled by (Veeling et al., 2018) to be a benchmark for histological image classification tasks. It consists of 327,680 RGB images extracted from 400 H&E whole slides images in the “Camelyon16” dataset at 10X magnification (Bejnordi et al., 2017). Each patch in the dataset is a non-overlapped 96X96 pixels extracted from a breast cancer lymph node section using a hard-negative mining regime. A binary label is assigned to it, indicating the presence of metastatic tissue (at least one pixel) in the 32X32 centre region of the patch. Any metastatic tissue in the patch on the outer region of that central region does not influence the label. The labels in the dataset are balanced (almost 50% normal examples and 50% tumour examples). The training set comprises 75% of the dataset, while the remaining is split equally between the validation and test sets. Figure 5.11 shows samples from the “PCam” dataset, where the green square in the centre of cancerous tissue is the 32X32 boxes that contain the cancerous tissue.



**Figure 5.11** Samples from the PCam dataset (Veeling et al., 2018)

### 5.5.3 Signature dataset

The signature dataset is a small dataset drawn from the training set of the target dataset to train the LOF classifier. The target dataset cannot be used in training LOF because it is a large-sized dataset, and it is infeasible to train such a classifier with a large dataset. As the signature dataset is used in the matching phase, the random selection of the patches will fail the model if the patches are selected from regions usually not participating in the diagnosis of dysplasia. For example, if the random selection leads to collecting samples from the lamina propria, then LOF will fail to recognise patches

from the epithelium. Nevertheless, the manual selection of the signature dataset samples is impossible because it is time-consuming and requires pathologists' involvement. Therefore, it is emphasised to include all the sampled patches from all the annotated regions in an NFD WSI. All the sampled patches (5487 in total) in the NFD WSI ("10570.svs") were used as the signature dataset.

#### 5.5.4 Low-level based classification dataset

Referring to Table 3.5, the number of extracted batches at 40X magnification is nine times larger than the LGD and HGD sample patches. Training the model using that distribution will lead to bias in the NFD class. Thus, the whole slides of the training set from the target dataset were excluded, as it contains the highest number of patches, and the validation and test set from the target dataset were included. Table 5.2 shows the number of patches from each grade. Also, the detected non-dysplastic patches from the dysplastic annotation were not involved in training the low-level-based classification network to avoid fuzzing the model.

**Table 5.2** The number of patches within the training and the validation sets that were drawn from the sampled patches at 40X magnification

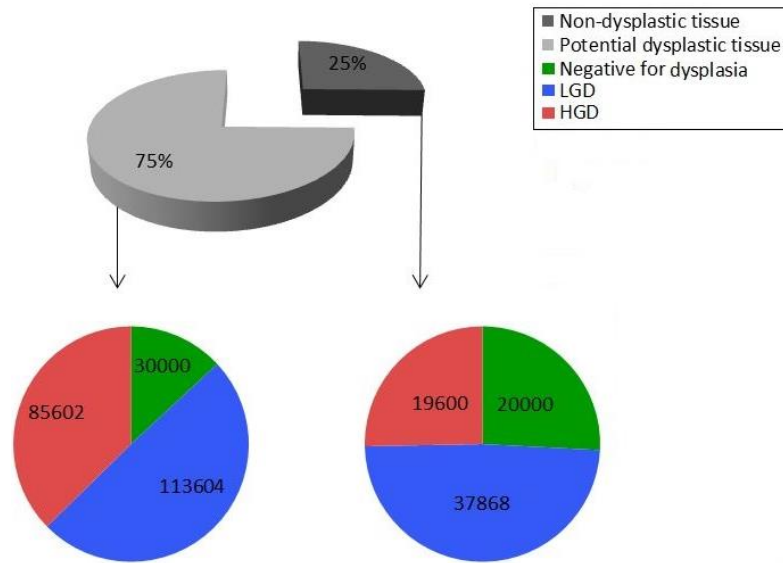
	Train	Validate	Total for each grade
NFD	204,031	36,629	240,660
LGD	113,604	25,368	138,972
HGD	82,902	30,368	113,270
Total for each set	400,537	92,365	492,902

## 5.6 Results

### 5.6.1 Potential dysplastic tissue detection

In the validation phase, the model was tested on the test set using the target dataset, which contains LGD and HGD slides (refer to Table 5.1 for more details about the number of sampled patches from each grade), 40%, 25%, and 18% from the patches in NFD, LGD and HGD annotations were detected as non-dysplastic tissue using LOF classifier (see Figure 5.12).

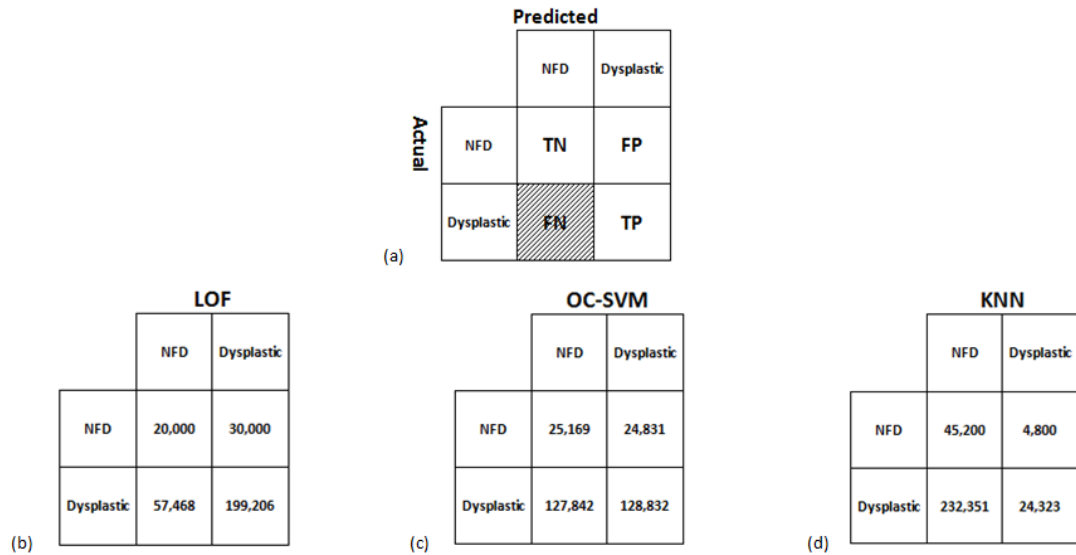
### Potential Dysplastic Tissue Detection



**Figure 5.12** Pie charts show the percentage of detected potential dysplastic tissue within the initial testing set and the number of detected patches from each grade

When designing the proposed model, several experiments were conducted using different one-class classifiers to assist in selecting the most suitable classifier. Those classifiers were OC-SVM (Schölkopf et al., 2001) and K-nearest neighbour (KNN), as it was employed by Perera and Patel (2019) with a threshold equal to zero. The confusion matrices for the classifiers, including LOF, are presented in Figure 5.13. The performance of this sub-model will be evaluated based on precision, recall and F1-score. The precision is the metric that considers the number of detected dysplastic patches from the dysplastic annotations and the number of misclassified NFD patches. As is shown in Figure 5.13 (a), the highlighted cell represents the metric that the model is not certain about its trueness. FN is the number of dysplastic patches that were predicted as NFD. In this task, it is suggested that LGD and HGD annotations might contain NFD patches. Results corresponding to LOF selection were summarised in Table 5.3. LOF outperformed the other classifiers.





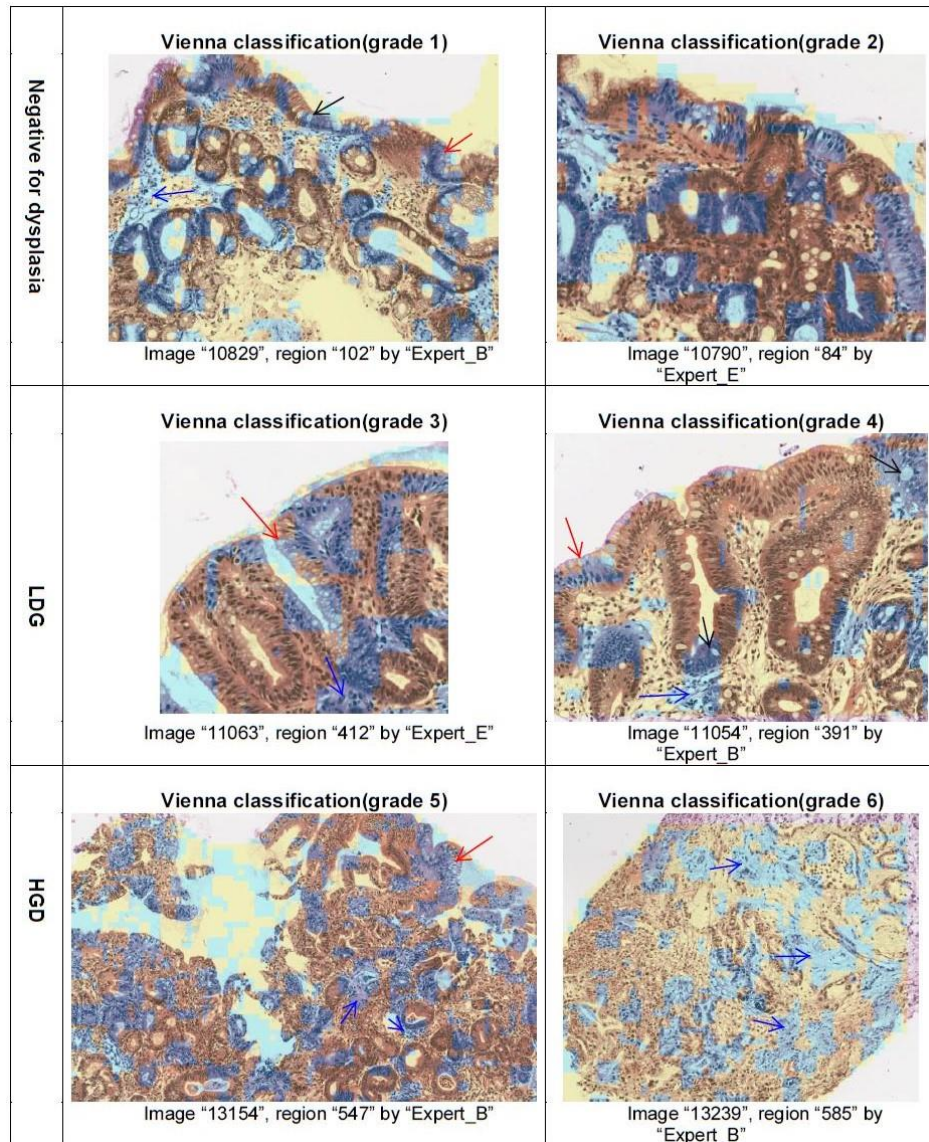
**Figure 5.13** Confusion matrices for different one-class classifiers

By using the target dataset, the fusion matrices for one-class classifiers were tested on the test set. (a) shows the correspondent metric in each cell, and (b), (c) and (d) are the LOF, OC-SVM and KNN confusion matrices, respectively.

**Table 5.3** Obtained results for design decisions relating to the one-class classifier

	OC-SVM	KNN	LOF
Precision	0.84	0.84	0.87
Recall	0.50	0.09	0.78
F1-score	0.63	0.17	0.82

Moreover, the model was tested on the test WSIs. In the NFD annotations, 36% of 75,848 patches were classified as the same class as their annotations' class, and 63% of 130,232 patches from dysplasia annotations were detected as potential abnormal changes occurring in the tissue. The percentage is equal for LGD and HGD. Figure 5.14 shows examples of detected non-dysplastic tissue from each grade. The light blue highlight indicates non-dysplastic tissue, whereas the orange indicates tissue with potential dysplastic changes. The model detected cells with mucinous (red arrow), different shapes of goblet cells (black arrow) and different textures from the lamina propria (blue arrow) as non-dysplastic tissue.



**Figure 5.14** Samples from the test annotations show the detected non-dysplastic tissues (in light blue)

### 5.6.2 Low-level based classification

The performance of this model, which is the trained "Inception-ResNet-v2" on the filtered dataset, was evaluated on the test set based on the patch, annotation, and slide levels. The model achieved an overall 0.50 precision and recall at the patch level, 0.74 specificity, 0.49 F1-score, and 52% accuracy. The model's performance was enhanced at the annotation and slide levels by testing the generated heatmaps on the trained random forest classifier (as discussed in section 4.3.2) to decide their grades. The model achieved 0.70 precision, 0.65 recall, 0.84 specificity, 0.63 F1- score and 63% accuracy at the annotations level, while it scored 0.87 precision, 0.90 recall, 0.94 specificity, 0.86 F1-score and 87% accuracy at the slide-level.

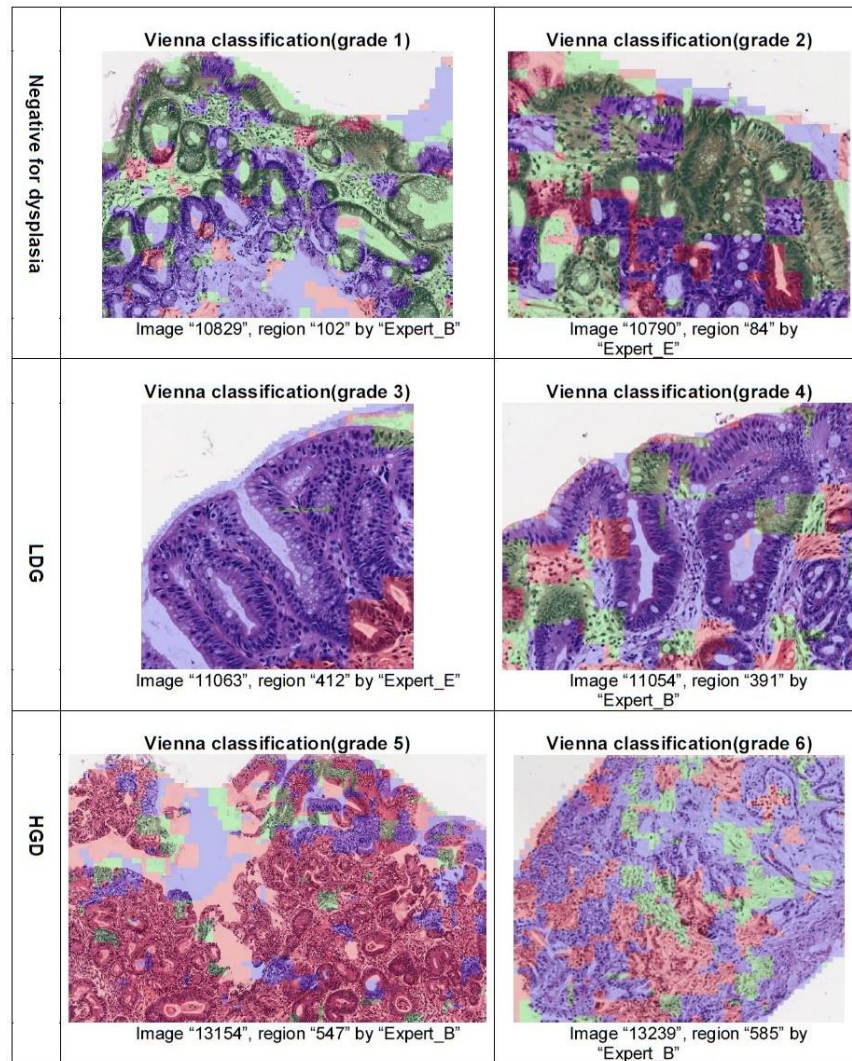
Table 5.4 shows the confusion matrices, and Table 5.5 shows the metrics for each grade. Predictions maps for samples from each class are shown in Figure 5.15.

**Table 5.4** The confusion matrices for the features classification model based on analysing slides at 40X magnification at the patch, annotation and slide levels (three-tier)

	Patch-wise			Annotation-wise			Slide-wise		
	NFD	LGD	HGD	NFD	LGD	HGD	NFD	LGD	HGD
NFD	21226	41243	13379	12	11	0	5	2	0
LGD	12723	65014	13739	2	7	1	0	3	0
HGD	5849	12646	20347	0	5	14	0	0	5

**Table 5.5** The performance measurements for the features classification model based on analysing slides at 40X magnification at the patch, annotation and slide levels (three-tier)

	NFD	LGD	HGD
<b>Patch-wise</b>			
<b>Precision</b>	0.53	0.55	0.43
<b>Recall</b>	0.28	0.71	0.52
<b>Specificity</b>	0.86	0.53	0.84
<b>F1-score</b>	0.37	0.62	0.47
<b>Accuracy</b>	64%	61%	78%
<b>Annotation-wise</b>			
<b>Precision</b>	0.86	0.30	0.93
<b>Recall</b>	0.52	0.70	0.74
<b>Specificity</b>	0.93	0.62	0.97
<b>F1-score</b>	0.65	0.42	0.82
<b>Accuracy</b>	75%	63%	88%
<b>Slide-wise</b>			
<b>Precision</b>	1.00	0.60	1.00
<b>Recall</b>	0.71	1.00	1.00
<b>Specificity</b>	1.00	0.83	1.00
<b>F1-score</b>	0.83	0.75	1.00
<b>Accuracy</b>	87%	87%	100%



**Figure 5.15** The prediction maps for annotations from the test set from each grade, following the Vienna classification

The colours green, blue and red indicates NFD, LGD, and HGD, respectively.

### 5.6.3 The proposed model

The assembled model from the two architectures in section 5.3.1 and section 5.3.2 was tested, and its results for each grade are listed in Table 5.6 and Table 5.7. The scored performance results for the model at the annotation-level/ slide-level are 0.72/0.83 precisions, 0.67/0.84 recalls, 0.84/0.92 specificities, 0.64/0.80 F1-scores and 63%/80% accuracies.

**Table 5.6** The confusion matrices for the proposed model on the annotation and slide levels (three-tier)

	Annotation-wise			Slide-wise		
	NFD	LGD	HGD	NFD	LGD	HGD
NFD	13	10	0	5	2	0
LGD	1	8	1	0	3	0
HGD	0	7	12	0	1	4

**Table 5.7** The performance measurements for the proposed model on the annotation and slide levels (three-tier)

	NFD	LGD	HGD
<b>Annotation-wise</b>			
<b>Precision</b>	0.93	0.32	0.92
<b>Recall</b>	0.57	0.80	0.63
<b>Specificity</b>	0.97	0.60	0.97
<b>F1-score</b>	0.70	0.46	0.75
<b>Accuracy</b>	79%	63%	85%
<b>Slide-wise</b>			
<b>Precision</b>	1.00	0.50	1.00
<b>Recall</b>	0.71	1.00	0.80
<b>Specificity</b>	1.00	0.75	1.00
<b>F1-score</b>	0.83	0.67	0.89
<b>Accuracy</b>	87%	80%	93%

## 5.7 Discussion

The proposed model for grading Barrett’s related dysplasia based on analysing WSIs at 40X magnification is composed of two parts that were separately trained and tested before they were assembled. The first part is a model that filters non-dysplastic tissues from the annotations using a novelty detection method. The performance of the novelty detection model can be measured and evaluated only in two ways, by computing the patch-wise performance and the annotation and slide performance. To compute the performance at the patch level, F1-score for the model, which is the average between precision and recall, was used. Although the performance of the model was used in the evaluation of the dysplastic tissues, it was not enough as FN contributed to the calculation, and in this task, FN is not certain for the reason that the dysplastic tissue might have normal regions



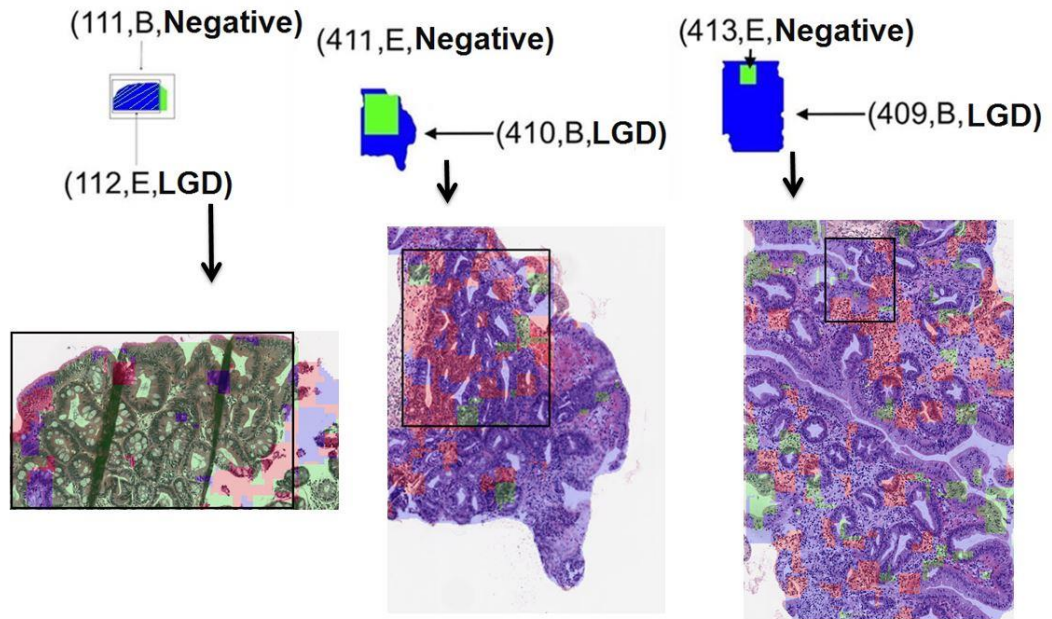
and the abnormalities are affecting some parts of the tissue; thus only non-dysplastic tissues are assumed to be healthy. As a result, the evaluation of the model is supported by measuring how the model enhanced the signature dataset, which was used in matching the new examples to decide their class. A potential solution to enhance the performance of the supervised classification at the annotation and slide levels when it is stacked on top of the low-level-based classification network. On the patch-wise performance, the F1-score of the model was 0.82. Many undetected patches belong to the NFD class, mainly caused by the limited number of templates in the signature dataset used to match the new examples to decide their class. A potential solution to enhance the performance of this part of the system could involve a guided selection of the signature dataset by pathologists. Such a method guarantees a very high degree of accuracy in selecting the template set that covers most of the non-dysplastic cases and decreases the redundancy in the template.

The second part was compared to the performance of an identical network trained in the same manner as training the proposed network, except that all the patches in the dataset were used without following the novelty detection approach to clean the data during the training time. It was found that the proposed network scored better accuracy and loss during the validation, as training with the uncleansed data achieved 80% accuracy and reduced the loss to 0.4 only. The proposed network also achieved better performance than the classification based on the analysis at 10X magnification (see Chapter 4). On the one hand, grading the dysplastic annotations and slides yielded a very high performance in grading HGD on the annotation-level and the slide-level, as F1-score reached 0.82 and 1.00, respectively. Also, the model's performance in grading LGD slides increased significantly from 0.31 to 0.42 F1-score on the annotation-level and from 0.25 to 0.75 F1-score on the slide-level. On the annotation-level, one annotation, region "258", the grade was corrected from high grade based on the analysis at 10X magnification features to low grade based on the analysis at 40X magnification features. In addition, region "257" was correctly upgraded from NFD to LGD. On the other hand, recognising of the metaplastic annotations and slides increased from 0.54 and 0.73 F1-score to 0.65 and 0.83 F1-score for annotations and slides.

Generally, the results show that limited success is achieved to grade LGD at the analysis of both magnifications compared to the two other grades, proving that the most challenging grade has the fuzziest extracted features at a low level for the classifier to discriminate. That might be attributed to two facts, the nature of this grade as it is located in the middle of a continuous spectrum with undefined boundaries between NFD and HGD grades. The other fact is that the annotations used in testing the model belong to misleading slides. A misleading slide is identified as one of the following:

- A slide containing non-overlapped regions that different pathologists annotated led to different diagnoses for the slide.
- A slide containing overlapped regions that different pathologists annotated have different grades. In the case of different labelled overlapped regions that one pathologist annotates, the overlapped area is assigned to the lowest grade, as we can consider that the pathologist excluded it.

For instance, Figure 5.16 shows annotation masks for overlapped LGD annotations with NFD. The overlapping in the annotations resulted from disagreements between two pathologists and not from the presence of a lower grade lesion within a higher annotation. The model classified region “111” and the contained “112” region as NFD and the classified region “413” and the contained “409” region as LGD, agreeing with “Expert\_B” in both cases. While the cytological abnormalities in region “411” (the surrounded tissue by a rectangle in Figure 5.16), which was annotated from the lamina propria layer and showed a crowded glandular arrangement, were classified as HGD. In contrast, the containing annotation “410”, which shows regions from the epithelial layer, was classified as LGD.



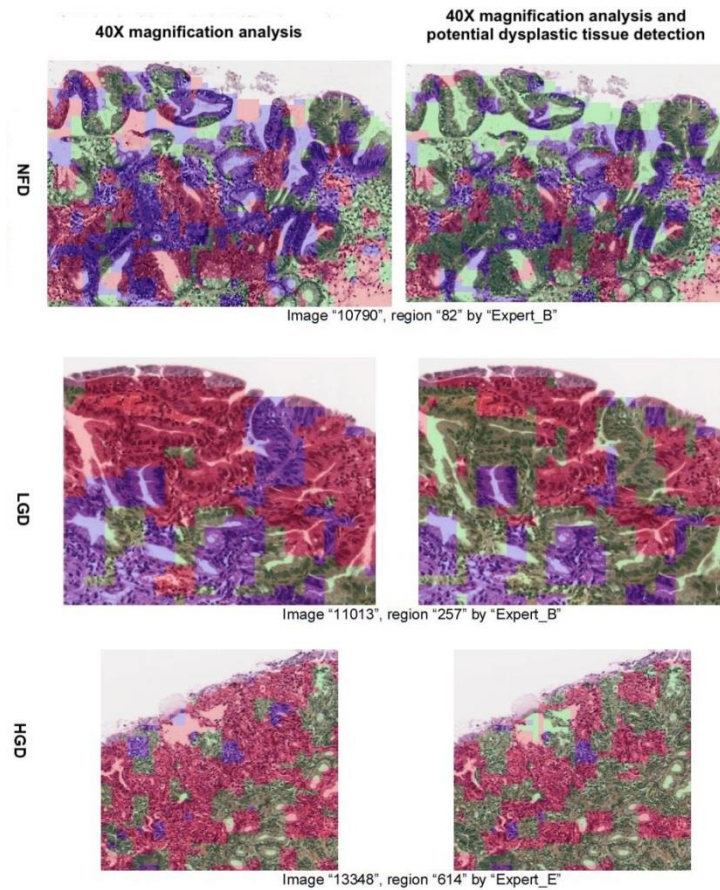
**Figure 5.16** Examples of misleading overlapped annotations

After adding the novelty detector to filter the non-dysplastic tissues within the dysplastic annotations, the results were slightly increased compared to the earlier Barrett's oesophagus related dysplasia classifier based on analysing slides at 40X magnification at the annotation-level by increasing the performance from 0.63 to 0.64 F1-score. However, the performance decreased from 0.86 to 0.80 F1-score at the slide-level. Also, it reduced the interobserver agreement with "Expert\_B" at the annotation-level from substantial agreement to moderate agreement. By investigating the effect of the added module on the performance for each grade, it was found that the addition enhanced the performance of grading NFD and LGD only. The model's performance in classifying HGD annotations was decreased from 0.82 to 0.75 F1-score as two annotations, and one slide was downgraded from high to low grade. In general, the addition is considered an excellent solution to shed light on the grey area where experts usually disagree; otherwise, it degrades the performance when detecting the highest degree of dysplasia. Table 5.8 and Figure 5.17 compare the results of the dysplastic classification based on the analysis at the 40X magnification model before and after adding the novelty detection part.



**Table 5.8** Performance measurements for features classification model based on analysing slides at 40X magnification solely against it is coupled with potential dysplastic tissue detection model

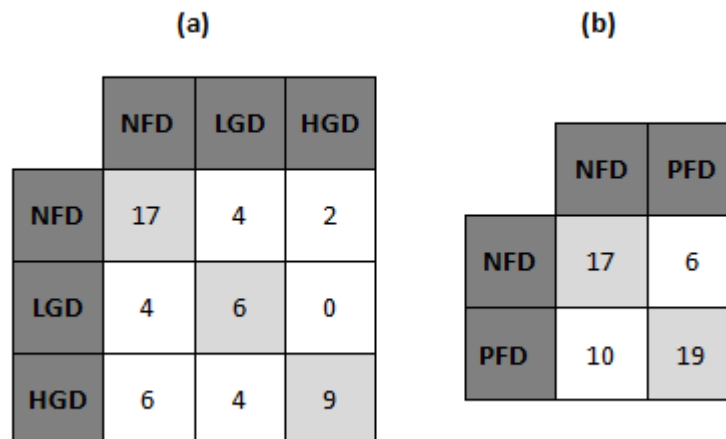
	Classification at 40X	The proposed model
	NFD	Dysplasia
	Annotations-wise	
Precision	0.70	0.72
Recall	0.65	0.67
Specificity	0.84	0.84
F1-score	0.63	0.64
Accuracy	63%	63%
Agreement with "Expert_B"	0.638	0.585
Agreement with "Expert_E"	0.299	0.336
	Slide-wise	
Precision	0.87	0.83
Recall	0.90	0.84
Specificity	0.94	0.92
F1-score	0.86	0.80
Accuracy	87%	80%
Agreement with "Expert_B"	0.80	0.704
Agreement with "Expert_E"	0.286	0.286



**Figure 5.17** The analysis provides the prediction maps for three test annotations from each grade at the 40X magnification model (on the left) against its predictions coupled with the potential dysplastic tissue detection model (on the right)

The evaluation includes comparing the proposed model results in this chapter against another work by Sali et al. (2020). They conducted a comparative study that included supervised, weakly supervised, and unsupervised deep-learning approaches to identify dysplastic and non-dysplastic Barrett's oesophagus tissue in WSIs. Their weakly-supervised approach investigated MIL as each annotation was considered a bag with a label and contained multiple instances. They trained a CNN on patches that were extracted from those annotations. Then, they used the trained model to output probabilities for each class in the task for each extracted patch. After that, a histogram of the class's probability distribution for each annotation was then generated. The generated histograms for the training set were used to train a support vector machine classifier. The trained classifier was employed to predict the annotation class for a test set. Their experiment was regenerated using the dataset used in this thesis and "Inception-ResNet-v2". The confusion matrices for their approach at three-tier and two-tier

classification are provided in Figure 5.18 (a) and (b), respectively. Table 5.9 presents results for their model against the proposed model using three-tier and two-tier classifications. It is important to mention that due to the small dataset, the strength of evidence the comparison results provide is weak. By running the experiments using the same test set and relying on precision, recall, specificity, F1-score and accuracy, the proposed model scored better performance compared to the applied approach by Sali et al. (2020). In addition, the proposed model has a moderate agreement at the annotation-level with the pathologists, while the other work has a fair agreement. The increase in the performance of the proposed model compared to the model of Sali et al. (2020) is attributed to the weakly-supervised approach used to filter the non-dysplastic tissue from the dysplastic training annotations. Also, the inclusion of the earlier mentioned approach to the proposed model to detect tissues where they are expected to be dysplasia. Finally, considering the distribution of each grade in each layer separately in the generated heatmaps during inferring the annotation-level grade might be one of the reasons for the performance enhancement.



**Figure 5.18** The annotation-level confusion matrices for Sali et al. (2020)

(a) follows the three-tier classification and (b) follows the two-tier classification

**Table 5.9** The three-tier and two-tier classification performances for the proposed model against the work proposed by Sali et al. (2020) at the annotation-level

	Three-tier classification		Two-tier classification	
	The proposed model 95% confidence interval	(Sali et al., 2020) 95% confidence interval	The proposed model 95% confidence interval	(Sali et al., 2020) 95% confidence interval
<b>Precision</b>	0.72 (+/-0.12)	0.63 (+/-0.13)	0.83 (+/-0.1)	0.69 (+/-0.13)
<b>Recall</b>	0.67 (+/-0.13)	0.60 (+/-0.14)	0.77 (+/-0.12)	0.70 (+/-0.13)
<b>Specificity</b>	0.84 (+/-0.1)	0.80 (+/-0.11)	0.77 (+/-0.12)	0.70 (+/-0.13)
<b>F1-score</b>	0.64 (+/-0.13)	0.59 (+/-0.14)	0.77 (+/-0.12)	0.69 (+/-0.13)
<b>Accuracy</b>	63% (+/-13)	62% (+/-14)	79% (+/-11)	69% (+/-13)
<b>Agreement with the experts</b>	0.476	0.400	0.553	0.387

Considering the interobserver agreement for the models of this chapter with the pathologists, Table 5.8 summarises the agreement levels on the annotation-level and the slide-level. On the one hand, based on the agreements at the annotation-level, the proposed model in this chapter did not change the agreement level with "Expert\_B" compared to the model proposed in Chapter 4; however, the model without the novelty detection module managed to upgrade the agreement with "Expert\_B" one level from moderate to substantial, and both models in this chapter upgraded the agreements with "Expert\_E" from slight to fair. On the other hand, based on the slide-level agreement, both models in this chapter increased the agreement with "Expert\_B" to a substantial level, and the agreement with "Expert\_E" was a fair agreement. The assemble model scored 0.704 and

0.286 KVs compared to the reported mean scores of 0.23 and 0.365 for the agreement with the two pathologists (Treanor et al., 2009).

Following the two-tier classification (dysplasia against non-dysplasia), the proposed model achieved the slide-level 0.83 precision, 0.77 for recall, specificity and F1-score and 79% accuracy. It reached 0.658 KV, a “substantial” agreement with “Expert\_B” similar to the agreement level between the pathologists, and 0.435 KV, a “moderate” agreement with “Exper\_E”. The proposed model scored a higher level of agreement with “Expet\_B” than “Expert\_E”.

## **5.8 Conclusion**

The grading of Barrett’s related dysplasia is complicated, with low interobserver and intraobserver agreements even amongst experienced gastrointestinal tract pathologists. The diagnostic process relies on a combination of morphological abnormalities and the architectural structures of the components of the different layers of the oesophagus. This chapter discussed a study that analysed the WSIs at high magnification, and it addressed the issue of coarse-grained labelling in the provided annotated regions. It presented a classification model relying on a low-level analysis (40X magnification). The proposed model consists of two stages. The first one is a novelty detection based model to filter the non-dysplastic tissues; in other words, to nominate tissues suspected to be abnormal. The first submodel was used to exclude the non-dysplastic patches from the dysplastic annotations for the training set, aiming to train the model in a supervised manner. Furthermore, the second stage is to classify the nominated patches. This chapter aims to determine the effect of adding the novelty detection approach to solve the weakly supervised problem in two ways. One uses the novelty model to cleanse the training dataset, which was the main contribution of this chapter, and the other adds it to the network to nominate dysplastic regions. Based on the enhanced performance after employing the proposed solution, as discussed in section 5.6.3, hypothesis H1 states that “the provided “positive for dysplasia” annotations might have non-dysplastic tissues” is accepted.

After comparing the classification based on the low-level analysis against the high-level, the low-level analysis approach outperforms the other significantly. Also, it surpasses the performance of recognising NFD and LGD when the novelty detection model accompanies it. That is the first attempt to investigate the novelty detection approach in a weakly supervised problem to the best of our knowledge. That is the first time this approach has been proposed to solve a weakly supervised task.

The proposed model offers a solution to recognise the most challenging grade where the pathologists disagree. Despite the success of adding the novelty detection module, a significant limitation is introduced when the model aims to recognise the HGD or “intramucosal carcinoma” for operational purposes. It is found that the model works better for higher grades without adding the novelty detection submodel, as the performance is slightly affected by the addition. Moreover, the proposed model scored high-performance measurements when used to detect dysplastic slides. It showed far higher interobserver agreement with one of the pathologists than the agreement between the two pathologists.

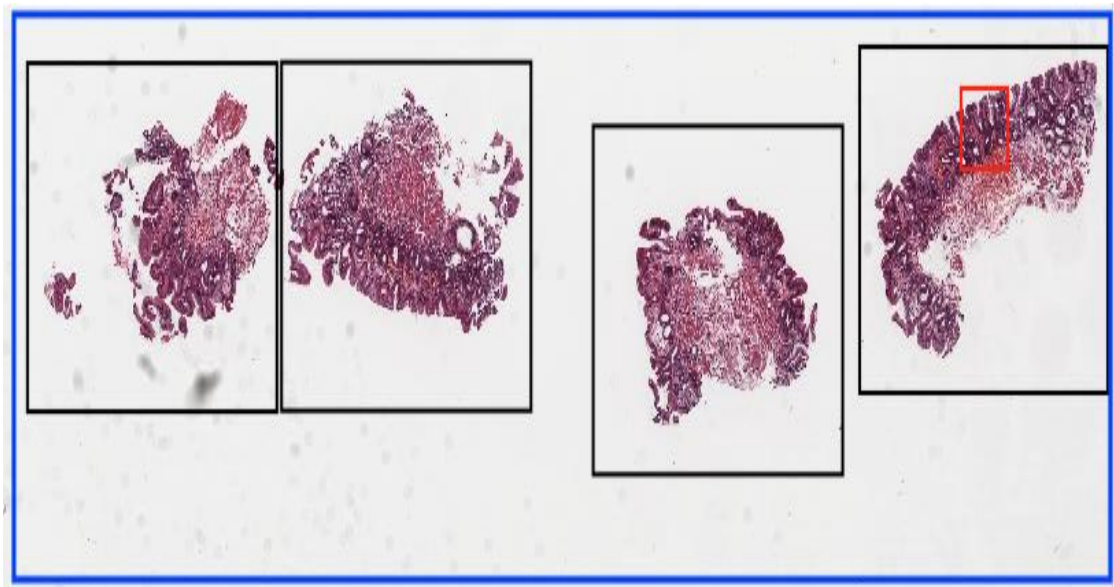
A comparative study between the proposed model in this chapter and the work proposed by Sali et al. (2020) was conducted to study the effect of the novel proposed solution on performance. In selecting a benchmark model, a weakly supervised model that operated on extracted images at 40X magnification and offered a solution to grade Barrett's oesophagus were the main criteria for a fair comparison. The same training and testing sets were used in regenerating the results using their model, and the model was trained until convergence. Based on the available small dataset, the proposed model in this chapter scored better than the other solution.

The next step of this research is to explore the effect of combining the classification models based on the high-level analysis, as discussed in Chapter 4 and on the low-level analysis, as discussed in this chapter. The next chapter will incorporate the discussed approaches in Chapter 3, Chapter 4 and Chapter 5 to automate the diagnosing process and offer a CAD system.

## Chapter 6. Automated Dysplasia Detection and Grading in the Whole Virtual Slides

### 6.1 Introduction

The models were tested on annotations, which represent parts of the biopsies extracted from patients, in Chapter 4 and Chapter 5. Multiple biopsies from each patient were gathered on a glass slide. Figure 6.1 illustrates the relationship between annotations (an example annotation bounded by the red box), biopsies (within black boxes) and WSI (bounded by the blue box). As a result, the tested annotations have smaller tissue contents and require a smaller memory size to be processed compared to their container whole virtual slides. While considering analysing the WSIs at high resolution requires high memory usage and processing time. For instance, the average size of the test set images has more than 130,000 patches at 40X magnification. Performing complex image analysis tasks on those patches will dramatically increase the computational cost and time. Therefore, decreasing the tissue area to be analysed at the highest magnification appears as the right solution.



**Figure 6.1** The relation between annotation (the red box), biopsies (the black boxes) and WSI (the blue box that surrounds the whole image)

This chapter discusses the proposed CAD system that detects the dysplastic tissue and then determines the degree of dysplasia in the tissue within the H&E stained WSI. The system analyses the WSIs in a pyramidal multi-magnification approach that processes the image at the lowest available magnification and switches to higher magnifications when more information is needed about the tissue to evaluate the lesion, mimicking the diagnosis of the pathologists. The proposed CAD system utilised the proposed systems in Chapter 4 and Chapter 5 to build the final system. The model in Chapter 4 was used to detect dysplastic regions to be classified later by the model in Chapter 5.

The main research contributions in this chapter are three solutions. The first solution is a novel attempt to find the consensus between the analyses of the WSIs at two different magnifications. That solution attempts to follow the pathology guidelines for diagnosing dysplasia in Barrett's oesophagus by assuming the extracted features at 10X magnification are architectural features. In contrast, the extracted ones at 40X magnification are cytological features. The second one is a distinctive solution that imitates the pathologists in diagnosing different grades of Barrett's oesophagus dysplasia. Finally, it contributes to the histopathology community by offering a novel, fully automated model that examines every region in the lesion instead of randomly selecting a region that risks neglecting critical regions.

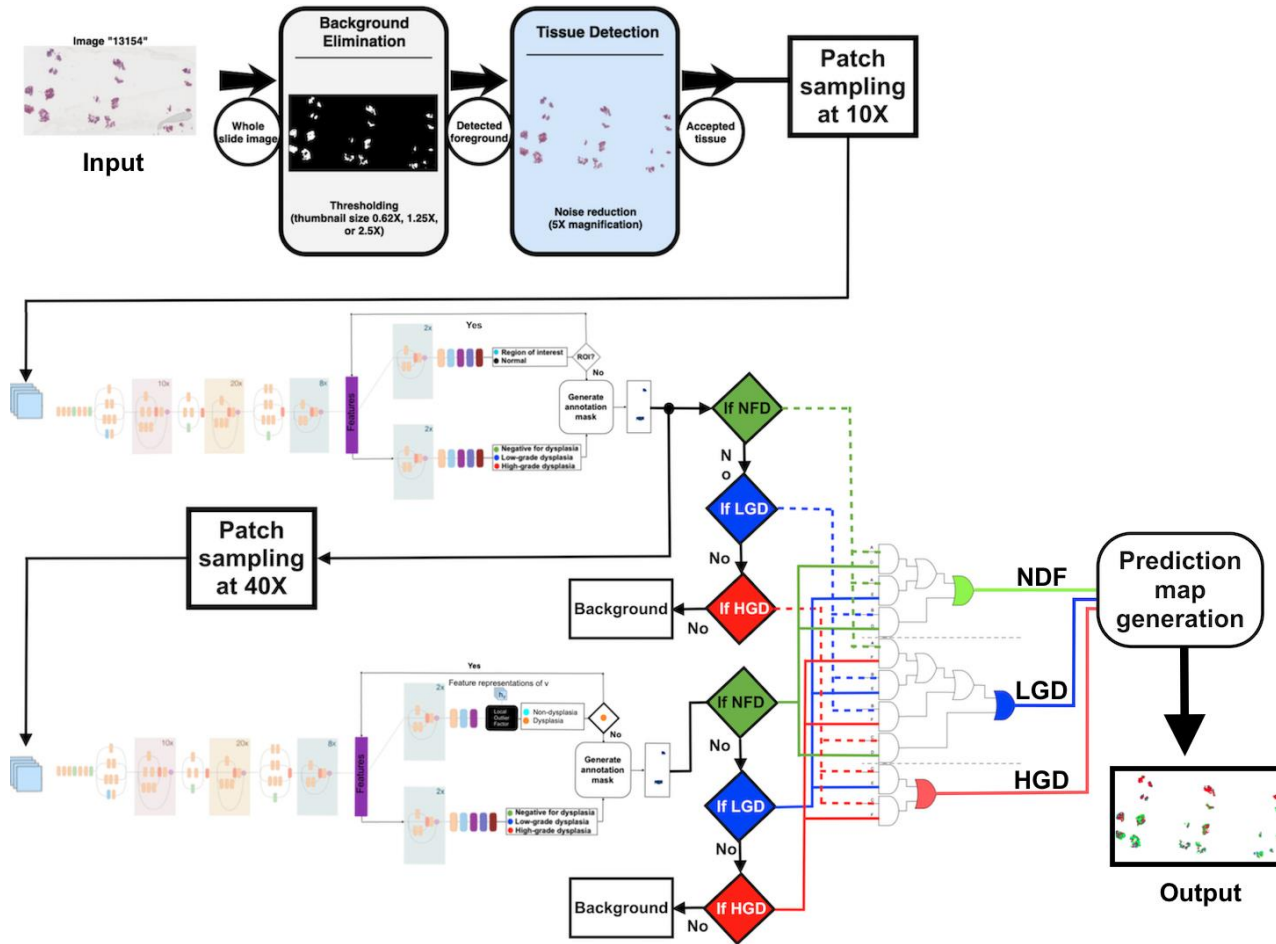
## **6.2 Methodology**

Many published works on histopathology CAD systems are limited to diagnosis annotations extracted by expert pathologists. Whereas a fully automated system is expected to take the raw WSI of a patient and produce all the clinically relevant tissues that the pathologist should focus on to diagnose the patient, generate a heatmap with the accurate assessment of tissues within each biopsy, or provide a diagnosis for the WSI and each biopsy within it. Implementing such a system is a challenging task. One reason why it is such a challenging task is that the system should have the ability to analyse a considerable amount of data that contains a large number of structures and various abnormalities.



Three possible approaches are considered for implementing such a system after detecting the tissue in the slide. These approaches use a high single-magnification scheme for detecting and classifying various tissue structures, using multi-magnifications to analyse the whole tissue or using the lowest magnification to analyse the tissue and go deeper whenever more information is needed. The first two solutions are likely to be more expensive because they need to evaluate each part in the detected tissue in the same way regardless of their importance; for example, they will process the epithelium and the lamina propria similarly despite the fact that the epithelial layer is the most contributing layer in the diagnosis. The high cost of the first two approaches makes the third solution a good alternative.

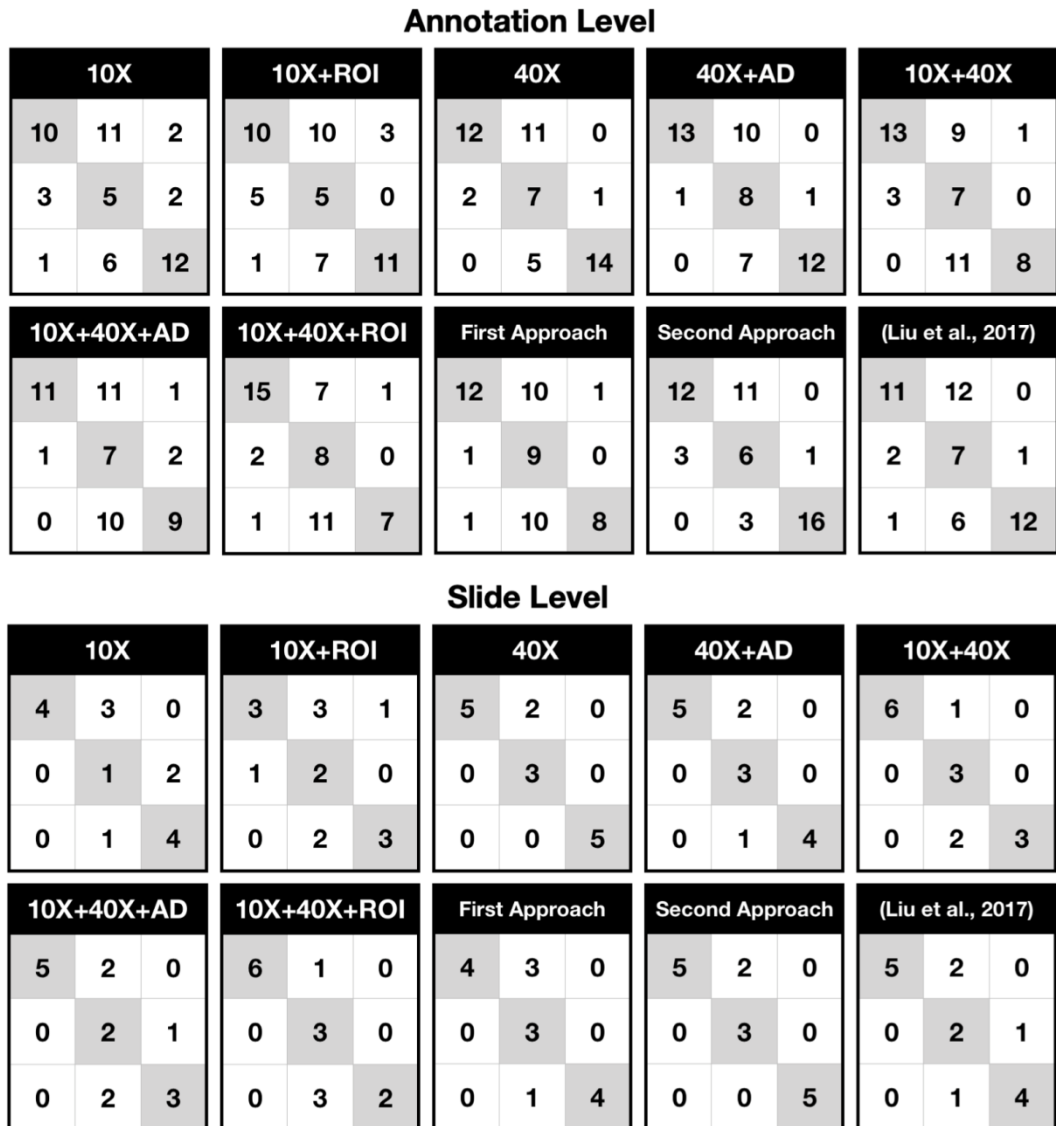
There are two options for building the CAD system from the networks in the previous chapters by considering the third approach. On the one hand, the first option is to assemble the two proposed models from Chapter 4 and Chapter 5, as illustrated in Figure 6.2, by sampling patches from the output prediction maps from the proposed model in Chapter 4 and feeding them as the input of the proposed model in Chapter 5. Some rules are set to decide the grade of the detected regions based on the classification at 10X and 40X magnifications.



**Figure 6.2** The architecture of the potential CAD system following the first approach  
The "first approach" is the first attempt to design a CAD system. More information is in section 6.2.1

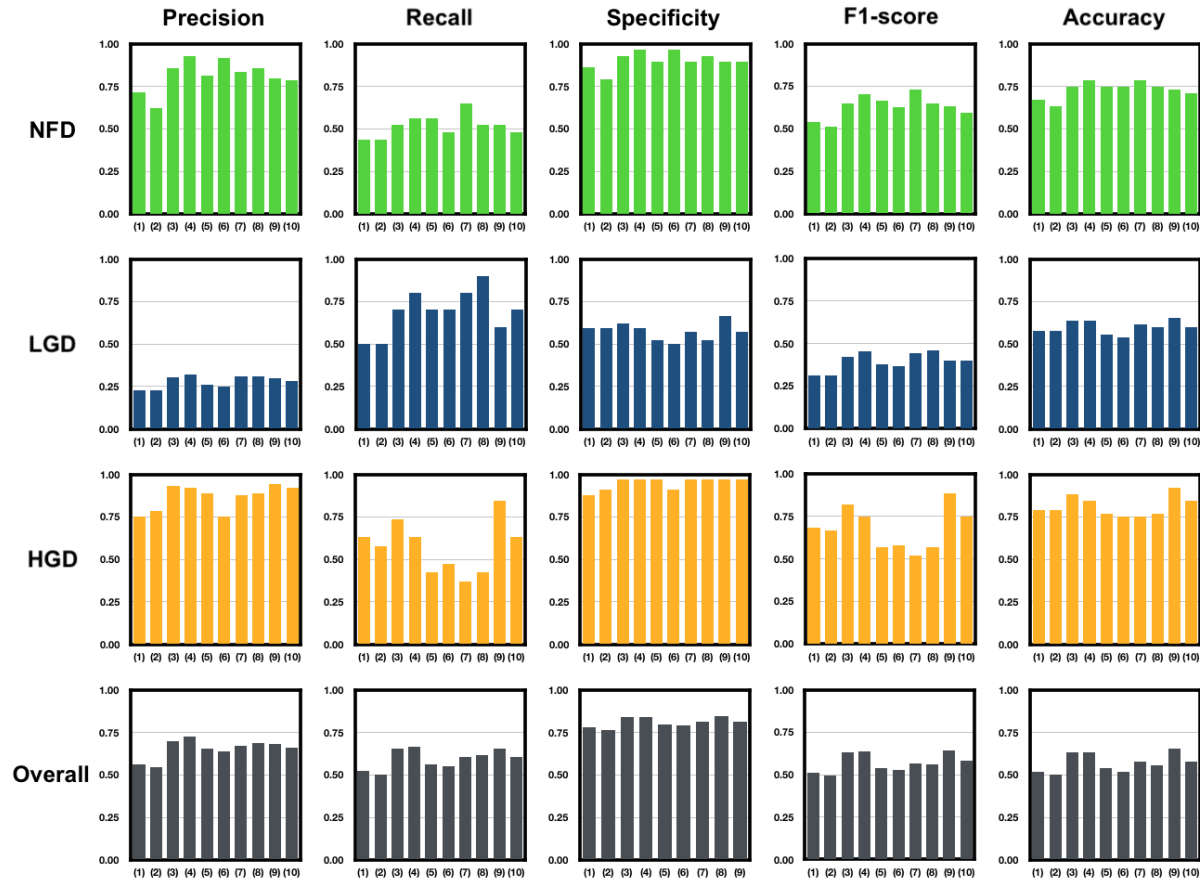
On the other hand, the second option was concluded from the following analysis. Figure 6.4 and Figure 6.5, which show bar charts for the performance metrics at the annotation-level and the slide-level for each grade (NFD, LGD and HGD in green, blue and red, respectively) and Figure 6.3, is concluded that the model in Chapter 4 has a very high specificity for NFD (0.86 at the annotation-level and one at the slide-level). That suggests that dysplasia is rarely misclassified as NFD only uses the analysis at 10X magnification. Whereas the analysis at 40X magnification increased the recall for that grade slightly (from 0.43 to 0.52 at the annotation-level and from 0.57 to 0.71 at the slide-level), suggesting that higher magnifications are not necessary for regions that were classified as NFD, as there is a small chance to be classified as higher grades using higher magnifications level. However, using the assumed cytological features ( at 40X magnification) increased the recall for the HGD (from 0.63 to 0.74 at the annotation-level and from 0.80 to 1 at the slide-level). It sharply increased the recall for LGD (from 0.50 to 0.70 at the annotation-level and from 0.33 to 1 at the slide-level), emphasising the need to use higher magnifications for those two grades. In addition, from our observation of the pathologists' techniques in diagnosing Barrett's related dysplasia, we find that the pathologists relied on 5X and 10X magnifications to diagnose NFD and used 20X and 40X in diagnosing LGD and HGD, the bar chart in Figure 3.8 summarises the number of annotations from each grade and the magnification levels that used to decide their grades by each pathologist. From those findings, a conclusion was drawn that using the high-level analysis model from Chapter 4 is sufficient and accurate as a dysplasia detector, and the tissue that is classified as NFD is accurately diagnosed and involving higher magnification in the diagnosis increases the computational time and cost without increasing the performance of the diagnosis. Thus, only regions classified as LGD and HGD at 10X magnification are processed at 40X magnification analysis. When choosing between the low-level analysis model (see section 5.3.2) or the proposed model in section 5.4.4, there is a medical consideration regarding the priority of detecting each LGD or HGD. For instance, adding the anomaly detector enhanced the recall for NFD and LGD, yet it decreased the performance of recognition HGD. Hence, using the classification network based on low-level analysis without adding the anomaly detector is the right choice. Section 2.1.7.2 discussed that in pathology, detecting HGD has a higher priority because remedial action (surgery) should be taken to prevent dysplasia from developing into cancer. While detecting LGD is vital to

schedule more endoscopic surveillance to monitor the progression of dysplasia (see Figure 2.7) (Wang and Sampliner, 2008). By weighing the pros and cons, detecting HGD has the highest priority, as prevention, while detecting LGD is a precaution.



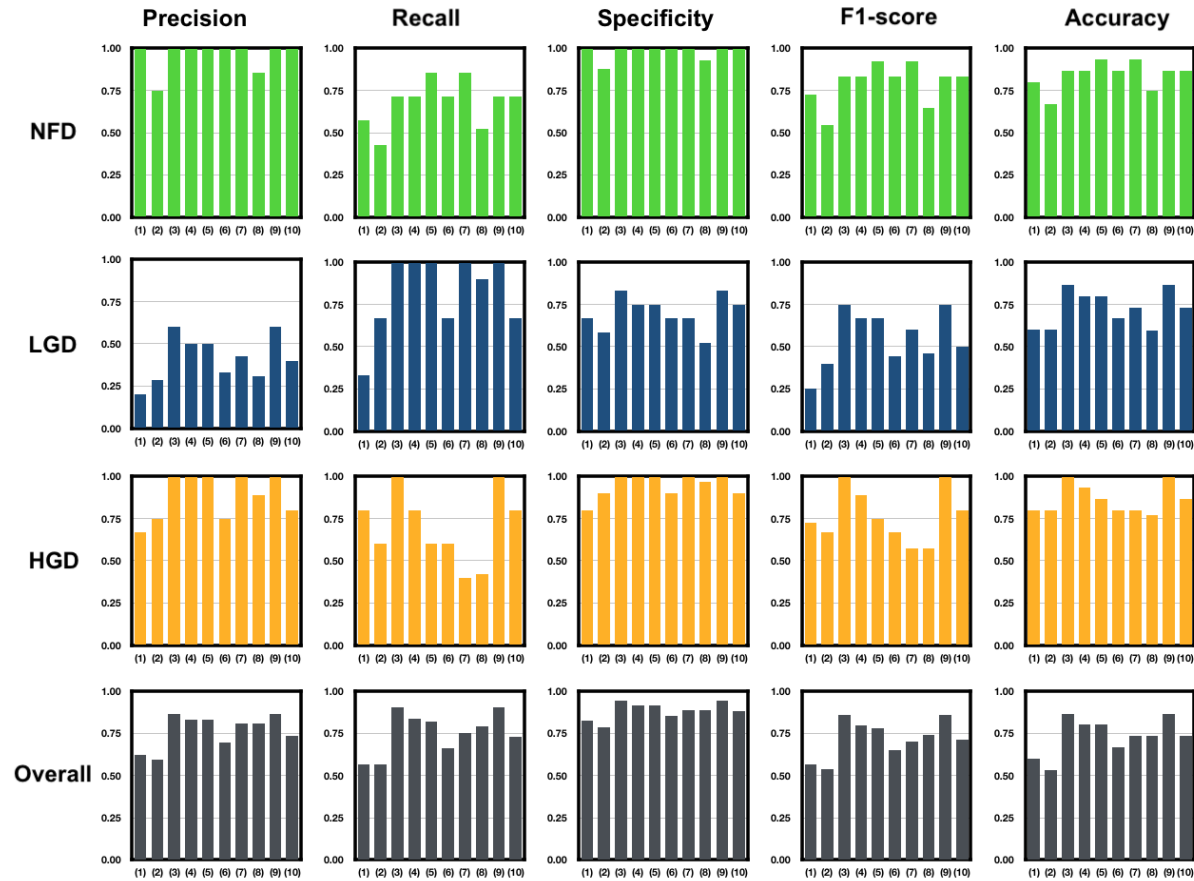
**Figure 6.3** Confusion matrices for different models at the annotation and the slide levels

(1)10X, (2)10X+ROI, (3) 40X, (4) 40X+AD, (5) 10X+40X, (6) 10X+40X+AD, (7) 10X+40X+ROI, (8) first approach, (9) second approach and (10) (Liu et al., 2017) will be discussed in section 0. Symbols of different architectures are clarified in Table 6.3.



**Figure 6.4** Bar charts for the performance measurements for each grade and the overall grades at the annotation-level for different models

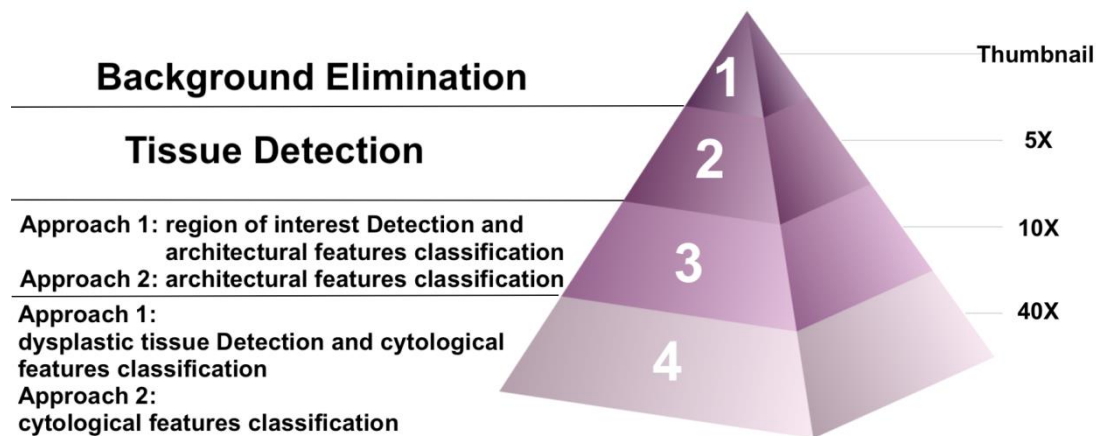
(1)10X, (2)10X+ROI, (3) 40X, (4) 40X+AD, (5) 10X+40X, (6) 10X+40X+AD, (7) 10X+40X+ROI, (8) first approach, (9) second approach and (10) (Liu et al., 2017) will be discussed in section 0. Symbols of different architectures are clarified in Table 6.3.



**Figure 6.5** Bar charts for the performance measurements for each grade and the overall grades at the slide-level for different models (1)10X, (2)10X+ROI, (3) 40X, (4) 40X+AD, (5) 10X+40X, (6) 10X+40X+AD, (7) 10X+40X+ROI, (8) first approach, (9) second approach and (10) (Liu et al., 2017) will be discussed in section 0. Symbols of different architectures are clarified in Table 6.3.

The two approaches were assembled from the trained models in Chapter 4 and Chapter 5 and tested on the annotations of the test set to decide which one is the best for the research CAD system. Both approaches follow a pyramidal architecture, as illustrated in Figure 6.6. The top of the pyramid takes WSIs at the lowest available magnification (0.6, 1.25, or 2.5) to detect the foreground and reduce noise, as discussed in section 3.5. Boxes bounded them after detecting tissues with a high probability of being the actual biopsies. More pixels are padded to make the bounding boxes divisible by 256 at 10X and 128 at 5X. Then, the bounding boxes are divided into non-overlapped tiles of 256x256x3 size. The two suggested approaches take these sampled patches as input to be analysed, as discussed in the two sections.

Finally, the two approaches were tested, and their performances were compared at the annotation-level using the proposed annotation-level inference system discussed in section 4.3.2. The selected CAD system, which improved performance, was tested at the slide-level using the same inference technique; however, it was trained on the detected tissues in the WSIs instead of the annotations. The used dataset for the slide-level inference is provided in section 6.4.



**Figure 6.6** The general pyramidal architecture of the first and second approaches

### **6.2.1 A potential CAD system based on the consensus grading for the analysis at 10X and 40X magnifications**

The proposed consensus system assumes that analysing WSIs at 10X magnification will make the model capture the architectural arrangements such as glandular distortion and crowding. In contrast, the analysis at 40X magnification will capture the cytological features like nuclei shape and size, as is discussed in the sampling process of the patches at different magnifications (refer to section 3.7). Through the sampling process, we observed that patches at 10X contain gland arrangements and the budding, the breaching, the crowding and the shape of crypts. In contrast, we observed that patches at 40X contain nuclei polarity, thickness, stratification, hyperchromatic, and nuclear number. Thus, the proposed consensus system relied on our observation and was not scientifically justified. This limitation will be discussed later, and a suggested future work (See Chapter 7).

For simplicity, this approach will be referred to as “the first approach” in the text, figures and tables. Figure 6.2 provides an overview of the first approach. It takes all the tiled biopsies, including all the layers from the oesophagus, using the proposed model in Chapter 4; it detects the region of interest (Barrett's tissue and not normal tissue), which are supposed to include regions similar to the annotated regions by the pathologists. Once a tile is marked as a region of interest, then its cached feature representation is fed to the subnetwork that classifies the dysplastic changes in this region based on the analysis at 10X magnification, which is assumed to be spatial arrangements of the glands and the nuclei and the amount of lamina propria between glands, as it was discussed in section 3.7 and visualised in Figure 3.6, Figure 3.7 and Figure 3.9. After that, a module gathers information about the detected tiles and their 10X level classification and generates a prediction map for the input WSI.

For further analysis, each detected region of interest at 10X magnification was divided into four tiles of 256x256x3 pixels at 40X magnification and fed into the proposed model in Chapter 5 to detect regions suspected of having dysplastic changes using the anomaly detector module. Then, their dysplastic changes are classified based on the extracted features at 40X magnification, which are assumed to be cytological features within the input



tiles. After using the earlier mentioned module, a prediction heatmap for the WSI is obtained at the 40X magnification.

A logical architecture was designed to follow the pathology criteria in grading Barrett's related dysplasia to find the consensus grading for tissues in WSI between the suggested prediction heatmaps at the two magnifications. In the logical system design, we assumed NFD10, LGD10 and HGD10 represent pixels predicted as NFD, LGD and HGD at 10X magnification, and NFD40, LGD40 and HGD40 represent them as NFD, LGD and HGD at 40X magnification. The following criteria are discussed later, and each logical expression and its corresponding pathology guidelines are summarised in Table 6.1.

**Table 6.1** Some of the inspiring pathology guidelines in designing the consensus system and their correspondence with a logical expression

Logical expression	Pathology Guidelines	Pathology Diagnosis
$(NFD10 \wedge NFD40)$	Architectural (at 10X) and cytological (at 40X) features are preserved at the NFD level.	NFD
$(NFD10 \wedge LGD40)$	Slight abnormality changes usually are observed in LGD cytological changes (at 40X) and are accepted as long as the architectural features (at 10X) are preserved as NFD architectural changes.	NFD
$(LGD10 \wedge NFD40)$	Slight changes for the architectural features (at 10X), categorised as LGD changes, might be witnessed with preserved cytological features at 40X and NFD levels.	NFD
$(LGD10 \wedge LGD40)$	The architectural abnormalities (at 10X) and cytological (at 40X) have LGD changes.	LGD
$(NFD10 \wedge HGD40)$	The architectural features (at 10X) for tissue are retained, as NFD changes, while there are massive abnormalities on	LGD

	the cytological level (at 40X) that are categorised as HGD cytological changes	
$(LGD10 \wedge HGD40)$	The architectural features (at 10X) are relatively retained (LGD), while there are massive abnormalities (HGD) at the cytological level (at 40X)	LGD
$(HGD10 \wedge NFD40)$	The tissue shows a high degree of architectural change (HGD changes at 10X), while the cytological features are retained (NFD features at 40X)	LGD
$(HGD10 \wedge HGD40)$	The architectural (at 10X) and cytological (at 40X) abnormalities, both at the HGD level	HGD
$(HGD10 \wedge LGD40)$	The cytological features (at 40X) for tissue are graded as LGD, and the architectural-level abnormalities (at 10X) for the tissue have HGD features	HGD

In pathology, NFD grade includes cases where architectural and cytological features are preserved. However, slight abnormality changes to some degree to the cytological features are accepted as long as the architectural features are preserved. In some cases where the epithelium is curing (generating epithelium), slight changes in the architectural features might be witnessed with preserved cytological features. Thus, the consensus pixel is set to NFD when the propositional logic  $(NFD10 \wedge NFD40) \vee (NFD10 \wedge LGD40) \vee (LGD10 \wedge NFD40)$  is true.

Suppose the architectural features for tissue are retained while there are massive abnormalities at the cytological level  $(NFD10 \wedge HGD40)$ . In that case, the tissue is graded as LGD because HGD architectural changes always accompany HGD tissue. This case is also applicable when the architectural features are relatively retained (a low degree of abnormalities)  $(LGD10 \wedge HGD40)$ . Also, when the tissue shows a high degree of architectural change while the cytological features are retained  $(HGD10 \wedge NFD40)$ , it cannot be graded as NFD as a high degree of architectural abnormalities is not accepted in this grade. Nevertheless, it can not be graded as HGD, as the

presence of cytological abnormalities is necessary for this grade. Thus, it is considered LGD for the pathology criteria refer to section 2.1.6, Table 2.2 and Table 2.3. A tissue is graded as LGD when the propositional logic  $(NFD10 \wedge HGD40) \vee (LGD10 \wedge LGD40) \vee (LGD10 \wedge HGD40) \vee (HGD10 \wedge NFD40)$  is true.

Finally, as discussed in Chapter 2, when the cytological features of tissue are graded as LGD, and the architectural-level abnormalities for tissue have HGD features, the tissue is diagnosed as HGD. Therefore, in the suggested consensus grading system, the tissue is HGD when  $(HGD10 \wedge LGD40) \vee (HGD10 \wedge HGD40)$  is true. The prediction map generator gathers the consensus grades and produces a final prediction map for the WSI. Table 6.2 shows the logical table for the consensus grading system, and Figure 6.2 shows the CAD system following this approach.

**Table 6.2** The propositional logic table for the proposed consensus grading system

Pixels at 10X			Pixels at 40X			The consensus grading for the output pixels		
NFD	LGD	HGD	NFD	LGD	HGD	NFD	LGD	HGD
NFD10	LGD10	HGD10	NFD40	LGD40	HGD40	A	B	C
1	0	0	1	0	0	1	0	0
1	0	0	0	1	0	1	0	0
1	0	0	0	0	1	0	1	0
0	1	0	1	0	0	1	0	0
0	1	0	0	1	0	0	1	0
0	1	0	0	0	1	0	1	0
0	0	1	1	0	0	0	1	0
0	0	1	0	1	0	0	0	1
0	0	1	0	0	1	0	0	1

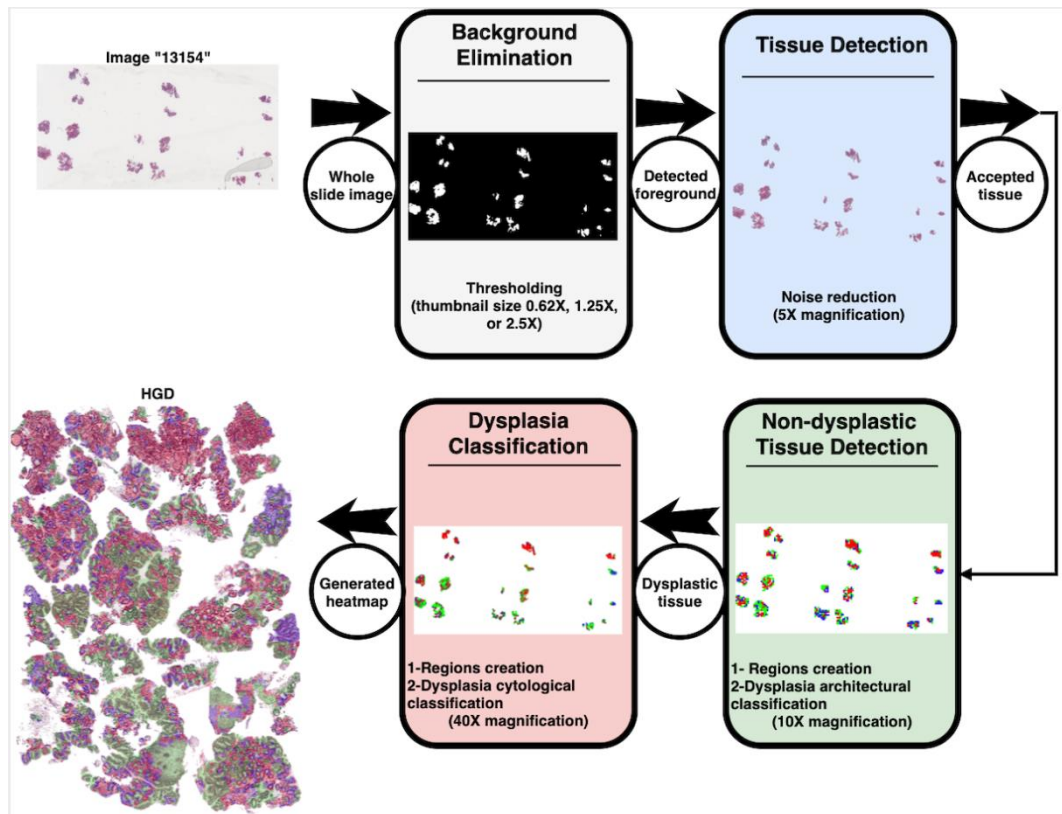
$$A = (NFD10 \wedge NFD40) \vee (NFD10 \wedge LGD40) \vee (LGD10 \wedge NFD40),$$

$$B = (NFD10 \wedge HGD40) \vee (NFD10 \wedge LGD40) \vee (LGD10 \wedge HGD40) \vee (HGD10 \wedge NFD40)$$

$$\text{and } C = (HGD10 \wedge LGD40) \vee (HGD10 \wedge HGD40).$$

## 6.2.2 The proposed CAD system

Figure 6.7 illustrates the CAD system following this approach that receives the sampled patches at 10X as input and then feeds them into the trained model in Chapter 4 to analyse WSIs and classify dysplasia at 10X magnification (discussed in section 4.6.2 without using the region of interest detector), to grade the patches as NFD, LGD and HGD considering their observed changes at 10X magnification only. The model produces three values that represent the probabilities that it belongs to each grade. These probabilities are transformed into a four-label map. Every pixel within the patch is coloured green, blue and red if the NFD, LGD and HGD probabilities are higher or equal to 50%; otherwise, its pixels are coloured white, indicating the model is uncertain about the grade. Regions were detected as dysplasia, either LGD or HGD, and uncertain regions are analysed at 40X magnification in the next phase; thus, the patches sampler module takes the generated prediction maps from the previous step and divides each LGD, HGD, and uncertain region tile into a further 16 tiles with 256x256x3 pixels at 40X magnification. The sampled patches are then fed into a higher magnification-based classification network implemented and trained in section 5.3.2. Finally, the prediction maps generator collects the NFD information at the 10X magnification analyser and LGD and HGD information at 40X magnification and generates the whole virtual slide prediction map using only probabilities more than or equal to 50%. A consensus grade among the two classification networks is designed to confirm tissues as NFD at 10X magnifications and LGD and HGD at 40X magnification.



**Figure 6.7** The proposed CAD system architecture following the second approach

The "Non-dysplastic Tissue Detection" is the fine-tuned Inception-ResNet-v2 using the sampled patches at 10X magnification (see section 4.6.2)., and the "Dysplasia Classification" module takes the generated heatmap from the "Non-dysplastic Tissue Detection", samples patches from the dysplastic tissue, then classifies them using the model in section 5.3.2. The final generated heatmap contains the NFD stain from the "Non-dysplastic Tissue Detection" module and the LGD and HGD stains from the "Dysplasia Classification" module.

### 6.3 Empirical evaluation and results comparison

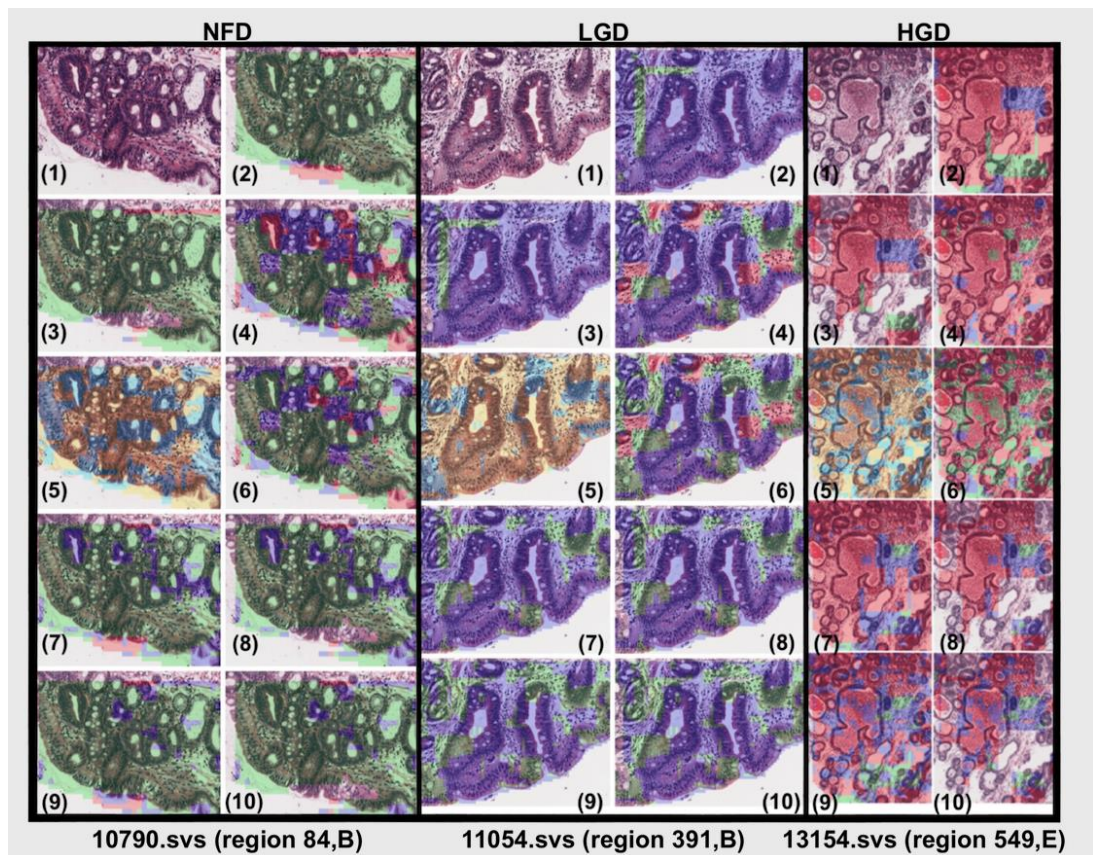
This section will discuss the conducted experiments to decide the final architecture of the CAD system for Barrett's related dysplasia diagnosis. Balancing two essential factors, high performance and low computational cost, are essential in selecting the best architecture. Therefore, different compositions of the trained networks in Chapter 4 and Chapter 5 were tested. All the experiments were conducted using the provided annotations for the test whole slides. The architecture compositions for the tested approaches are provided in Table 6.3.

**Table 6.3** The architecture compositions for the tested approaches in the empirical evaluation.

Approach symbol	Description
10X	The dysplasia classification is based on the <b>analysis at 10X magnification</b> (section 4.6.2)
10X+ROI	The dysplasia classification is based on the <b>analysis at 10X magnification</b> and the <b>region of interest detection</b> (section 4.6.3)
40X	The dysplasia classification is based on the <b>analysis at 40X magnification</b> (section 5.3.2)
40X+AD	The dysplasia classification is based on the <b>analysis at 40X magnification</b> combined with <b>potential dysplastic tissue detection</b> (section 5.4.4)
10X+40X	The <b>consensus</b> grading system between the <b>analysis at 10X magnification</b> and the <b>analysis at 40X magnification</b> based grading (section 6.2.1)
10X+40X+AD	The <b>consensus</b> grading system between <b>10X</b> and <b>40X+AD</b>
10X+40X+ROI	The <b>consensus</b> grading system between <b>10X+ROI</b> and <b>40X</b>
First approach	The <b>consensus</b> grading system between <b>10X+ROI</b> and <b>40X+AD</b> (section 6.2.1)
Second approach	The proposed <b>CAD system</b> uses 10X for dysplasia detection and 40X to classify the detected regions (section 6.2.2)

Figure 6.4 and Figure 6.5 show bar charts to compare the performance metrics for the different suggested architectures at the annotation-level and the slide-level, respectively. At both levels, the 10X + ROI approach slightly decreased the performance of the 10X approach, and both of them are not sufficient and as accurate as the model performing at 40X magnification. Similar to these are the results of the 40X and 40X + AD, as adding the anomaly detection module slightly declined the performance for the model classifying HGD at the annotation-level and the overall performance at the slide-level. The consensus grading system (10X + 40X) performed better than 10X and 10X + ROI. Also, by adding the region of interest module to the consensus grading system, it outperformed it. Applying the consensus rules between the results of 10X and 40X + AD instead of 40X had better results in grading the dysplastic annotations, in contrast, lower results in grading NFD annotations. In general, adding the region of interest module

enhanced the performance of  $10X + 40X + AD$  based on the provided annotations results. From the previous comparisons, it was found that amongst all the approaches, except the first and second proposed approaches, approach  $40X$  is the best performing architecture followed by  $40X + AD$  approach. Further comparison, between the architectures performances for each grade, was taken into consideration. Also, Figure 6.8 provides prediction maps for three annotations belonging to NFD, LGD and HGD using different architectures.



**Figure 6.8** NFD, LGD and HGD annotations with their grades using different architectures

Green, blue and red colours indicate NFD, LGD and HGD, respectively, in all images except (5) marked images where light blue indicates non-dysplastic tissue and orange indicates potential dysplastic tissue. (1) the provided annotations, (2) are predictions using  $10X$  architecture, (3) are predictions using  $10X+ROI$  architecture, (4) are predictions using  $40X$  architecture, (5) are potential dysplastic tissue detections, (6) are predictions using  $40X+AD$  architecture, (7) are predictions using  $10X+40X$  architecture, (8) are predictions using  $10X+40X+ROI$  architecture, (9) are predictions using  $10X+40X+AD$  architecture, (10) are predictions using the first approach architecture.



This paragraph will focus on  $10X + 40X$  and  $10X + 40X + AD$  approaches. At the annotation-level, the  $10X + 40X + AD$  approach slightly increased the performance for HGD and slightly decreased it for NFD while keeping the approach's performance at the same level for LGD. Clinically, remedial actions are taken for HGD patients, while patients with lower grades are put on surveillance schedules. Therefore, this research focuses on models that best detect and classify HGD. As a result,  $10X+40X+AD$  was nominated as the proposed CAD system. One major drawback of this model is the high cost of running it on the whole slide. All the regions within a tissue will be analysed at both magnifications; thus, the region of interest detection was added to form the first proposed approach for the CAD system.

The final step is to compare the first and second proposed approaches as an intermediate evaluation. The two proposed approaches were tested using the provided annotations from the test set. At the annotation-level, the first approach correctly classifies 12, 9 and 8 annotations belonging to NFD, LGD and HGD, respectively. Whereas the second approach correctly classifies 12 NFD, 6 LGD and 16 HGD annotations. From Figure 6.3, Figure 6.4, Figure 6.5, and Table 6.4, it can be seen that the second approach performed better for dysplasia and is similar to the first approach for NFD.

Additionally, it has a higher weighted KV, indicating the seriousness of the disagreement with the first approach (see section 2.3). In other words, the second approach has a higher weighted KV because whenever it misclassifies an annotation, it confuses its grade with an adjacent grade, as in Barrett's related dysplasia, the definitions of the boundaries between grades are absent. For instance, when it misclassified three annotations belonging to HGD, the second approach downgraded them to one grade (to LGD). The first approach downgraded one HGD annotation to two grades (to NFD) and ten HGD annotations to LGD.

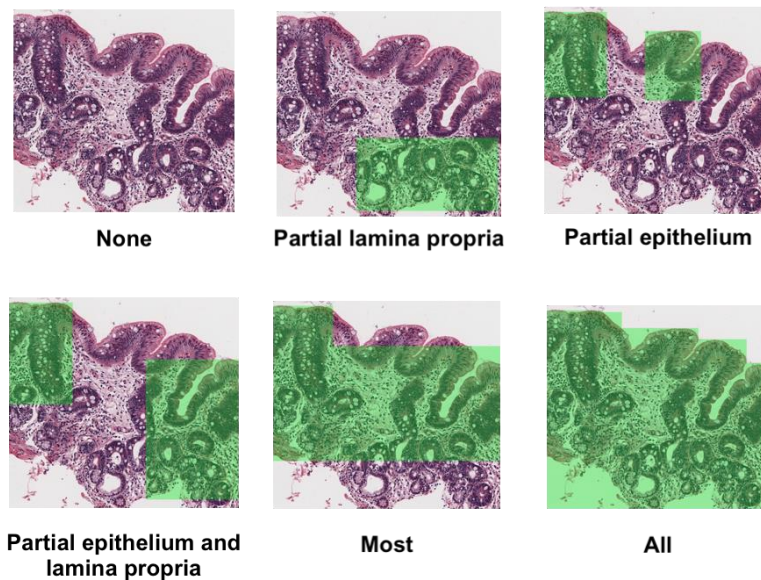


**Table 6.4** The interobserver agreements for different approaches at the annotation and slide levels

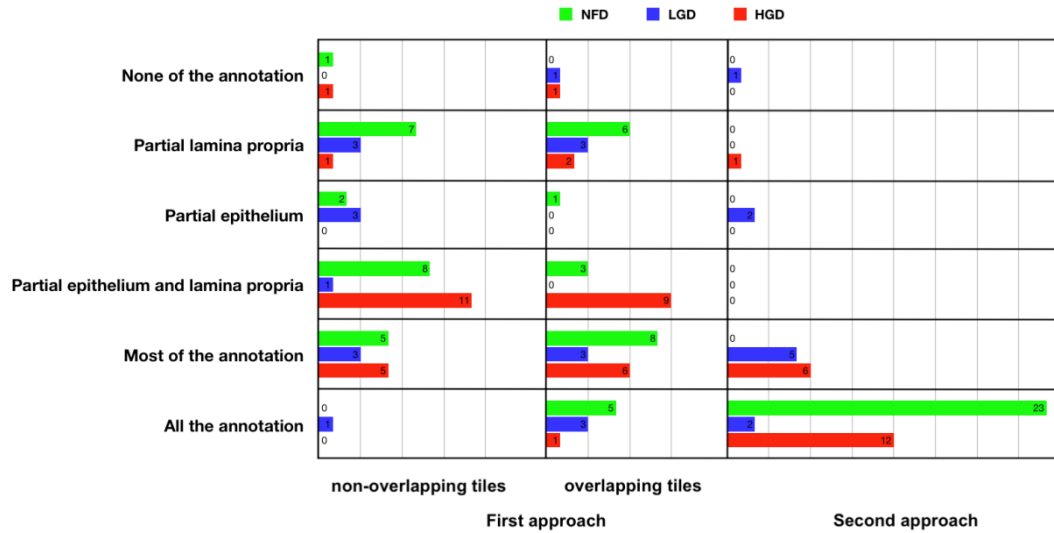
	Annotation-level		Slide-level	
	KV	Weighted KV	KV	Weighted KV
10X	0.300	0.416	0.408	0.579
10X+ROI	0.269	0.370	0.327	0.412
40X	0.471	0.601	<b>0.800</b>	<b>0.857</b>
40X+AD	0.476	0.597	0.704	0.780
10X+40X	0.341	0.460	0.700	0.776
10X+40X+AD	0.323	0.442	0.507	0.634
10X+40X+ROI	0.391	0.481	0.605	0.694
First approach	0.377	0.457	0.615	0.706
Second approach	<b>0.490</b>	<b>0.627</b>	<b>0.800</b>	<b>0.857</b>
(Liu et al., 2017)	0.393	0.512	0.600	0.714

Moreover, the region of interest detector for both systems, which is the region of interest module to discriminate Barrett's tissue from normal tissue in the first approach and the module to analyse dysplasia at 10X magnification in the second approach, was used as a discriminator for dysplastic and non-dysplastic tissue, were tested on the WSIs to evaluate the approaches abilities to capture dysplastic regions. The measurement for this experiment is the approach's ability to detect "parts of", "most of", or "all" of the given annotations. Also, partial detection of an annotation is further categorised as "partial lamina propria detection", "partial epithelium detection", and "partial epithelium and lamina propria detection". Following that order, the performance of the region of interest is increased. Figure 6.9 gives examples of the measurement categories. In addition, Figure 6.10 compares the results of the region of interest modules from the first approach and the second approach. In the first approach, two techniques were used to divide the whole slide into tiles at 10X magnification, using the non-overlapped and overlapping techniques. The overlapped technique (by 64 pixels) increased the number of the sampled patches 16 times more than the number of sampled patches following the non-overlapped technique. That is, the same number of sampling non-overlapping tiles at 40X for the

whole slide; thus, the overlapped technique is not applicable. However, it was used to study the effect of overlapping in the first approach. The bar charts in Figure 6.10 show that 9 and 13 dysplastic annotations were mostly or entirely detected using the non-overlapping tiles and overlapping tiles from the first approach, respectively, while 25 dysplastic annotations were mostly or entirely detected in the second approach. From the bar charts, the second approach leaves one LGD annotation undetected. It is essential to clarify that the region (Figure 4.12) was diagnosed differently by the two experts, as “Expert\_B” graded it as NFD while “Expert\_E” graded it as LGD and the proposed CAD system agrees with “Expert\_B”. In addition, the HGD annotation, in which the second approach detected that some of its regions belong to the lamina propria, is a tiny region annotated from the lamina propria layer only (see Figure 4.13 (c)).



**Figure 6.9** Examples for the overall region of interest detection measurement categories for annotations



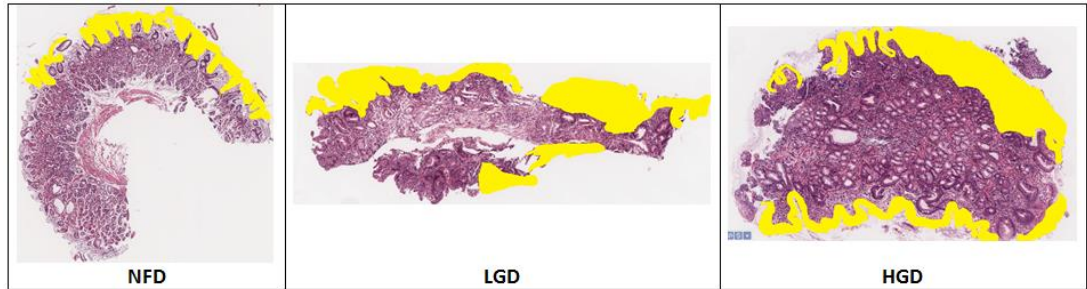
**Figure 6.10** The results comparison for the region of interest’s module for the first and second approaches

The results show that the overlapping affected the detection process positively. Also, they show that the second approach enhanced dysplastic detection compared to the first approach. Based on those observations, the second approach is selected as the CAD system architecture for diagnosing Barrett’s oesophagus related dysplasia.

## 6.4 Datasets

In this chapter, all the deep-learning models were trained in Chapter 4 and Chapter 5. The used datasets are the annotated dataset (refer to section 4.5) employed to train the annotation-level inference system for the different approaches. The annotated dataset will be used to train the slide-level inference system for the second CAD approach. For evaluating, the proposed CAD system, all tissues within the WSIs were analysed. As a result, the annotation-level inference system does not appear adequate to determine the slide-level grade. Thus, a similar module is trained in the same manner as training the submodule in sections 4.3.2 and 4.6.2, instead, using a dataset that the researcher manually annotated to separate the epithelial layer from the lamina propria layer for each tissue in the WSIs that contain any annotation in the train set and all the tissues in the test set. Then, label the tissues with the highest grade of any annotation. The annotation process was performed on the training and testing WSIs, and the

labelling is for the training set only. From the 128 training WSIs, 329 tissues were annotated. 201, 50, and 78 tissues belong to NFD, LGD and HGD, respectively, and from the test WSIs, 191 tissues were annotated. Figure 6.11 provides examples of the annotated tissues from each grade.



**Figure 6.11** Samples of the epithelial layer mask dataset for the tissues

## 6.5 Results of the selected CAD system

After comparing the results of the two approaches on the test annotations only, the approach with the more enhanced performance was applied to all the regions within the WSIs from the test set. The following three sections provide the result of each module in the selected CAD system (the second approach) following the slide-level inference system discussed in the previous section.

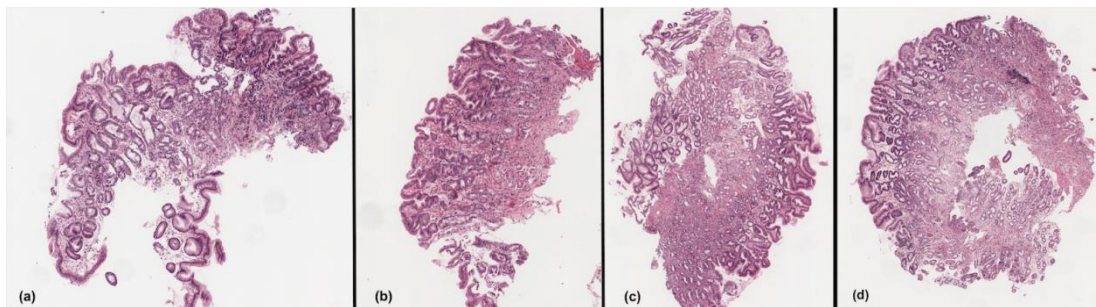
### 6.5.1 Background elimination and tissue detection

This module focuses on detecting tissues that are more likely to be identified as biopsies. It accepted an amount of tissue in each slide equal to the number of biopsies or more. The accepted additional tissues are detached tissues from the biopsies. Table 6.6 shows the actual number of biopsies in each test slide, the number of the detected foreground boxes, and finally, the number of accepted tissue boxes after applying the filtration rules. The table shows that the module successfully detected the actual biopsies and neglected the noise and any small detached tissue.

### 6.5.2 NFD classification and dysplasia detection (10X)

Using 10X magnification to detect NFD regions successfully classified all tissues in 5 NFD test slides and 6 and 3 tissues in two NFD slides as LGD

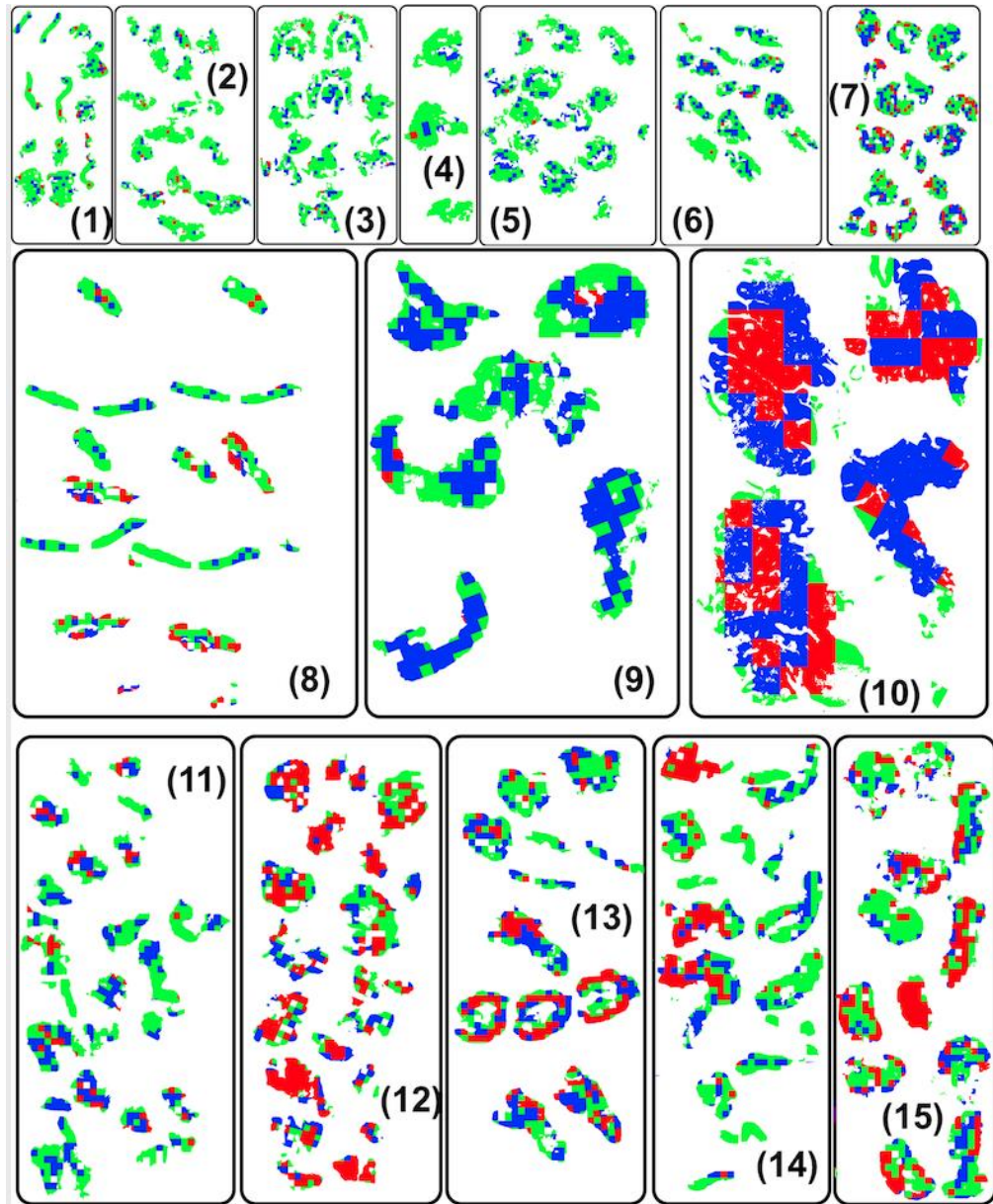
and HGD, respectively. All the detected regions as, LGD and HGD, belong to the lamina propria. None belong to the epithelial layer, except for slides "11035" and "11040", which were misclassified as LGD and HGD, respectively. Three biopsies in the misclassified NFD slides were predicted as HGD (see Figure 6.12 (a) and (b) for examples), and six biopsies as LGD (for a sample, refer to Figure 6.12 (c)). The other detected dysplastic regions, alongside the three biopsies, belong to the lamina propria. The misclassification for those regions is attributed to the glandular disarrangement and the unique shape of the crypts in this slide, as shown in Figure 6.12 (d). This unique shape was not observed in any of the training sets. As shown in Table 6.6, all HGD slides were predicted correctly, whereas eight biopsies from two LGD slides were misclassified as dysplastic slides. Generally, the CAD system successfully detected dysplastic biopsies at that level except for one NFD slide. Predictions maps for all the test slides are provided in Figure 6.13. the slide-level confusion matrices for the prediction for this submodel using the three-tier and two-tier classifications are shown in Figure 6.14.



**Figure 6.12** Samples of the misclassified biopsies in slide "11040.svs" using the analysis at 10X magnification

(a), (b), (c), and (d) are visualised at 5X magnification.





**Figure 6.13** Prediction maps for the test set slides produced by the analysis at 10X magnification submodel

Green, blue and red colours indicate NFD, LGD and HGD, respectively. The slides are “10586.svs”, “10790.svs”, “10829.svs”, “10857.svs”, “11014.svs”, “11035.svs”, “11040.svs”, “11013.svs”, “11054.svs”, “11063.svs”, “13083.svs”, “13154.svs”, “13239.svs”, “13303.svs” and “13348.svs”, which are presented in the ascending order.

	NFD	LGD	HGD
NFD	5	1	1
LGD	0	1	2
HGD	0	0	5

	NFD	PFD
NFD	5	2
PFD	0	8

**Figure 6.14** The slide-level confusion matrices for the NFD classification and dysplasia detection submodel

(a) follows the three-tier classification and (b) follows the two-tier classification.

### 6.5.3 Detected dysplastic tissue classification (40X)

Further higher resolution sampling for the detected dysplastic regions to predict their dysplastic degree enhanced the overall prediction for the WSIs. It downgraded the 8 misclassified LGD biopsies and correctly graded all the dysplastic slides. However, the system still incorrectly grades the NFD slides ("11035.svs" and "11040") as LGD. Table 6.6 shows the prediction after the 40X magnification to form the CAD system in the diagnosing. Also, Table 6.5 shows the overall grade for each test slide against pathologists' diagnosis for the glass slides and the virtual slides.

**Table 6.5** List of grades for the test set provided by two pathologists for the virtual and glass slides and their associated diagnosis by the proposed CAD system

	Virtual slide				Glass slide	
	"Expert_B"	"Expert_E"	(Adam, 2015)	The proposed CAD system	"Expert_B"	"Expert_E"
10586	NFD (g1)	NFD (g1)	LGD	NFD	NFD (g1)	NFD (g1)
10790	NFD (g2)	NFD (g1)	NFD	NFD	LGD (g4)	LGD (g3)
10829	NFD (g1)	NFD (g1)	NFD	NFD	NFD (g1)	NFD (g1)
11014	NFD (g1)	NFD (g1)	NFD	NFD	NFD (g2)	NFD (g1)
11035	NFD (g1)	NFD (g1)	NFD	LGD	NFD (g1)	NFD (g1)
11040	NFD (g1)	NFD (g1)	NFD	LGD	NFD (g1)	NFD (g1)
10857	NFD (g1)	LGD (g3)	LGD	NFD	LGD (g4)	LGD (g3)
11013	LGD (g4)	LGD (g4)	LGD	LGD	LGD (g4)	LGD (g4)
11054	LGD (g3)	NFD (g1)	LGD	LGD	LGD (g4)	LGD (g3)
11063	LGD (g3)	NFD (g2)	HGD	LGD	LGD (g4)	LGD (g3)
13083	HGD (g5)	HGD (g6)	HGD	HGD	HGD (g5)	HGD (g6)
13154	HGD (g5)	HGD (g6)	HGD	HGD	HGD (g5)	HGD (g6)
13239	HGD (g6)	HGD (g6)	HGD	HGD	HGD (g6)	HGD (g6)
13303	HGD (g5)	HGD (g6)	HGD	HGD	HGD (g5)	HGD (g6)
13348	HGD (g5)	HGD (g6)	HGD	HGD	HGD (g5)	HGD (g6)

Cells highlighted in red are considered an error, and yellow highlighted cells indicate no consensus among pathologists in diagnosing the virtual slides.



**Table 6.6** Results for different modules in the proposed CAD system

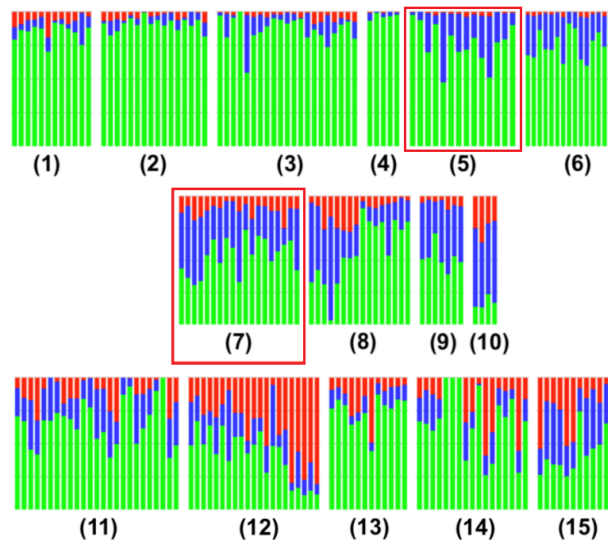
Slide ID		Number of			Number of detected (NFD, LGD, HGD) biopsies		The CAD Process time
		Real biopsies	Detected foreground	Accepted tissue	At 10X magnification	CAD system	hours:mins:secs
10586	NFD	12	39	12	12,0,0	12,0,0	00:08:22
10790	NFD	12	89	16	12,0,0	12,0,0	00:09:17
10829	NFD	12	156	21	12,0,0	12,0,0	00:10:24
10857	NFD	5	13	5	5,0,0	5,0,0	00:06:02
11035	NFD	12	58	14	9,3,0	10,2,0	00:13:40
11014	NFD	15	168	16	15,0,0	15,0,0	00:18:39
11040	NFD	15	41	19	9,3,3	10,5,0	00:17:36
11013	LGD	16	31	16	8,2,6	8,8,0	00:11:54
11054	LGD	6	45	7	1,5,0	1,5,0	00:38:50
11063	LGD	4	185	4	0,2,2	0,4,0	00:39:07
13083	HGD	23	71	25	5,10,8	11,6,6	00:15:45
13154	HGD	18	62	20	1,5,12	0,5,13	00:46:08

13239	HGD	12	25	12	3,5,4	3,5,4	00:30:34
13303	HGD	17	46	17	10,4,3	13,1,3	00:36:40
13348	HGD	12	76	12	2,4,6	9,0,3	00:41:18
Total		191	191	1105	111,33,47	125,40,26	05:44:16

Numbers in bold are the number of predicted tissues that led to misclassifying their correspondence WSIs.

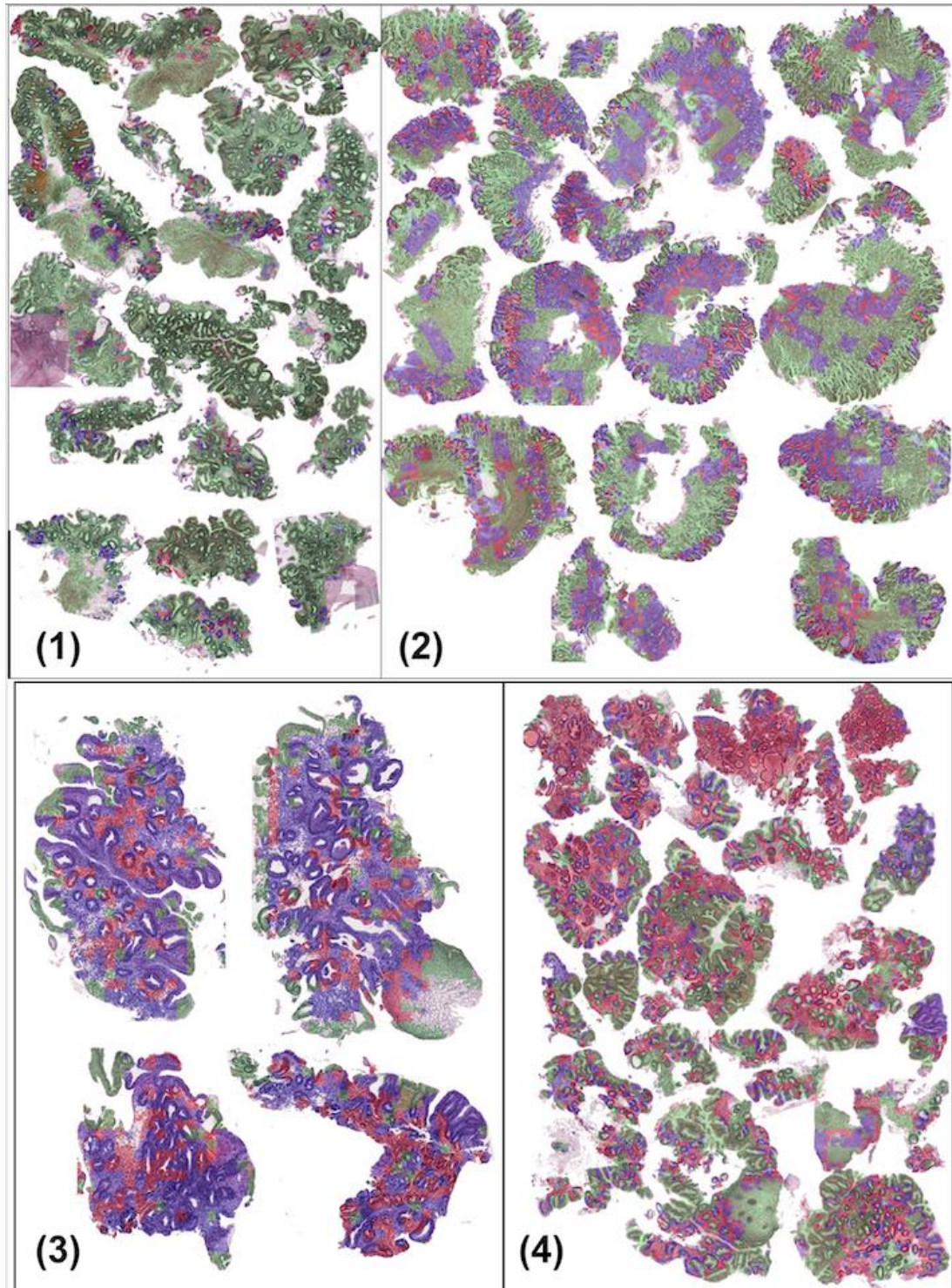
Figure 6.15 shows the distribution of each grade within each slide. Each set represents a virtual slide, and each column in the set is a biopsy within the slide. Green, blue and red colours, representing the percentage of NFD, LGD and HGD pixels within the detected biopsy, could be present in each column unless the colour is assigned a 0% value. It shows that the diagnosed NFD slides, the first row in the figure except Figure 6.15 (5), have small percentages of dysplasia. As mentioned before, these dysplastic tissues were detected in the lamina propria layer. The second row and Figure 6.15 (5) shows the diagnosed LGD slides, including the misclassified slides ("11035.svs" and "11040.svs") are represented by sets (5) and (7) in Figure 6.15; the sets have at least one biopsy with the higher percentage being for LGD and none being for HGD.

Similarly to that, the classified HGD slides are in the third row. Although the stacked bar charts visualise the results and give a good understanding of the different grade distributions, they were not relied on in the whole virtual slides grading decisions. Prediction maps for three correctly predicted slides and a misclassified slide are visualised at 5X magnification in Figure 6.16 (the prediction maps for the other test slides are provided in Appendix C.1).



**Figure 6.15** The distribution of each dysplasia grade within each test slide using the proposed CAD system

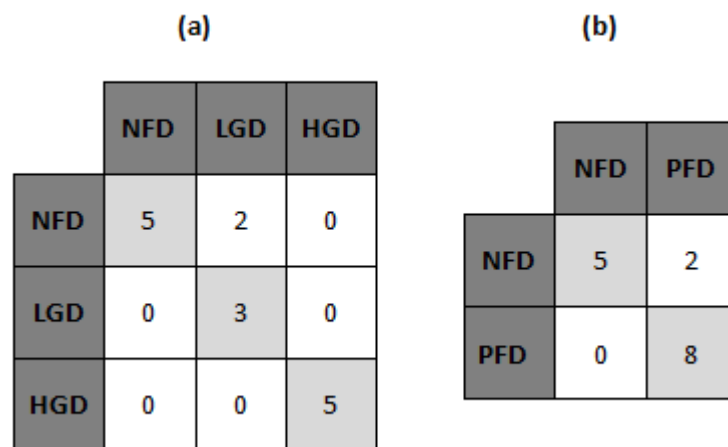
Green, blue and red colours indicate NFD, LGD and HGD, respectively. The slides are (1)“10586.svs”, (2)“10790.svs”, (3)“10829.svs”, (4)“10857.svs”, (5)“11035.svs”, (6)“11014.svs”, (7)“11040.svs”, (8)“11013.svs”, (9)“11054.svs”, (10)“11063.svs”, (11)“13083.svs”, (12)“13154.svs”, (13)“13239.svs”, (14)“13303.svs” and (15)“13348.svs”, which are presented in the ascending order.



**Figure 6.16** Visualised prediction maps for some test slides using the proposed CAD system (the second approach)

(1) slide “10790.svs”, NFD, which is correctly predicted, (2) slide “11040.svs”, NFD which is predicted as LGD, and the correctly predicted, (3) slide “11063.svs” as LGD and (4) slide “13154.svs” as HGD.

Generally, the CAD system has confusion matrices at the three-tier and two-tier classifications, as shown in Figure 6.17. It scored 0.87 precision, 0.90 recall, 0.94 specificity, 0.86 F1-score, 87% accuracy and 0.8 and 0.5 agreements with “Expert\_B” and “Expert\_E”, respectively.



**Figure 6.17** The slide-level confusion matrices for the CAD system

(a) follows the three-tier classification and (b) follows the two-tier classification

#### 6.5.4 Computational time

The computation time for analysing each slide in the test set is provided in Table 6.6. Overall, the CAD system took 5 hours to analyse all the test sets. The size of the slide and the number of biopsies within each slide affect the processing time. The NFD slides took less time to be processed as they have a few regions that need to be analysed at 40X magnification. The average times for each NFD, LGD and HGD slide are 12, 30, and 34 minutes. The implementation is done in python and uses the Keras library for all the experiments. This work was undertaken on ARC3, part of the High-Performance Computing facilities at the University of Leeds in the UK. ARC3 offers different GPU nodes, and for this work, we used K80 GPU nodes and P100 GPU nodes, depending on the availability. The K80 GPU node has two K80 cards, 24 CPU cores and 128GB of system memory,

while the P100 GPU node has Four P100 cards, 24 CPU cores and 256GB of system memory.

## 6.6 Discussion

This section will discuss the proposed CAD system results and compare them against the outputs of domain-related research. It is worth stressing that the strength of evidence the comparison results provide is weak as the available dataset for this thesis is considered a small dataset. By starting from the first two stages in the CAD system, tissue detection, the system detected all the biopsies in the test slides. Then, the third stage successfully filtered all non-dysplastic slides except two slides, which is believed to be the failure attributed to merging the “indefinite for dysplasia” class with NFD. This merging might force the model to learn some confusing features that make the pathologists indefinite in their decisions. The final stage classified all detected dysplastic slides correctly except the failed slides in the previous stage. The proposed CAD system will be compared to two pieces of research. The first one is a domain-related work introduced by Adam (2015). It was an attempt to detect and grade dysplasia in Barrett's oesophagus. The other work is the relevant comparator approach, adopting a multi-scale analysis.

The proposed CAD system achieved 0.90 recall and 0.94 specificity scores, indicating that the system performs well in detecting the disease and avoiding false alarms. In comparison, the proposed CAD system by Adam (2015) performed well in avoiding false alarms yet had a lower performance in detecting the disease (0.84 recall and 0.94 specificity). Generally, the proposed CAD system scored higher performance metrics than (Adam, 2015), except for the agreement with “Expert\_E”. It is important to emphasise that for both CAD systems, the diagnosis for each slide is considered correct whenever it agrees with the diagnosis of any of the experts for that virtual slide. For instant, Table 6.5 shows that slide “10857.svs” was categorised as NFD by the proposed CAD system and was diagnosed with LGD by Adam’s approach and both diagnoses were considered correct as they agree with whether “Expert\_B” or “Expert\_E” diagnosis. In addition, it should be noted that the diagnosis of both the glass and virtual slides are valid and reliable, and the disagreement between a

pathologist's diagnoses is considered intraobserver disagreement. In this thesis, only the virtual slide diagnoses are considered in calculating the models' performances because the pathologists annotated the used annotations in training the models while examining the virtual slides.

Furthermore, the proposed CAD system is compared against the approach in the Liu et al. (2017) paper. They proposed frameworks to detect and localise tumours in breast cancer using the Camelyon16 dataset. One of the frameworks is a multi-scale approach to imitating pathologists examining tissue. The model comprises two supervised deep-learning networks (Inception) that work parallel. One accepts sampled patches from a region at 10X magnification, while the other accepts patches from the same region at 40X. Then, the fully connected layers from both networks are concatenated before being fed into the Softmax classifier to generate heat maps for the WSIs. Their approach was reimplemented and trained on this thesis dataset and tested on the test set (annotated regions for the train and test set) to compare the performance of the proposed multi-scale CAD system with their work. In the comparative study, the annotation-level inference system was employed to decide the grade of the annotations for both systems. Figure 6.3 provides confusion matrices at the annotation-level for their approach against the second approach CAD system following the three-tier classification. They achieved 58% accuracy, 0.66 precision, 0.60 recall, 0.81 specificity, and 0.58 F1-score against 65%, 0.68, 0.65, 0.84 and 0.64 metrics for the proposed CAD system. The two approaches performed the same in terms of two-tier classification, and the proposed CAD system outperformed the other in the three-tier classification. Even though both systems have close performances, the proposed CAD system has better runtime. Higher magnification is employed when the low magnification analysis only indicates a region as dysplasia. In conclusion, the proposed CAD system outperformed the two other CAD systems.

After comparing the proposed CAD system with others, a comparison of the agreement of the proposed CAD system against the agreement of pathologists is conducted as a final assessment. For this assessment, on the one hand, biopsies were selected from this research dataset for comparison. That includes all biopsies in the NFD slides and only those with annotations from dysplastic slides. The results of the proposed CAD system using the selected biopsies are summarised in the confusion matrix, as shown in Figure 6.18 (a). Also, the confusion matrixes for the biopsies labelled by



“Expert\_B” and “Expert\_E” are provided in Figure 6.18 (b) and Figure 6.18 (c), respectively. By considering biopsies that contain the annotated regions by the pathologists, the proposed CAD system managed to predict all the slides correctly except for slide "11040.svs", which is misclassified as LGD while it is NFD.

On the other hand, two studies were used in this comparison. The first study (Treanor et al., 2009) used the same dataset used in this research, yet not necessarily the same test subset. Forty-six biopsies were selected from Barrett's oesophagus and were graded, following the Vienna classification system, by two trainees (Trainee D and G) and two expert pathologists (“Expert\_B” and “Expert\_E”). Agreements between each trainee and each expert are provided in Table 6.7, besides the agreement between the proposed CAD system and each pathologist using their labelled biopsies. Eleven biopsies have annotations provided by both pathologists, and some of them are overlapped. Thus, the pathologists' agreement was calculated using those biopsies, as shown in the fifth row in Table 6.7. The proposed CAD system has a substantial agreement with “Expert\_B” and a fair agreement with the other pathologist. The CAD agreements are higher than the Trainees', except for "Trainee G" with "Expert\_E". Although the CAD system yielded the best agreement, this comparison suffers from two factors, the uneven sample size for the comparison parties and the usage of different categorisation systems. Hence, this study cannot be used as a final assessment, and the comparison is made to have a general overview of the performance. For the previously mentioned 11 biopsies, the agreements between the proposed CAD system and the pathologists are provided in the fourth row in Table 6.7, with an almost perfect agreement with “Expert\_B” and a moderate agreement with “Expert\_E”. In contrast, the agreement between the pathologists for those biopsies is moderate.



(a)				(b)				(c)			
	NFD	LGD	HGD		NFD	LGD	HGD		NFD	LGD	HGD
NFD	79	7	0	NFD	7	1	0	NFD	8	8	0
LGD	4	5	0	LGD	2	5	0	LGD	2	0	0
HGD	0	1	10	HGD	0	1	9	HGD	0	1	6

**Figure 6.18** The confusion matrices for the CAD system in the term of biopsies following three-tier classification

(a) The confusion matrix for the proposed CAD system for the available labelled biopsies, (b) The confusion matrix for the proposed CAD system for the provided labelled biopsies by "Expert\_B", and (c) The confusion matrix for the proposed CAD system for the provided labelled biopsies by "Expert\_E".

**Table 6.7** The interobserver agreements for trainee pathologists, expert pathologists and the proposed CAD system against the expert pathologists

	Expert_B KV (95% confidence interval)	Expert_E KV (95% confidence interval)
Trainee D (n=46) (6 categories)	0.17 (0.02-0.32)	0.27 (0.12-0.42)
Trainee G (n=46) (6 categories)	0.29 (0.11-0.47)	0.46 (0.28-0.65)
The proposed CAD (n=25) (3 categories)	0.758 (0.545-0.972)	0.321 (0.054-0.588)
The proposed CAD (n=11) (3 categories)	1 (1-1)	0.56 (0.275-0.845)
Expert_B/Expert_E (n=11) (3 categories)	0.484 (0.17-0.80)	

The second study was conducted by (Salomao et al., 2018). It focuses on the interobserver agreement between 3 pathologists in grading Barrett's oesophagus related dysplasia using four categories (NFD, indefinite for dysplasia, LGD and HGD) and three categories (NFD, indefinite for dysplasia and PFD) classifications. It also aims to find the effect of

diagnosing patients by examining all their biopsies. Five hundred forty-nine biopsies were used that belonged to 129 unique patients. Each pathologist examined each biopsy individually and recorded its grade. The pathologist is provided with all the biopsies that belong to each patient so that the pathologist can diagnose the patient. The agreements for each class are provided in Table 6.8. The overall agreement was calculated after deleting the result of the “indefinite for dysplasia” class. According to the study, the agreement is enhanced by diagnosing per patient instead of diagnosing per biopsy. Compared to the proposed CAD system, 106 biopsies from 15 patients were used. The agreement for diagnosing HGD and NFD is slightly higher in the per-patient analysis while far higher for LGD. The agreement for the proposed CAD system is substantial for both the diagnosis per patient and per biopsy analysis, while the study recorded moderate agreements.

**Table 6.8** The interobserver agreements in the diagnosing of dysplasia per biopsy and per patient in (Salomao et al., 2018) paper and the proposed CAD system

	Diagnosed per biopsy		Diagnosed per patient	
	KV (95% confidence interval)		KV (95% confidence interval)	
	The proposed CAD (n=106)	(Salomao et al., 2018) (n=549)	The proposed CAD (n=15)	(Salomao et al., 2018) (n=129)
Overall	0.671 (0.503-0.840)	0.523 (0.48-0.57)	0.8 (0.549-1)	0.587 (0.48-0.68)
NFD	0.679 (0.504-0.855)	0.61 (0.57-0.66)	0.727 (0.389-1)	0.66 (0.56-0.76)
LGD	0.394 (0.118-0.669)	0.3 (0.25-0.35)	0.667 (0.261-1)	0.31 (0.21-0.41)
HGD	0.947 (0.844-1)	0.66 (0.61-0.71)	1 (1-1)	0.79 (0.66-0.86)

## 6.7 Conclusion

This chapter presents an essential contribution to this thesis: a multi-scale CAD system for Barrett’s related dysplasia detection and grading using histological digitalised tissue samples that simulate the pathologists’ cognition in selecting magnifications during their examination. The proposed

CAD system was built from the trained networks in the previous chapters. Also, many experiments were conducted to pick components for the proposed CAD. In this chapter, two architectures were introduced as solutions. The first one is to integrate the proposed networks from Chapter 4 to detect the regions of interest and grade them twice based on extracted features by the analysis at 10X (refer to Chapter 4) and 40X magnifications, using networks from Chapter 5. Then, a consensus grade for each detected region is computed using a novel solution designed based on the assumption that the learnt features at 10X and 40X magnifications are architectural and cytological, respectively. Also, it was inspired by the histopathology guidelines in grading Barrett's related dysplasia that state: that a region is graded as NFD when its architectural and cytological features are preserved or has mild changes to either of the architectural or cytological features, but not both as it is conceded as LGD. In addition, it is graded as LGD if either architectural or cytological features have extreme changes. In cases where extreme changes occur to the cytological features while the architectural features have mild abnormal changes, they are graded as LGD and HGD and vice versa. The first solution did not produce the optimum performance due to involving the region of interest detector as it neglected important information. Therefore, the second solution sought another reliable way to detect regions of interest. By reviewing information about the provided annotations, it was found that pathologists tend to rely on lower magnification levels, such as 5X and 10X, to grade NFD. In contrast, higher magnifications are used to grade PFD regions. By mimicking their way, patches at 10X magnification were used to detect the PFD. Then, sampled patches from those regions at 40X magnification were examined to decide the severity of dysplasia.

The proposed CAD system is a multi-scale histological image processing and analysis with a pyramid structure. By starting from the pyramidal peak, the system detects foregrounds from the lowest magnification version of the slide. It reduces noises to find foregrounds that are most likely to become biopsies. That module managed to detect all the biopsies within the test slides. Then, the system processes the image at a higher scale (10X magnification) to extract features and classify them under one of the three grades of dysplasia. During conducting the experiments on the different approaches, it was noted that the performance of that module is high for NFD, and dysplastic slides are rarely graded as NFD. Thus, regions

detected as dysplasia by the previous stage are fed through further analysis at 40X magnification to be reclassified. Up to that phase, the CAD system detects and classifies relevant regions that might be located in any layer of the oesophagus. It visualises the grade for each region within the slide. Even though the distribution of each grade in each biopsy within a slide can successfully decide the grade of that slide, this is not a scientific way of grading dysplasia because grading dysplasia relies on a complex process of balancing the morphological changes in the architectural and cytological levels.

Moreover, the location of the changes contributes to the grade decision. For instance, in tissue, when changes affect the base of crypts, that can be tolerated, but not if they affect the epithelium surface. Hence, to ensure that the CAD system is fully automated and does not need any human interaction at any stage, a slide grading inference system was utilised following the annotation grading inference system (refer to section 4.3.2), given the grades of detected regions in tissue. Comparing the proposed CAD system against two CAD systems using a small-size dataset. The proposed CAD scored better than both proposed works by Adam (2015) and Liu et al. (2017) as it achieved 65% accuracy, 0.68 precision, 0.65 recall, 0.84 specificity, and 0.64 F1-score. Moreover, it has a substantial agreement with pathologists similar to the level of agreement between the pathologists themselves in detecting, localising and grading dysplasia in lesions.

Although the previous findings indicated a well-performing tool to detect, localise and grade Barrett's related dysplasia in histological images, it has two limitations. The proposed consensus system for the first approach in this chapter was designed based on an assumption and was not scientifically justified. Furthermore, the selection of the components of the proposed CAD system was derived from the analysis of the pathologists' behaviours and the conducted experiments for different modules. In future work, methods for averaging the predictions of multiple classifiers, majority voting, or ensemble classifiers should be considered to improve the accuracy of the CAD system.

## **Chapter 7. Conclusion and Future Work**

### **7.1 Thesis summary**

This research aims to develop a fully automated CAD system that assists pathologists in detecting and grading dysplasia in Barrett's oesophagus patients. This aim was fulfilled through the chapters of this thesis.

Chapter 1 provided brief information about Barrett's oesophagus and its potential development into cancer through different degrees of dysplasia. Also, it provided the aim and objectives, which were motivated by two facts. The first fact relates particularly to dysplasia in Barrett's oesophagus, which suffers from fair to moderate interobserver agreement. While the second fact relates to the nature of histopathological assessment of biopsies, as pathologists tend to search for regions of interest at a low-power magnification to be examined at high magnifications only, which might increase the chance of losing valuable information from missed small regions that cannot be observed at low-power magnification, finally, it provides an overall structure of the thesis framework.

Chapter 2 identified Barrett's oesophagus dysplasia and the guidelines to grade it, and it explained the preparation and the structure of the histological virtual slides. Besides, it presented background material on deep-learning and its different approaches and some of its network architectures. They were used to detect and classify dysplasia and discussed the one-class classification tasks.

The aim of this research is fulfilled by developing a CAD system, which involves a module for foreground detection and noise reduction and a module for the region of interest detection. Then, to emulate the pathologists' way of assessing the grade of dysplasia by evaluating the abnormal alteration in architectural and cytological features of the examined tissue. Changes are analysed at 10X and 40X magnifications. The concepts of the proposed CAD system are presented in Chapter 3, Chapter 4, Chapter 5 and Chapter 6.

Chapter 3 described this thesis's used dataset, including the whole virtual slides and the manually selected and labelled annotations. The chapter also explained the pre-processing phase to detect the foreground and reduce noise. Finally, it explained the process of patch sampling to generate datasets suitable for training and testing the deep-learning approaches.

Chapter 4 proposed a model for the region of interest detection and dysplasia grading based on the analysis at a low-power magnification (10X). The model followed the transfer learning approach to fine-tune several layers for two separate "Inception-ResNet-v2" networks. One network detects the critical regions, and the other is to extract features from the detected regions and classifies them. The fine-tuned layers from the second network are connected to the first one at the endpoint of the frozen layers to reduce the computational time. Also, a histogram-based random forest classifier was employed to infer the grade of the annotations.

Chapter 5 discussed the main challenge in the provided annotations of Barrett's oesophagus related dysplasia. The challenge was the coarse-grain dysplastic annotations, as these annotations contain a mixture of dysplastic and not dysplastic tissue attributed to the nature of the rectangular annotations. Many approaches were tested to tackle this issue, and the best solution was to train a novelty detector on the non-dysplastic annotations. It was tested on the dysplastic annotations to filter the non-dysplastic tissue. That detector was implemented by fine-tuning two identical "Inception-ResNet-v2" networks that share the weight set. One network aims to minimise the distance between the representations of the samples in Barrett's oesophagus dataset using one of the nearest neighbour algorithms, and the other aims to classify a new public dataset, a fine-grained breast cancer dataset, using cross-entropy. The two networks were updated with the total of the two losses produced by the networks. Using the total loss increases the fine-tuned model's compactness and descriptiveness. Then, the networks' classifiers were removed

In Chapter 6, detailed experiments were discussed to find the best architecture for a CAD system. As a result, two architectures were proposed and tested. The first one is a combination of the models from Chapter 4 and Chapter 5. The region of interest network focuses on finding important

regions similar to the pathologists annotated. A set of features are extracted at 10X and 40X magnifications, and each candidate region is graded twice at two magnifications. Then, a consensus grading solution between the two grading systems was found. The second approach is the high-level-based network combined with the low-level-based network. This approach utilises a multi-magnifications strategy to detect dysplastic regions and grade them in whole slides after detecting the tissue and eliminating the background and any artefacts.

Contrary to the first approach, it does not utilise a region of interest detector network, which discriminates normal oesophagus tissue against Barrett's oesophagus tissue. Alternatively, it uses the high-level-based dysplasia classification network (at 10X magnification) to discriminate between NFD and dysplastic regions by processing the sampled patches at low magnification, reducing the computation burden. The corresponding patches at higher magnification are processed whenever more sophisticated information is needed to decide the degree of dysplasia or where the network at lower magnification produces a low confidence score in predicting the grade. In other words, the high-level based network from Chapter 4 was used as a region of interest detector as the performance of the model showed that the model rarely misclassified dysplastic tissue as NFD, emulating the pathologists' way of grading NFD, and the low-level based network, without the potential dysplastic one-class classifier, was used for the sophisticated analysis for the dysplastic regions. To this end, a thumbnail image for each slide is generated with a white background and three colours, green for NFD, blue for LGD and red for HGD, representing the prediction map for the tissue where the probability for the prediction is equal or more than 50%.

## **7.2 Key contributions and findings**

This thesis presents approaches, techniques, and experiments searched, enhanced, and applied for this research over the last four years. The chapters were presented in chronological order starting from data pre-processing and cleaning in Chapter 3 and ending with the integration of the parts of the proposed CAD system in Chapter 6. This section will summarise the findings of the held experiments and clarify the contributions of this

research. This research made innovative contributions in several areas: image processing, computer vision, deep-learning, and histopathology society. The main contribution of this research is experimenting with the state-of-art deep-learning approaches to provide a tool for grading Barrett's oesophagus in histological images and comparing their performances against the proposed approach that is presented by Adam (2015), which adopted a traditional machine learning approach to provide an automatic diagnosis for the same disease. That comparison was held twice, in Chapter 4 to compare against a benchmark that analyses changes at the same level (10X magnification) and in Chapter 6 as a benchmark for a CAD system. For both comparative studies, the proposed work by this thesis had higher results.

Furthermore, another main contribution is investigating the efficiency of involving a novelty detection approach to address the issue of coarse-grained labelling in histological images. The addition of the novelty detection approach affected the model positively. Finally, this thesis sought a solution that mimics the pathologists during their examining oesophageal tissues and proposed a novel CAD system that meets that goal. The CAD system was compared with a multi-scale approach proposed by Liu et al. (2017) and showed better results.

Moreover, experiments in Chapter 3 were focused on detecting the regions that include tissue to release loads of computational burdens that can be caused by analysing backgrounds and noises. The used approach successfully detected all the sampled biopsies in WSIs. Also, the proposed technique for guiding the process of sampling the tiles is by surrounding each detected foreground with a box dividable by the tile size to guarantee it is not wasting any information from the tissue. This chapter contributes to image pre-processing in the field of histopathology by enhancing a tool for WSIs preparation and sampling. It can be applied to any H&E stained histological WSIs dataset for biopsies extracted from any organ such as colon, prostate, breast, or oral tissue.

In Chapter 4, the proposed model for classifying the extracted features at 10X magnification into three grades of dysplasia achieved an overall 0.52 recall, 0.78 specificity, 0.51 F1-score and 52% accuracy at the annotation



level. After introducing the region of interest detection model, these measurements increased to reach an overall 0.50 recall, 0.77 specificity, 0.50 F1-score and 50% accuracy. Both models work alongside each other at the slide level to yield a 0.57 recall. Even though the region of interest detector decreased the performance slightly when it was attached to the model, it is essential to reduce the computational costs, increase the speed of the system when further higher magnification analyses are required, and enhance the performance of the classifier by obviating regions where might mislead the such as the muscular mucosa or the healthy oesophagus tissue which were not present in the training the annotations. This chapter contributed to the pathology society by introducing an acceptable performing, fast and low-budget tool to detect and grade dysplasia in H&E stained WSIs. Also, it provided a novel annotation and slide levels inference system to decide the grade of bags (annotations or slides) of instances (sampled patches), taking into account the presence of abnormal changes in each layer of the oesophagus.

Furthermore, Chapter 5 tackles the main challenge we faced in Barrett's oesophagus dataset, which is the presence of non-dysplastic tissues in the provided dysplastic annotations. This chapter contributes to the deep-learning community by providing a novel solution for the MIL weakly supervised problem. This solution combines a one-class classifier for non-dysplastic tissue with the regular multi-class classifier for dysplasia to eliminate instances of non-dysplastic, so they do not fog the multi-class classifier. By examining this solution with the provided annotations, the main findings emerge that the overall measurements were increased to 0.67 recall, 0.84 specificity, 0.64 F1-score, and 63% accuracy compared to the results using the multi-class classifier solely, which are 0.65 recall, 0.84 specificity, 0.63 F1-score and 63% accuracy. In general, the solution achieved the expected goal of enhancing performance.

Finally, Chapter 6 shows the attempt to integrate the fully automated CAD system components and proposes the two applicable approaches to grade the slide based on the highest annotation. After rigorous examinations, it was concluded that the grading performance for the analysis at 10X magnification is not as good as its performance for the analysis at 40X magnification. Besides, even though adding the novelty detector had the same overall performance as the analysis at 40X magnification, it was

discovered that grading HGD annotations and slides had better performance using Barrett's related dysplasia classification based on the analysis at 40X magnification only. In contrast, the NFD and LGD grades performed better by adding the novelty detection module to filter the non-dysplastic tissues from the annotations. According to (Wang and Sampliner, 2008), the "Practice Parameters Committee of the American College of Gastroenterology" recommends a treatment plan for diagnosing Barrett's oesophagus-related dysplasia, including mucosal resection only for patients diagnosed with HGD. In contrast, it recommends endoscopic surveillance every specific interval of months or years. Since diagnosing HGD requires medical intervention, the choice of grading dysplasia using the cytological features without the novelty detector seems to be a wise decision. However, if a CAD system is interested in grading the most challenging grade (LGD), then adding the novelty detector meets that need. As a result of these experiments, two architectures were proposed. The first one, as discussed before, needs a solution to calculate the consensus grade between the two levels of features.

The novel proposed solution presents the equal importance of both analyses at the two magnifications to detect abnormalities in grading dysplasia. Thus when the grading at two magnifications disagrees, new rules concluded from the grading guidelines are engaged. These rules are: (1) when one of the grades belongs to NFD and the other to LGD, then the calculated grade is NFD, (2) when one of the grades belongs to NFD and the other to HGD. The calculated grade is LGD, (3) and (4) are when the grading based on the architectural features is LGD, and the other grade is HGD, then the grade is LGD and HGD if vice versa. Employing this solution with the first architecture for the CAD system scored 0.79 recall, 0.89 specificity, 0.74 F1-score and 73% accuracy. Even though the results are promising, when the architecture was applied with a non-overlapped technique on the WSIs, it neglected many regions from the provided annotations. Therefore, it encouraged us to seek an alternative solution for the region of interest detection. The second proposed solution was drawn from the observation of the behaviour of the pathologists in grading dysplasia. It was concluded that NFD does not need higher magnifications to be diagnosed, and diagnosing dysplastic regions (LGD or HGD) are analysed better at a higher magnification. After applying this solution to detect the dysplastic regions, it performed flawlessly and then classified the detected regions based on the extracted features at 40X

magnification on the WSIs. The proposed CAD system achieved 0.90 recall, 0.94 specificity, 0.86 F1-score and 87% accuracy.

To sum up, Chapter 6 is a crucial part of the thesis contributions. Its contributions are presented as follows: (1) it contributes to the histopathology community by providing a novel consensus grading system between the grading systems based on the assumed extracted architectural and cytological abnormalities. (2) it provides a unique solution for the region of interest detection that mimics the way of the pathologists. (3) It contributed by providing a fully automated CAD system that engages every lesion in the slide in the diagnosing contrary to (Adam, 2015), where regions were selected randomly.

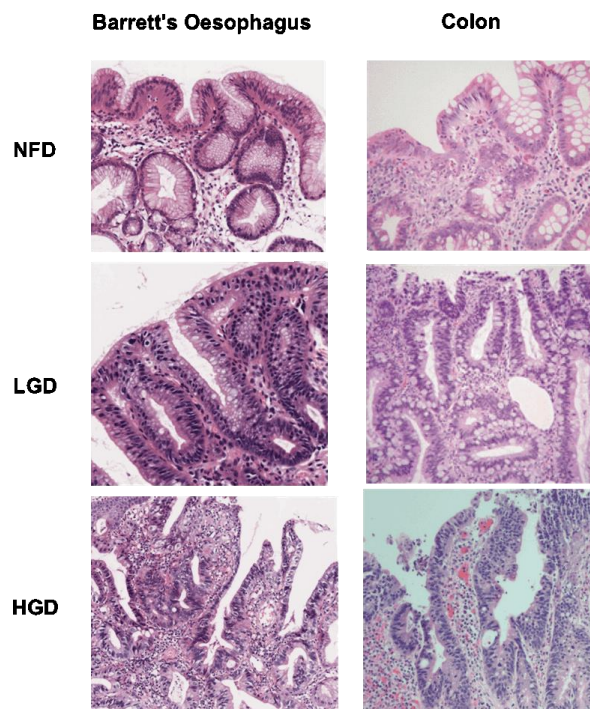
### **7.3 Research strengths, limitations and opportunities for future research**

CAD systems can help pathologists in three ways: they can find interesting regions in the whole slides where the pathologist should look to save their time and guarantee to capture all the critical regions and not neglect fateful regions with valuable information; they can diagnose each region in the lesion and provide an accurate assessment for it, and they can accept the WSIs and produce diagnosis and grades for diseases. The second and third types of CAD systems reduce the diversity in experts' interpretations. Unfortunately, most of the published research that proposed the CAD system and its application is limited to analysing either annotations that expert pathologists manually selected or regions that were arbitrarily picked and not on the level of the whole slides. Contrary to that, the proposed CAD system in this research is a fully automated system that engages each pixel within the detected biopsies in the diagnosing.

The proposed CAD system showed the ability of transfer learning to handle a wide range of heterogeneous structures and layers. For instance, examples of the normal oesophagus were not sufficiently included in the provided annotations. However, the system has graded them as NFD because some NFD annotations include a minor amount of such tissue. Moreover, the pyramidal structure of the multi-resolution CAD system has remarkably cut down the computational cost of diagnosing the whole slides

by eliminating the background and artefacts and analysing tiles of the tissue at low magnification. Then it applies more sophisticated analysis at higher magnification to tiles where the system is not particular about or suspects dysplastic. Also, when the CAD system misclassified an annotation, it assigned it to an adjacent class, indicating that the system is not guessing. In addition, the CAD system provides visualisation maps for the input WSI using a colouring scheme to illustrate the grade of each region in the virtual slide enabling pathologists to find regions where they should examine.

The proposed CAD system includes an annotation and slide grading inference system. That system is based on the histogram of the occurrence of each grade in each layer of epithelial and lamina propria layers. Then the histogram of either the annotation or the slide is classified by a trained random forest classifier to decide the final grade. Lastly, the proposed CAD system was mainly designed to assist pathologists in detecting and grading Barrett's oesophagus related dysplasia; however, there is a potential for using it with colon related dysplasia. Due to the similarities between the structure of Barrett's oesophagus tissue and the colon tissue and the similarities in their dysplastic changes, as shown in Figure 7.1.



**Figure 7.1** NFD, LGD, and HGD samples show similarities between Barrett's oesophagus tissue and colon tissue

Despite the success demonstrated, this CAD system suffers from several limitations. The major limitation is that sampled patches from both the epithelial and lamina propria layers were used to train the proposed CAD system without considering that in diagnosing each grade of dysplasia, the abnormalities degrees in each layer have different representations. Likewise, the annotation and slide inference systems employ a dataset that manually segmented the epithelial layer apart from the lamina propria layer. A potential solution would be detecting the epithelial layer to subdivide the tissue into layers. As most of the dysplastic abnormal changes occur in the epithelial layer, it will be possible to predict the grade of the whole slide based on the dominant higher grade in that layer. In addition, another distinct network could be trained on the extracted features from the lamina propria to support the decision of the overall grade.

Further processing on the generated heatmaps could be employed to analyse the distribution of the changes and the spatial relation between them or their sequence. Possible algorithms include hierarchical clustering, Voronoi diagram or recurrent networks. That will be a significant direction in the future. Another limitation is dealing with IMC cases, as the system does not show a robust performance. That is not necessarily a challenging task for the proposed system, but it is attributed to the underrepresentation of such cases in the training set. Also, IMC's nature changes that extend beyond the epithelial layer and make it unrecognisable.

One of the limitations of this thesis is the proposed consensus system to grade dysplasia, which was proposed for the first approach in section 6.2.1. The rules were formulated for the consensus between the predictions of the models using different magnifications (40X and 10X) based on assumptions. It was assumed that the extracted features at 10X and 40X magnifications are architectural and cytological features based on the researcher's observations. In addition, it was assumed that the solution for finding a consensus between the classifications at different levels is imitating the pathology guidelines in grading dysplasia in Barrett's oesophagus using the architectural and cytological changes, as explained in section 2.1.6. Future researchers should consider investigating the employment of ensemble classifiers that rely on averaging the classifiers' confidences or probabilities or voting for the final grade to accomplish the consensus task. Better yet, training consensus learning approaches should be considered to draw the

final decision. Those approaches focus on the heterogeneity of the used algorithms for classification other than the input data representations. For instance, (Chakraborty et al., 2017) and (Plewczynski, 2009) presented consensus learning approaches that train ensemble classifiers.

This research used H&E stained whole virtual slides, which the same laboratory provided. Thus, skipping the phase of stain normalisation did not affect the performance of the CAD system. However, using this CAD system with virtual slides from different laboratories would remarkably deteriorate the performance. Such behaviour is because laboratories follow different staining protocols resulting in a different amount of eosin and hematoxylin in the biopsies. Human eyes can easily cope with variations in colours; on the contrary, CAD systems could be hampered by such variations.

In histopathology, during the preparation of biopsies, some tissue might defect in the phase of sectioning. That results in a small part of tissue detaching to the primary biopsy. From our observation, we noticed that the CAD system misclassifies those detached tissues, usually to higher grades, and results in affecting the overall grade. That highlights the potential usefulness of developing an algorithm to stitch tissues.

## List of References

- ABDEL-HAMID, O., MOHAMED, A.-R., JIANG, H. & PENN, G. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. 2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP), 2012. IEEE, 4277-4280.
- ADAM, A. 2015. *Computer Aided Dysplasia Grading for Barrett's Oesophagus Virtual Slides*. University of Leeds.
- ADAM, A., BULPITT, A. & TREANOR, D. Grading Dysplasia in Barrett's Oesophagus Virtual Pathology Slides with Cluster Co-occurrence Matrices. Proc. of Histopathology Image Analysis: Image Computing in Digital Pathology in conjunction with The 15th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2012.
- ADAM, A., BULPITT, A. J. & TREANOR, D. Texture Analysis of Virtual Slides for Grading Dysplasia in Barretts Oesophagus. MIUA, 2011. 269-274.
- AKCAY, S., ATAPOUR-ABARGHOUEI, A. & BRECKON, T. P. Ganomaly: Semi-supervised anomaly detection via adversarial training. Asian Conference on Computer Vision, 2018. Springer, 622-637.
- ALTSCHULER, S. J. & WU, L. F. 2010. Cellular heterogeneity: do differences make a difference? *Cell*, 141, 559-563.
- ARESTA, G., ARAÚJO, T., KWOK, S., CHENNAMSETTY, S. S., SAFWAN, M., ALEX, V., MARAMI, B., PRASTAWA, M., CHAN, M. & DONOVAN, M. 2019. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56, 122-139.
- BAHDANAU, D., CHO, K. & BENGIO, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1609.07744*.
- BASU, K. K. & DE CAESTECKER, J. S. 2002. Surveillance in Barrett's oesophagus: a personal view. *The Postgraduate Medical Journal*, 78, 263-268.
- BEJNORDI, B. E., VETA, M., VAN DIEST, P. J., VAN GINNEKEN, B., KARSSEMEIJER, N., LITJENS, G., VAN DER LAAK, J. A., HERMSEN, M., MANSON, Q. F. & BALKENHOL, M. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA network open*, 318, 2199-2210.
- BHAT, S., COLEMAN, H. G., YOUSEF, F., JOHNSTON, B. T., MCMANUS, D. T., GAVIN, A. T. & MURRAY, L. J. 2011. Risk of malignant progression in Barrett's esophagus patients: results from a large population-based study. *Journal of the National Cancer Institute*, 103, 1049-1057.
- BIRD-LIEBERMAN, E. L., DUNN, J. M., COLEMAN, H. G., LAO-SIRIEIX, P., OUKRIF, D., MOORE, C. E., VARGHESE, S., JOHNSTON, B. T., ARTHUR, K. & MCMANUS, D. T. 2012. Population-based study reveals new risk-stratification biomarker panel for Barrett's esophagus. *Gastroenterology*, 143, 927-935. e3.

- BRANCATI, N., FRUCCI, M. & RICCIO, D. Multi-classification of breast cancer histology images by using a fine-tuning strategy. *International conference image analysis and recognition*, 2018. Springer, 771-778.
- BREUNIG, M. M., KRIEGEL, H.-P., NG, R. T. & SANDER, J. LOF: identifying density-based local outliers. *ACM sigmod record*, 2000. ACM, 93-104.
- CAMPANELLA, G., HANNA, M. G., GENESLAW, L., MIRAFLORES, A., SILVA, V. W. K., BUSAM, K. J., BROGI, E., REUTER, V. E., KLIMSTRA, D. S. & FUCHS, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25, 1301-1309.
- CHAKRABORTY, T., CHANDHOK, D. & SUBRAHMANYAN, V. MC3: A multi-class consensus classification framework. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2017. Springer, 343-355.
- CHALAPATHY, R. & CHAWLA, S. 2019. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:03407*.
- CHALAPATHY, R., MENON, A. K. & CHAWLA, S. 2018. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:06360*.
- CHEN, H., WANG, X. & HENG, P. A. Automated mitosis detection with deep regression networks. *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016. IEEE, 1204-1207.
- CHENNAMSETTY, S. S., SAFWAN, M. & ALEX, V. Classification of breast cancer histology image using ensemble of pre-trained neural networks. *International conference image analysis and recognition*, 2018. Springer, 804-811.
- CHEPLYGINA, V., DE BRUIJNE, M. & PLUIM, J. P. 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54, 280-296.
- CIREŞAN, D. C., GIUSTI, A., GAMBARDELLA, L. M. & SCHMIDHUBER, J. Mitosis detection in breast cancer histology images with deep neural networks. *International conference on medical image computing and computer-assisted intervention*, 2013. Springer, 411-418.
- CLEVERT, D.-A., UNTERTHINER, T. & HOCHREITER, S. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:07289*.
- COCO, D. P., GOLDBLUM, J. R., HORNICK, J. L., LAUWERS, G. Y., MONTGOMERY, E., SRIVASTAVA, A., WANG, H. & ODZE, R. D. 2011. Interobserver variability in the diagnosis of crypt dysplasia in Barrett esophagus. *The American journal of surgical pathology*, 35, 45-54.
- COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educational psychological measurement*, 20, 37-46.
- COHEN, J. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70, 213.
- CUI, D., LIU, Y., LIU, G. & LIU, L. 2020. A Multiple-Instance Learning-Based Convolutional Neural Network Model to Detect the IDH1 Mutation in the Histopathology Images of Glioma Tissues. *Journal of Computational Biology*.
- DAS, K., CONJETI, S., ROY, A. G., CHATTERJEE, J. & SHEET, D. Multiple instance learning of deep convolutional neural networks for breast



- histopathology whole slide classification. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018. IEEE, 578-581.
- DELPISHEH, A., VEISANI, Y., SAYEHMIRI, K. & RAHIMI, E. 2014. Esophageal carcinoma: long-term survival in consecutive series of patients through a retrospective cohort study. *Gastroenterology and hepatology from bed to bench*, 7, 101.
- DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. & FEI-FEI, L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition, 2009. Ieee, 248-255.
- DENG, L. & YU, D. 2014. Deep learning: methods and applications. *Foundations Trends® in Signal Processing*, 7, 197-387.
- DIETTERICH, T. G., LATHROP, R. H. & LOZANO-PEREZ, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89, 31-71.
- DUCHI, J., HAZAN, E. & SINGER, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12, 2121-2159.
- ELEFTHERIADIS, N., INOUE, H., IKEDA, H., ONIMARU, M., YOSHIDA, A., MASELLI, R., SANTI, G. & KUDO, S.-E. 2014. Definition and staging of early esophageal, gastric and colorectal cancer. *Journal of Tumor*, 2, 161-178.
- FERHATOGLU, M. F. & KIVILCIM, T. 2017. Anatomy of Esophagus. In: CHAI, J. (ed.) *Esophageal Abnormalities*. IntechOpen.
- FISICHELLA, P. M. & PATTI, M. G. 2001. Normal physiology of the esophagus. In: TILANUS, H. W. & ATTWOOD, S. E. A. (eds.) *Barrett's Esophagus*. Springer.
- FLEJOU, J. 2005. Barrett's oesophagus: from metaplasia to dysplasia and cancer. *Gut*, 54, i6-i12.
- GALAL, S. & SANCHEZ-FREIRE, V. Candy cane: Breast cancer pixel-wise labeling with fully convolutional densenets. International Conference Image Analysis and Recognition, 2018. Springer, 820-826.
- GARRETT, B. 2014. *Study Guide to Accompany Bob Garrett's Brain & Behavior: An Introduction to Biological Psychology*, Sage Publications.
- GARUD, S. S., KEILIN, S., CAI, Q. & WILLINGHAM, F. F. 2010. Diagnosis and management of Barrett's esophagus for the endoscopist. *Therapeutic advances in gastroenterology*, 3, 227-238.
- GATENBY, P. A., RAMUS, J. R., CAYGILL, C. P., SHEPHERD, N. A. & WATSON, A. 2008. Relevance of the detection of intestinal metaplasia in non-dysplastic columnar-lined oesophagus. *Scandinavian journal of gastroenterology*, 43, 524-530.
- GERON, A. 2017. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*, " O'Reilly Media, Inc.".
- GOLDSTEIN, M. & UCHIDA, S. 2016. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS one*, 11, e0152173.
- GRIN, A. & STREUTKER, C. J. 2014. Histopathology in Barrett esophagus and Barrett esophagus-related dysplasia. *Clinical endoscopy*, 47, 31.

- GRUBBS, F. E. 1969. Procedures for detecting outlying observations in samples. *Technometrics*, 11, 1-21.
- GU, J., SCHUBERT, M. & TRESP, V. 2018. Semi-supervised Outlier Detection using Generative And Adversary Framework.
- GURCAN, M. N., BOUCHERON, L., CAN, A., MADABHUSHI, A., RAJPOOT, N. & YENER, B. 2009. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2, 147.
- GUTOSKI, M., AQUINO, N. M. R., RIBEIRO, M., LAZZARETTI, E. & LOPES, H. S. Detection of video anomalies using convolutional autoencoders and one-class support vector machines. XIII Brazilian Congress on Computational Intelligence, 2017, 2017.
- HAGGERTY, J. M., WANG, X. N., DICKINSON, A., O'MALLEY, C. J. & MARTIN, E. B. 2014. Segmentation of epidermal tissue with histopathological damage in images of haematoxylin and eosin stained human skin. *BMC medical imaging*, 14, 7.
- HAGGITT, R. C. 1994. Barrett's esophagus, dysplasia, and adenocarcinoma. *Human pathology*, 25, 982-993.
- HARALICK, R. M., SHANMUGAM, K., DINSTEN, I. H. & CYBERNETICS 1973. Textural features for image classification. *IEEE Transactions on systems, man*, 610-621.
- HAVAEI, M., DAVY, A., WARDE-FARLEY, D., BIARD, A., COURVILLE, A., BENGIO, Y., PAL, C., JODOIN, P.-M. & LAROCHELLE, H. 2017. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35, 18-31.
- HE, K., ZHANG, X., REN, S. & SUN, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. 770-778.
- HEBB, D. O. 1952. *The organisation of behaviour: a neuropsychological theory*, Wiley.
- HINTON, G., SRIVASTAVA, N. & SWERSKY, K. J. C. O. 2012. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. 14.
- HOU, L., NGUYEN, V., KANEVSKY, A. B., SAMARAS, D., KURC, T. M., ZHAO, T., GUPTA, R. R., GAO, Y., CHEN, W. & FORAN, D. 2019. Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. *Pattern recognition*, 86, 188-200.
- HOU, L., SAMARAS, D., KURC, T. M., GAO, Y., DAVIS, J. E. & SALTZ, J. H. Patch-based convolutional neural network for whole slide tissue image classification. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. 2424-2433.
- HU, B., TANG, Y., ERIC, I., CHANG, C., FAN, Y., LAI, M. & XU, Y. 2018. Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks. *IEEE journal of biomedical health informatics*, 23, 1316-1328.
- HUBEL, D. H. & WIESEL, T. N. 1959. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148, 574-591.
- ILSE, M., TOMCZAK, J. M. & WELLING, M. 2018. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1808.04712*.
- KANDEMIR, M., FEUCHTINGER, A., WALCH, A. & HAMPRECHT, F. A. Digital pathology: Multiple instance learning can detect Barrett's

- cancer. 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), 2014. IEEE, 1348-1351.
- KANDEMIR, M., HAMPRECHT, F. A. & GRAPHICS 2015. Computer-aided diagnosis from weak supervision: A benchmarking study. *Computerized medical imaging*, 42, 44-50.
- KEELER, J. D., RUMELHART, D. E. & LEOW, W. K. Integrated segmentation and recognition of hand-printed numerals. *Advances in neural information processing systems*, 1991. 557-563.
- KELTY, C. J., GOUGH, M. D., VAN WYK, Q., STEPHENSON, T. J. & ACKROYD, R. 2007. Barrett's oesophagus: intestinal metaplasia is not essential for cancer risk. *Scandinavian journal of gastroenterology*, 42, 1271-1274.
- KERKHOF, M., VAN DEKKEN, H., STEYERBERG, E., MEIJER, G., MULDER, A., DE BRUINE, A., DRIESSEN, A., TEN KATE, F., KUSTERS, J. & KUIPERS, E. 2007. Grading of dysplasia in Barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists. *Histopathology*, 50, 920-927.
- KHAN, A. M., RAJPOOT, N., TREANOR, D. & MAGEE, D. 2014. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61, 1729-1738.
- KINGMA, D. P. & BA, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.0441*.
- KOHL, M., WALZ, C., LUDWIG, F., BRAUNEWELL, S. & BAUST, M. Assessment of breast cancer histology using densely connected convolutional networks. *International Conference Image Analysis and Recognition*, 2018. Springer, 903-913.
- KRAUS, O. Z., BA, J. L. & FREY, B. J. 2016. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32, i52-i59.
- KRIZHEVSKY, A., SUTSKEVER, I. & HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012. 1097-1105.
- KWOK, S. Multiclass classification of breast cancer in whole-slide images. *International conference image analysis and recognition*, 2018. Springer, 931-940.
- LECUN, Y. 1998. The MNIST database of handwritten digits.
- LECUN, Y., BOTTOU, L., BENGIO, Y. & HAFFNER, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 2278-2324.
- LECUN, Y. A., BOTTOU, L., ORR, G. B. & MULLER, K.-R. 2012. Efficient backprop. *Neural networks: Tricks of the trade*. Springer.
- LEE, B. & PAENG, K. A robust and effective approach towards accurate metastasis detection and pn-stage classification in breast cancer. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018. Springer, 841-850.
- LI, H., LIU, H., JI, X., LI, G. & SHI, L. 2017. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11, 309.
- LITJENS, G., BANDI, P., EHTESHAMI BEJNORDI, B., GEESINK, O., BALKENHOL, M., BULT, P., HALILOVIC, A., HERMSEN, M., VAN DE LOO, R. & VOGELS, R. 2018. 1399 H&E-stained sentinel lymph

- node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7, giy065.
- LIU, W., HAHN, H., ODZE, R. D. & GOYAL, R. K. 2009. Metaplastic esophageal columnar epithelium without goblet cells shows DNA content abnormalities similar to goblet cell-containing epithelium. Nature Publishing Group.
- LIU, Y., GADEPALLI, K., NOROUZI, M., DAHL, G. E., KOHLBERGER, T., BOYKO, A., VENUGOPALAN, S., TIMOFEEV, A., NELSON, P. Q. & CORRADO, G. S. 2017. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*.
- MALON, C., MILLER, M., BURGER, H. C., COSATTO, E. & GRAF, H. P. Identifying histological elements with convolutional neural networks. Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, 2008. 450-456.
- MONTGOMERY, E. 2005. Is there a way for pathologists to decrease interobserver variability in the diagnosis of dysplasia? *Archives of pathology and laboratory medicine*, 129, 174-176.
- MONTGOMERY, E., BRONNER, M. P., GOLDBLUM, J. R., GREENSON, J. K., HABER, M. M., HART, J., LAMPS, L. W., LAUWERS, G. Y., LAZENBY, A. J. & LEWIN, D. N. 2001. Reproducibility of the diagnosis of dysplasia in Barrett esophagus: a reaffirmation. *Human pathology*, 32, 368-378.
- NAINI, B. V., SOUZA, R. F. & ODZE, R. D. 2016. Barrett's esophagus: a comprehensive and contemporary review for pathologists. *The American journal of surgical pathology*, 40, e45.
- ODZE, R. 2006. Diagnosis and grading of dysplasia in Barrett's oesophagus. *Journal of clinical pathology*, 59, 1029-1038.
- ODZE, R. D. 2008. Update on the diagnosis and treatment of Barrett esophagus and related neoplastic precursor lesions. *Archives of pathology and laboratory medicine*, 132, 1577-1585.
- OPENSLIDE. 2013. *OpenSlide* [Online]. Available: <https://openslide.org/> [Accessed October 16 2019].
- OTSU, N. & CYBERNETICS 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man*, 9, 62-66.
- OZA, P. & PATEL, V. M. 2018. One-class convolutional neural network. *IEEE Signal Processing Letters*, 26, 277-281.
- PASCHALI, M., NAEEM, M. F., SIMSON, W., STEIGER, K., MOLLENHAUER, M. & NAVAB, N. 2019. Deep Learning Under the Microscope: Improving the Interpretability of Medical Imaging Neural Networks. *arXiv preprint arXiv:03127*.
- PECKHAM, M., KNIBBS, A. & PAXTON, S. 2003. *The Leeds Histology Guide* [Online]. Available: <https://www.histology.leeds.ac.uk/oral/oesophagus.php> [Accessed 17 May 2019].
- PERERA, P. & PATEL, V. M. 2019. Learning deep features for one-class classification. *IEEE Transactions on Image Processing*.
- PLEWCZYNSKI, D. 2009. BRAINSTORMING: consensus learning in practice. *arXiv preprint arXiv:0910.0949*.
- QUELLEC, G., CAZUGUEL, G., COCHENER, B. & LAMARD, M. 2017. Multiple-instance learning for medical image and video analysis. *IEEE reviews in biomedical engineering*, 10, 213-234.

- QUIROS, A. C., MURRAY-SMITH, R. & YUAN, K. 2019. Pathology GAN: Learning deep representations of cancer tissue. *arXiv preprint arXiv:02644*.
- REINHARD, E., ADHIKHMEN, M., GOOCH, B. & SHIRLEY, P. 2001. Color transfer between images. *IEEE Computer graphics and applications*, 21, 34-41.
- RICE, T. W., MENDELIN, J. E. & GOLDBLUM, J. R. Barrett's esophagus: Pathologic considerations and implications for treatment. *Seminars in thoracic and cardiovascular surgery*, 2005. Elsevier, 292-300.
- RIDDELL, R. H., GOLDMAN, H., RANSOHOFF, D. F., APPELMAN, H. D., FENOGLIO, C. M., HAGGITT, R. C., AHREN, C., CORREA, P., HAMILTON, S. & MORSON, B. 1983. Dysplasia in inflammatory bowel disease: standardized classification with provisional clinical applications. *Human pathology*, 14, 931-968.
- ROBBINS, H. & MONRO, S. 1951. A stochastic approximation method. *The annals of mathematical statistics*, 400-407.
- RONY, J., BELHARBI, S., DOLZ, J., AYED, I. B., MCCAFFREY, L. & GRANGER, E. 2019. Deep weakly-supervised learning methods for classification and localization in histology images: a survey. *arXiv preprint arXiv:03354*.
- ROUX, L., RACOCEANU, D., LOMENIE, N., KULIKOVA, M., IRSHAD, H., KLOSSA, J., CAPRON, F., GENESTIE, C., LE NAOUR, G. & GURCAN, M. N. 2013. Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of pathology informatics*, 4.
- RUFF, L., VANDERMEULEN, R., GOERNITZ, N., DEECKE, L., SIDDIQUI, S. A., BINDER, A., MULLER, E. & KLOFT, M. Deep one-class classification. *International Conference on Machine Learning*, 2018. 4393-4402.
- RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A. & BERNSTEIN, M. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115, 211-252.
- SALI, R., MORADINASAB, N., GULERIA, S., EHSAN, L., FERNANDES, P., SHAH, T. U., SYED, S. & BROWN, D. E. 2020. Deep learning for whole-slide tissue histopathology classification: A comparative study in the identification of dysplastic and non-dysplastic Barrett's esophagus. *Journal of Personalized Medicine*, 10, 141.
- SALOMAO, M. A., LAM-HIMLIN, D. & PAI, R. K. 2018. Substantial interobserver agreement in the diagnosis of dysplasia in Barrett esophagus upon review of a patient's entire set of biopsies. *The American journal of surgical pathology*, 42, 376-381.
- SCHLEMPER, R., RIDDELL, R., KATO, Y. E. A., BORCHARD, F., COOPER, H., DAWSEY, S., DIXON, M., FENOGLIO-PREISER, C., FLEJOU, J. & GEBOES, K. 2000. The Vienna classification of gastrointestinal epithelial neoplasia. *Gut*, 47, 251-255.
- SCHÖLKOPF, B., PLATT, J. C., SHAW-TAYLOR, J., SMOLA, A. J. & WILLIAMSON, R. C. 2001. Estimating the support of a high-dimensional distribution. *Neural computation*, 13, 1443-1471.
- SHABAN, M. T., BAUR, C., NAVAB, N. & ALBARQOUNI, S. Staining: Stain style transfer for digital histological images. *2019 IEEE 16th*

- international symposium on biomedical imaging (Isbi 2019), 2019. IEEE, 953-956.
- SHAHEEN, N. J. & RICHTER, J. E. 2009. Barrett's oesophagus. *The Lancet*, 373, 850-861.
- SHAHEEN, O., GHIBOUR, A. & ALSAID, B. 2017. Esophageal cancer metastases to unexpected sites: a systematic review. *Gastroenterology Research and Practice*, 2017.
- SIMONYAN, K. & ZISSERMAN, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.0052*.
- SIRINUKUNWATTANA, K., RAZA, S. E. A., TSANG, Y.-W., SNEAD, D. R., CREE, I. A. & RAJPOOT, N. M. 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35, 1196-1206.
- SPANHOL, F. A., OLIVEIRA, L. S., PETITJEAN, C. & HEUTTE, L. 2015. A dataset for breast cancer histopathological image classification. 63, 1455-1462.
- SPECHLER, S. J. 2002. Barrett's esophagus. *New England Journal of Medicine*, 346, 836-842.
- SPECHLER, S. J. 2007. Screening and surveillance for Barrett's esophagus—an unresolved dilemma. *Nature Clinical Practice Gastroenterology & Hepatology*, 4, 470-471.
- SRINIDHI, C. L., CIGA, O. & MARTEL, A. L. 2019. Deep neural network models for computational histopathology: A survey. *arXiv preprint arXiv:1903.08238*.
- SUDHARSHAN, P., PETITJEAN, C., SPANHOL, F., OLIVEIRA, L. E., HEUTTE, L. & HONEINE, P. 2019. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117, 103-111.
- SUN, M., HAN, T. X., LIU, M.-C. & KHODAYARI-ROSTAMABAD, A. Multiple instance learning convolutional neural networks for object recognition. 2016 23rd International Conference on Pattern Recognition (ICPR), 2016. IEEE, 3270-3275.
- SZEGEDY, C., IOFFE, S., VANHOUCHE, V. & ALEMI, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-First AAAI Conference on Artificial Intelligence, 2017.
- SZEGEDY, C., LIU, W., JIA, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 1-9.
- TAN, C., SUN, F., KONG, T., ZHANG, W., YANG, C. & LIU, C. A survey on deep transfer learning. International conference on artificial neural networks, 2018. Springer, 270-279.
- TOMCZAK, K., CZERWIŃSKA, P. & WIZNEROWICZ, M. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, 19, A68.
- TOMITA, N., ABDOLLAHI, B., WEI, J., REN, B., SURIAWINATA, A. & HASSANPOUR, S. 2019. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. *JAMA network open*, 2.

- TORREY, L. & SHAVLIK, J. 2010. Transfer learning. *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI Global.
- TOSTA, T. A. A., DE FARIA, P. R., NEVES, L. A. & DO NASCIMENTO, M. Z. 2019. Computational normalization of H&E-stained histological images: Progress, challenges and future potential. *Artificial intelligence in medicine*, 95, 118-132.
- TREANOR, D., LIM, C. H., MAGEE, D., BULPITT, A. & QUIRKE, P. 2009. Tracking with virtual slides: a tool to study diagnostic error in histopathology. *Histopathology*, 55, 37-45.
- VAN DER WALT, S., SCHÖNBERGER, J. L., NUNEZ-IGLESIAS, J., BOULOGNE, F., WARNER, J. D., YAGER, N., GOUILLART, E. & YU, T. 2014. scikit-image: image processing in Python. *PeerJ*, 2, e453.
- VEELING, B. S., LINMANS, J., WINKENS, J., COHEN, T. & WELLING, M. Rotation equivariant cnns for digital pathology. International Conference on Medical image computing and computer-assisted intervention, 2018. Springer, 210-218.
- VENNALAGANTI, P., KANAKADANDI, V., GOLDBLUM, J. R., MATHUR, S. C., PATIL, D. T., OFFERHAUS, G. J., MEIJER, S. L., VIETH, M., ODZE, R. D. & SHREYAS, S. 2017. Discordance among pathologists in the United States and Europe in diagnosis of low-grade dysplasia for patients with Barrett's esophagus. *Gastroenterology Research and Practice*, 152, 564-570. e4.
- VLADIMIROV, B., IVANOVA, R. & TERZIEV, I. 2013. Diagnosis and Management of Barrett's Esophagus with and Without Dysplasia. *Endoscopy of GI Tract*. IntechOpen.
- WANG, D., KHOSLA, A., GARGEYA, R., IRSHAD, H. & BECK, A. H. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:05718*.
- WANG, F., OH, T. W., VERGARA-NIEDERMAYR, C., KURC, T. & SALTZ, J. Managing and querying whole slide images. Medical Imaging 2012: Advanced PACS-Based Imaging Informatics and Therapeutic Applications, 2012. International Society for Optics and Photonics, 83190J.
- WANG, H., ROA, A. C., BASAVANHALLY, A. N., GILMORE, H. L., SHIH, N., FELDMAN, M., TOMASZEWSKI, J., GONZALEZ, F. & MADABHUSHI, A. 2014. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1, 034003.
- WANG, K. K. & SAMPLINER, R. E. J. A. J. O. G. 2008. Updated guidelines 2008 for the diagnosis, surveillance and therapy of Barrett's esophagus. 103, 788-797.
- WANI, S., RUBENSTEIN, J. H., VIETH, M. & BERGMAN, J. 2016. Diagnosis and management of low-grade dysplasia in Barrett's esophagus: expert review from the Clinical Practice Updates Committee of the American Gastroenterological Association. *Gastroenterology*, 151, 822-835.
- WASHINGTON, K. 2010. of the AJCC cancer staging manual: stomach. *Annals of surgical oncology*, 17, 3077-3079.
- XU, J., LUO, X., WANG, G., GILMORE, H. & MADABHUSHI, A. 2016. A deep convolutional neural network for segmenting and classifying

- epithelial and stromal regions in histopathological images. *Neurocomputing*, 191, 214-223.
- XU, J., XIANG, L., LIU, Q., GILMORE, H., WU, J., TANG, J. & MADABHUSHI, A. 2015. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35, 119-130.
- XU, Y., MO, T., FENG, Q., ZHONG, P., LAI, M., ERIC, I. & CHANG, C. Deep learning of feature representation with multiple instance learning for medical image analysis. 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2014a. IEEE, 1626-1630.
- XU, Y., ZHU, J.-Y., ERIC, I., CHANG, C., LAI, M. & TU, Z. 2014b. Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis*, 18, 591-604.
- YAMASHITA, R., NISHIO, M., DO, R. K. G. & TOGASHI, K. 2018. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9, 611-629.
- ZARELLA, M. D., BOWMAN, D., AEFNER, F., FARAHANI, N., XTHONA, A., ABSAR, S. F., PARWANI, A., BUI, M., HARTMAN, D. J. & MEDICINE, L. 2018. A practical guide to whole slide imaging: a white paper from the digital pathology association. *Archives of pathology*, 143, 222-234.
- ZEILER, M. D. & FERGUS, R. Visualizing and understanding convolutional networks. European conference on computer vision, 2014. Springer, 818-833.
- ZHANG, X., SU, H., YANG, L. & ZHANG, S. Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 5361-5368.
- ZHOU, Z.-H. 2018. A brief introduction to weakly supervised learning. *National Science Review*, 5, 44-53.
- ZHOU, Z.-H., SUN, Y.-Y. & LI, Y.-F. Multi-instance learning by treating instances as non-iid samples. Proceedings of the 26th annual international conference on machine learning, 2009. 1249-1256.



## List of Abbreviations

<b>2D</b>	Two Dimensional
<b>3D</b>	Three Dimensional
<b>AdaGrad</b>	Adaptive Gradient Algorithm
<b>Adam</b>	Adaptive Moment Estimation
<b>ARC3</b>	Advanced Research Computing 3
<b>AUC</b>	Area Under Curve
<b>AUROC</b>	Receiver Operating Characteristics
<b>BACH</b>	BreAst Cancer Histology
<b>BACH</b>	Breast Cancer Histology Images
<b>BreaKHis</b>	The Breast Cancer Histopathological Image Classification
<b>CAD</b>	Computer-aided Diagnosis
<b>CCE</b>	Categorical Cross-Entropy
<b>CCI</b>	Cluster coded images
<b>CE</b>	Cross-Entropy
<b>CIFAR</b>	Canadian Institute for Advanced Research
<b>cm</b>	Centimetre
<b>CNN</b>	Convolutional Neural Network
<b>ELU</b>	Exponential Linear Unit
<b>FN</b>	False-negative
<b>FP</b>	False-positive
<b>GAN</b>	Generative Adversarial Network
<b>GB</b>	Gigabyte
<b>GERD</b>	Gastroesophageal reflux disease
<b>GLCM</b>	Grey-level co-occurrence matrices
<b>H&amp;E</b>	Hematoxylin and Eosin
<b>HGD</b>	High-grade dysplasia
<b>IBD</b>	inflammatory bowel disease

<b>ICPR</b>	International Conference on Pattern Recognition
<b>ILSVRC</b>	ImageNet Large Scale Visual Recognition Challenge
<b>IMC</b>	Intramucosal carcinoma
<b>KNN</b>	K-nearest neighbour
<b>KV</b>	Cohen kappa coefficient
<b>LGD</b>	Low-grade dysplasia
<b>LOF</b>	Local Outlier Factor
<b>MB</b>	Megabyte
<b>MIL</b>	Multiple Instances Learning
<b>mm</b>	Millimetre
<b>MNIST</b>	Mixed National Institute of Standards and Technology
<b>NFD</b>	Negative for dysplasia
<b>OC-SVM</b>	One-class Support Vector Machine
<b>PCam</b>	Patch Camelyon
<b>PFD</b>	Positive for dysplasia
<b>PPV</b>	Positive Predictive Value
<b>ReLU</b>	Rectified Linear Unit
<b>RMSprop</b>	Root Mean Square Propagation
<b>RNN</b>	Recurrent Neural Network
<b>ROC</b>	Receiver Operating Characteristic
<b>SVDD</b>	Support Vector Data Description
<b>TCGA</b>	The Cancer Genome Atlas
<b>TN</b>	True-negative
<b>TP</b>	True-positive
<b>t-SNE</b>	t-Distributed Stochastic Neighbour Embedding
<b>XML</b>	eXtensible Markup Language
<b>µm</b>	Micrometre

## Appendix A

### Materials for Chapter 3

#### A.1 Image pre-processing results for the whole virtual slides in the test set

Slide ID	Foreground pixels at a lowest available magnification	Background pixels at a lowest available magnification	Total number of pixels at a lowest available magnification	Background Percentage	Available magnifications	Available dimensions	Whole virtual slide image size in GB	Size after segmentation
10790	177028	2039222	2216250	92.0122729836435	40X 10X 2.5X 0.6X	144000X63066 36000X15766 9000X3941 2250X985	25.37	2.02648634404963
11040	217900	1808960	2026860	89.2493808156459	40X 10X 2.5X 0.6X	159360X52108 39840X13027 9960X3256 2490X814	23.20	2.49414365077016
11014	472319	4884481	5356800	91.1828143667861	40X 10X 2.5X 1.25X	115200X47646 28800X11911 7200X2977 3600X1488	15.34	1.35255627613501

11013	79733	1743307	1823040	95.6263713357908	40X 10X 2.5X 0.6X	138240X54029 34560X13507 8640X3376 2160X844	20.87	0.912776302220467
11035	127006	2096714	2223720	94.2885794974188	40X 10X 2.5X 0.6X	136320X66872 34080X16718 8520X4179 2130X1044	25.47	1.45469880200744
11063	408546	16280094	16688640	97.551951507133	40X 10X 2.5X	78720X54280 19680X13570 4920X3392	11.94	0.29229699004832
13154	141417	2243823	2385240	94.0711626502993	40X 10X 2.5X 0.6X	137280X71212 34320X17803 8580X4450 2145X1112	27.31	1.61916548020325
10829	231870	2234580	2466450	90.5990391047862	40X 10X 2.5X 0.6X	155520X65020 38880X16255 9720X4063 2430X1015	28.25	2.65577145289789
10586	105931	2345864	2451795	95.6794511776066	40X 10X 2.5X 0.6X	143040X70217 35760X17554 8940X4388 2235X1097	28.06	1.21234599956359
10857	125486	1667914	1793400	93.0028995204639	40X 10X 2.5X 1.25X	26880X68349 6720X17087 1680X4271 840X2135	5.13	0.358951254600201
13348	157512	2264163	2421675	93.4957415838211	40X 10X 2.5X 0.6X	135360X73325 33840X18331 8460X4582 2115X1145	27.73	1.8036308588064

13239	114473	2748622	2863095	96.0017743036819	40X 10X 2.5X 0.6X	156480X74993 39120X18748 9780X4687 2445X1171	32.79	1.31101820582272	
13083	115070	3099325	3214395	96.4201661587951	40X 10X 2.5X 0.6X	175680X74956 43920X18739 10980X4684 2745X1171	36.79	1.3170208701793	
13303	129551	1579009	1708560	92.4175328931966	40X 10X 2.5X 0.6X	161280X434114032 0X10852 10080X2713 2520X678	19.56	1.48313056609074	
11054	258580	1568060	1826640	85.8439539263347	40X 10X 2.5X 1.25X	23040X81184 5760X20296 1440X5074 720X2537	5.23	0.740361209652696	
Avg.	190828.133333333	3240275.86666667	3431104	93.1628727883602	-	-	22.2026666666667	1.51802456437634	
Total								333.04	21.0343542630478

## A.2 XML file for one of the whole virtual slides annotation

```
<All_Annotations>
  <Annotations>
    <Annotation Color="F4FA58" Name="Annotation 0" PartOfGroup="G3" Type="Polygon">
      <Coordinates>
        <Coordinate Order="0" X="13534.0" Y="65898.0"/>
        <Coordinate Order="1" X="15210.0" Y="65898.0"/>
        <Coordinate Order="2" X="15210.0" Y="68126.0"/>
        <Coordinate Order="3" X="13534.0" Y="68126.0"/>
      </Coordinates>
    </Annotation>
    <Annotation Color="F4FA58" Name="Annotation 1" PartOfGroup="G5" Type="Polygon">
      <Coordinates>
        <Coordinate Order="0" X="24259.0" Y="61139.0"/>
        <Coordinate Order="1" X="26645.0" Y="61139.0"/>
        <Coordinate Order="2" X="26645.0" Y="62311.0"/>
        <Coordinate Order="3" X="24259.0" Y="62311.0"/>
      </Coordinates>
    </Annotation>
  </Annotations>
  <AnnotationGroups>
    <Group Color="#0000ff" Name="G3" PartOfGroup="None">
      <Attributes/>
    </Group>
    <Group Color="#ff0000" Name="G5" PartOfGroup="None">
      <Attributes/>
    </Group>
  </AnnotationGroups>
</All_Annotations>
```

## Appendix B

### Materials for Chapters 4 and 5

#### B.1 A summary of the structure of Keras “Inception-ResNet-v2.”

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 256, 256, 3)	0	
conv2d_1 (Conv2D)	(None, 127, 127, 32)	864	input_1[0][0]
batch_normalization_1(BatchNo)	(None, 127, 127, 32)	96	conv2d_1[0][0]
activation_1 (Activation)	(None, 127, 127, 32)	0	batch_normalization_1[0][0]
conv2d_2 (Conv2D)	(None, 125, 125, 32)	9216	activation_1[0][0]
batch_normalization_2(BatchNo)	(None, 125, 125, 32)	96	conv2d_2[0][0]
activation_2 (Activation)	(None, 125, 125, 32)	0	batch_normalization_2[0][0]
conv2d_3 (Conv2D)	(None, 125, 125, 64)	18432	activation_2[0][0]
batch_normalization_3(BatchNo)	(None, 125, 125, 64)	192	conv2d_3[0][0]
activation_3 (Activation)	(None, 125, 125, 64)	0	batch_normalization_3[0][0]
max_pooling2d_1 (MaxPooling2D)	(None, 62, 62, 64)	0	activation_3[0][0]
conv2d_4 (Conv2D)	(None, 62, 62, 80)	5120	max_pooling2d_1[0][0]
batch_normalization_4(BatchNo)	(None, 62, 62, 80)	240	conv2d_4[0][0]
activation_4 (Activation)	(None, 62, 62, 80)	0	batch_normalization_4[0][0]
conv2d_5 (Conv2D)	(None, 60, 60, 192)	138240	activation_4[0][0]
batch_normalization_5(BatchNo)	(None, 60, 60, 192)	576	conv2d_5[0][0]
activation_5 (Activation)	(None, 60, 60, 192)	0	batch_normalization_5[0][0]
max_pooling2d_2 (MaxPooling2D)	(None, 29, 29, 192)	0	activation_5[0][0]
conv2d_9 (Conv2D)	(None, 29, 29, 64)	12288	max_pooling2d_2[0][0]
batch_normalization_9(BatchNo)	(None, 29, 29, 64)	192	conv2d_9[0][0]
activation_9 (Activation)	(None, 29, 29, 64)	0	batch_normalization_9[0][0]
conv2d_7 (Conv2D)	(None, 29, 29, 48)	9216	max_pooling2d_2[0][0]
conv2d_10 (Conv2D)	(None, 29, 29, 96)	55296	activation_9[0][0]
batch_normalization_7(BatchNo)	(None, 29, 29, 48)	144	conv2d_7[0][0]
batch_normalization_10(BatchNo)	(None, 29, 29, 96)	288	conv2d_10[0][0]
activation_7 (Activation)	(None, 29, 29, 48)	0	batch_normalization_7[0][0]
activation_10 (Activation)	(None, 29, 29, 96)	0	batch_normalization_10[0][0]
average_pooling2d_1 (AveragePoo)	(None, 29, 29, 192)	0	max_pooling2d_2[0][0]
conv2d_6 (Conv2D)	(None, 29, 29, 96)	18432	max_pooling2d_2[0][0]
conv2d_8 (Conv2D)	(None, 29, 29, 64)	76800	activation_7[0][0]
conv2d_11 (Conv2D)	(None, 29, 29, 96)	82944	activation_10[0][0]
conv2d_12 (Conv2D)	(None, 29, 29, 64)	12288	average_pooling2d_1[0][0]
batch_normalization_6(BatchNo)	(None, 29, 29, 96)	288	conv2d_6[0][0]
batch_normalization_8(BatchNo)	(None, 29, 29, 64)	192	conv2d_8[0][0]
batch_normalization_11(BatchNo)	(None, 29, 29, 96)	288	conv2d_11[0][0]
batch_normalization_12(BatchNo)	(None, 29, 29, 64)	192	conv2d_12[0][0]
activation_6 (Activation)	(None, 29, 29, 96)	0	batch_normalization_6[0][0]
activation_8 (Activation)	(None, 29, 29, 64)	0	batch_normalization_8[0][0]
activation_11 (Activation)	(None, 29, 29, 96)	0	batch_normalization_11[0][0]
activation_12 (Activation)	(None, 29, 29, 64)	0	batch_normalization_12[0][0]
mixed_5b (Concatenate)	(None, 29, 29, 320)	0	activation_6[0][0] activation_8[0][0] activation_11[0][0] activation_12[0][0]

Stem(end)

				Inception-resnet-A(start) Block#1
conv2d_16 (Conv2D)	(None, 29, 29, 32)	10240		mixed_5b[0][0]
batch_normalization_16 (BatchNo)	(None, 29, 29, 32)	96		conv2d_16[0][0]
activation_16 (Activation)	(None, 29, 29, 32)	0		batch_normalization_16[0][0]
conv2d_14 (Conv2D)	(None, 29, 29, 32)	10240		mixed_5b[0][0]
conv2d_17 (Conv2D)	(None, 29, 29, 48)	13824		activation_16[0][0]
batch_normalization_14 (BatchNo)	(None, 29, 29, 32)	96		conv2d_14[0][0]
batch_normalization_17 (BatchNo)	(None, 29, 29, 48)	144		conv2d_17[0][0]
activation_14 (Activation)	(None, 29, 29, 32)	0		batch_normalization_14[0][0]
activation_17 (Activation)	(None, 29, 29, 48)	0		batch_normalization_17[0][0]
conv2d_13 (Conv2D)	(None, 29, 29, 32)	10240		mixed_5b[0][0]
conv2d_15 (Conv2D)	(None, 29, 29, 32)	9216		activation_14[0][0]
conv2d_18 (Conv2D)	(None, 29, 29, 64)	27648		activation_17[0][0]
batch_normalization_13 (BatchNo)	(None, 29, 29, 32)	96		conv2d_13[0][0]
batch_normalization_15 (BatchNo)	(None, 29, 29, 32)	96		conv2d_15[0][0]
batch_normalization_18 (BatchNo)	(None, 29, 29, 64)	192		conv2d_18[0][0]
activation_13 (Activation)	(None, 29, 29, 32)	0		batch_normalization_13[0][0]
activation_15 (Activation)	(None, 29, 29, 32)	0		batch_normalization_15[0][0]
activation_18 (Activation)	(None, 29, 29, 64)	0		batch_normalization_18[0][0]
block35_1_mixed (Concatenate)	(None, 29, 29, 128)	0		activation_13[0][0] activation_15[0][0] activation_18[0][0]
block35_1_conv (Conv2D)	(None, 29, 29, 320)	41280		block35_1_mixed[0][0]
block35_1 (Lambda)	(None, 29, 29, 320)	0		mixed_5b[0][0] block35_1_conv[0][0]
block35_1_ac (Activation)	(None, 29, 29, 320)	0		block35_1[0][0]
				Inception-resnet-A(end) Block#1
: : Inception-resnet-A Block#2,3,4,5,6,7,8 and 9 : :				
				Inception-resnet-A(start) Block#10
conv2d_70 (Conv2D)	(None, 29, 29, 32)	10240		block35_9_ac[0][0]
batch_normalization_70 (BatchNo)	(None, 29, 29, 32)	96		conv2d_70[0][0]
activation_70 (Activation)	(None, 29, 29, 32)	0		batch_normalization_70[0][0]
conv2d_68 (Conv2D)	(None, 29, 29, 32)	10240		block35_9_ac[0][0]
conv2d_71 (Conv2D)	(None, 29, 29, 48)	13824		activation_70[0][0]
batch_normalization_68 (BatchNo)	(None, 29, 29, 32)	96		conv2d_68[0][0]
batch_normalization_71 (BatchNo)	(None, 29, 29, 48)	144		conv2d_71[0][0]
activation_68 (Activation)	(None, 29, 29, 32)	0		batch_normalization_68[0][0]
activation_71 (Activation)	(None, 29, 29, 48)	0		batch_normalization_71[0][0]
conv2d_67 (Conv2D)	(None, 29, 29, 32)	10240		block35_9_ac[0][0]
conv2d_69 (Conv2D)	(None, 29, 29, 32)	9216		activation_68[0][0]
conv2d_72 (Conv2D)	(None, 29, 29, 64)	27648		activation_71[0][0]
batch_normalization_67 (BatchNo)	(None, 29, 29, 32)	96		conv2d_67[0][0]
batch_normalization_69 (BatchNo)	(None, 29, 29, 32)	96		conv2d_69[0][0]
batch_normalization_72 (BatchNo)	(None, 29, 29, 64)	192		conv2d_72[0][0]
activation_67 (Activation)	(None, 29, 29, 32)	0		batch_normalization_67[0][0]



activation_69 (Activation)	(None, 29, 29, 32)	0	batch_normalization_69[0][0]
activation_72 (Activation)	(None, 29, 29, 64)	0	batch_normalization_72[0][0]
block35_10_mixed (Concatenate)	(None, 29, 29, 128)	0	activation_67[0][0] activation_69[0][0] activation_72[0][0]
block35_10_conv (Conv2D)	(None, 29, 29, 320)	41280	block35_10_mixed[0][0]
block35_10 (Lambda)	(None, 29, 29, 320)	0	block35_9_ac[0][0] block35_10_conv[0][0]
block35_10_ac (Activation)	(None, 29, 29, 320)	0	block35_10[0][0]
<b>Inception-resnet-A(end) Block#10</b>			
<b>Reduction A(start)</b>			
conv2d_74 (Conv2D)	(None, 29, 29, 256)	81920	block35_10_ac[0][0]
batch_normalization_74 (BatchNo)	(None, 29, 29, 256)	768	conv2d_74[0][0]
activation_74 (Activation)	(None, 29, 29, 256)	0	batch_normalization_74[0][0]
conv2d_75 (Conv2D)	(None, 29, 29, 256)	589824	activation_74[0][0]
batch_normalization_75 (BatchNo)	(None, 29, 29, 256)	768	conv2d_75[0][0]
activation_75 (Activation)	(None, 29, 29, 256)	0	batch_normalization_75[0][0]
conv2d_73 (Conv2D)	(None, 14, 14, 384)	1105920	block35_10_ac[0][0]
conv2d_76 (Conv2D)	(None, 14, 14, 384)	884736	activation_75[0][0]
batch_normalization_73 (BatchNo)	(None, 14, 14, 384)	1152	conv2d_73[0][0]
batch_normalization_76 (BatchNo)	(None, 14, 14, 384)	1152	conv2d_76[0][0]
activation_73 (Activation)	(None, 14, 14, 384)	0	batch_normalization_73[0][0]
activation_76 (Activation)	(None, 14, 14, 384)	0	batch_normalization_76[0][0]
max_pooling2d_3 (MaxPooling2D)	(None, 14, 14, 320)	0	block35_10_ac[0][0]
mixed_6a (Concatenate)	(None, 14, 14, 1088)	0	activation_73[0][0] activation_76[0][0] max_pooling2d_3[0][0]
<b>Reduction A(end)</b>			
<b>Inception-resnet-B(Start) Block#1</b>			
conv2d_78 (Conv2D)	(None, 14, 14, 128)	139264	mixed_6a[0][0]
batch_normalization_78 (BatchNo)	(None, 14, 14, 128)	384	conv2d_78[0][0]
activation_78 (Activation)	(None, 14, 14, 128)	0	batch_normalization_78[0][0]
conv2d_79 (Conv2D)	(None, 14, 14, 160)	143360	activation_78[0][0]
batch_normalization_79 (BatchNo)	(None, 14, 14, 160)	480	conv2d_79[0][0]
activation_79 (Activation)	(None, 14, 14, 160)	0	batch_normalization_79[0][0]
conv2d_77 (Conv2D)	(None, 14, 14, 192)	208896	mixed_6a[0][0]
conv2d_80 (Conv2D)	(None, 14, 14, 192)	215040	activation_79[0][0]
batch_normalization_77 (BatchNo)	(None, 14, 14, 192)	576	conv2d_77[0][0]
batch_normalization_80 (BatchNo)	(None, 14, 14, 192)	576	conv2d_80[0][0]
activation_77 (Activation)	(None, 14, 14, 192)	0	batch_normalization_77[0][0]
activation_80 (Activation)	(None, 14, 14, 192)	0	batch_normalization_80[0][0]
block17_1_mixed (Concatenate)	(None, 14, 14, 384)	0	activation_77[0][0] activation_80[0][0]
block17_1_conv (Conv2D)	(None, 14, 14, 1088)	418880	block17_1_mixed[0][0]
block17_1 (Lambda)	(None, 14, 14, 1088)	0	mixed_6a[0][0] block17_1_conv[0][0]
block17_1_ac (Activation)	(None, 14, 14, 1088)	0	block17_1[0][0]
<b>Inception-resnet-B(end) Block#1</b>			

: Inception-resnet-B Block#2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18 and 19 :				
<b>Inception-resnet-B(Start) Block#20</b>				
conv2d_154 (Conv2D)	(None, 14, 14, 128)	139264		block17_19_ac[0][0]
batch_normalization_154 (BatchN	(None, 14, 14, 128)	384		conv2d_154[0][0]
activation_154 (Activation)	(None, 14, 14, 128)	0		batch_normalization_154[0][0]
conv2d_155 (Conv2D)	(None, 14, 14, 160)	143360		activation_154[0][0]
batch_normalization_155 (BatchN	(None, 14, 14, 160)	480		conv2d_155[0][0]
activation_155 (Activation)	(None, 14, 14, 160)	0		batch_normalization_155[0][0]
conv2d_153 (Conv2D)	(None, 14, 14, 192)	208896		block17_19_ac[0][0]
conv2d_156 (Conv2D)	(None, 14, 14, 192)	215040		activation_155[0][0]
batch_normalization_153 (BatchN	(None, 14, 14, 192)	576		conv2d_153[0][0]
batch_normalization_156 (BatchN	(None, 14, 14, 192)	576		conv2d_156[0][0]
activation_153 (Activation)	(None, 14, 14, 192)	0		batch_normalization_153[0][0]
activation_156 (Activation)	(None, 14, 14, 192)	0		batch_normalization_156[0][0]
block17_20_mixed (Concatenate)	(None, 14, 14, 384)	0		activation_153[0][0] activation_156[0][0]
block17_20_conv (Conv2D)	(None, 14, 14, 1088)	418880		block17_20_mixed[0][0]
block17_20 (Lambda)	(None, 14, 14, 1088)	0		block17_19_ac[0][0] block17_20_conv[0][0]
block17_20_ac (Activation)	(None, 14, 14, 1088)	0		block17_20[0][0]
<b>Inception-resnet-B(end) Block#20</b>				
<b>Reduction-B(start)</b>				
conv2d_161 (Conv2D)	(None, 14, 14, 256)	278528		block17_20_ac[0][0]
batch_normalization_161 (BatchN	(None, 14, 14, 256)	768		conv2d_161[0][0]
activation_161 (Activation)	(None, 14, 14, 256)	0		batch_normalization_161[0][0]
conv2d_157 (Conv2D)	(None, 14, 14, 256)	278528		block17_20_ac[0][0]
conv2d_159 (Conv2D)	(None, 14, 14, 256)	278528		block17_20_ac[0][0]
conv2d_162 (Conv2D)	(None, 14, 14, 288)	663552		activation_161[0][0]
batch_normalization_157 (BatchN	(None, 14, 14, 256)	768		conv2d_157[0][0]
batch_normalization_159 (BatchN	(None, 14, 14, 256)	768		conv2d_159[0][0]
batch_normalization_162 (BatchN	(None, 14, 14, 288)	864		conv2d_162[0][0]
activation_157 (Activation)	(None, 14, 14, 256)	0		batch_normalization_157[0][0]
activation_159 (Activation)	(None, 14, 14, 256)	0		batch_normalization_159[0][0]
activation_162 (Activation)	(None, 14, 14, 288)	0		batch_normalization_162[0][0]
conv2d_158 (Conv2D)	(None, 6, 6, 384)	884736		activation_157[0][0]
conv2d_160 (Conv2D)	(None, 6, 6, 288)	663552		activation_159[0][0]
conv2d_163 (Conv2D)	(None, 6, 6, 320)	829440		activation_162[0][0]
batch_normalization_158 (BatchN	(None, 6, 6, 384)	1152		conv2d_158[0][0]
batch_normalization_160 (BatchN	(None, 6, 6, 288)	864		conv2d_160[0][0]
batch_normalization_163 (BatchN	(None, 6, 6, 320)	960		conv2d_163[0][0]
activation_158 (Activation)	(None, 6, 6, 384)	0		batch_normalization_158[0][0]
activation_160 (Activation)	(None, 6, 6, 288)	0		batch_normalization_160[0][0]
activation_163 (Activation)	(None, 6, 6, 320)	0		batch_normalization_163[0][0]
max_pooling2d_4 (MaxPooling2D)	(None, 6, 6, 1088)	0		block17_20_ac[0][0]



mixed_7a (Concatenate)	(None, 6, 6, 2080)	0	activation_158[0][0] activation_160[0][0] activation_163[0][0] max_pooling2d_4[0][0]
			<b>Reduction-B(end)</b>
			<b>Inception-resnet-C(start) Block#1</b>
conv2d_165 (Conv2D)	(None, 6, 6, 192)	399360	mixed_7a[0][0]
batch_normalization_165 (BatchN	(None, 6, 6, 192)	576	conv2d_165[0][0]
activation_165 (Activation)	(None, 6, 6, 192)	0	batch_normalization_165[0][0]
conv2d_166 (Conv2D)	(None, 6, 6, 224)	129024	activation_165[0][0]
batch_normalization_166 (BatchN	(None, 6, 6, 224)	672	conv2d_166[0][0]
activation_166 (Activation)	(None, 6, 6, 224)	0	batch_normalization_166[0][0]
conv2d_164 (Conv2D)	(None, 6, 6, 192)	399360	mixed_7a[0][0]
conv2d_167 (Conv2D)	(None, 6, 6, 256)	172032	activation_166[0][0]
batch_normalization_164 (BatchN	(None, 6, 6, 192)	576	conv2d_164[0][0]
batch_normalization_167 (BatchN	(None, 6, 6, 256)	768	conv2d_167[0][0]
activation_164 (Activation)	(None, 6, 6, 192)	0	batch_normalization_164[0][0]
activation_167 (Activation)	(None, 6, 6, 256)	0	batch_normalization_167[0][0]
block8_1_mixed (Concatenate)	(None, 6, 6, 448)	0	activation_164[0][0] activation_167[0][0]
block8_1_conv (Conv2D)	(None, 6, 6, 2080)	933920	block8_1_mixed[0][0]
block8_1 (Lambda)	(None, 6, 6, 2080)	0	mixed_7a[0][0] block8_1_conv[0][0]
block8_1_ac (Activation)	(None, 6, 6, 2080)	0	block8_1[0][0]
			<b>Inception-resnet-C(end) Block#1</b>
<div style="border: 1px solid blue; padding: 10px; margin: 10px auto; width: fit-content;"> <p style="text-align: center;">                     .                      .                      Inception-resnet-C Block#2,3,4,5,6 and 7                      .                      .                 </p> </div>			
			<b>Inception-resnet-C(start) Block#8</b>
conv2d_193 (Conv2D)	(None, 6, 6, 192)	399360	block8_7_ac[0][0]
batch_normalization_193 (BatchN	(None, 6, 6, 192)	576	conv2d_193[0][0]
activation_193 (Activation)	(None, 6, 6, 192)	0	batch_normalization_193[0][0]
conv2d_194 (Conv2D)	(None, 6, 6, 224)	129024	activation_193[0][0]
batch_normalization_194 (BatchN	(None, 6, 6, 224)	672	conv2d_194[0][0]
activation_194 (Activation)	(None, 6, 6, 224)	0	batch_normalization_194[0][0]
conv2d_192 (Conv2D)	(None, 6, 6, 192)	399360	block8_7_ac[0][0]
conv2d_195 (Conv2D)	(None, 6, 6, 256)	172032	activation_194[0][0]
batch_normalization_192 (BatchN	(None, 6, 6, 192)	576	conv2d_192[0][0]
batch_normalization_195 (BatchN	(None, 6, 6, 256)	768	conv2d_195[0][0]
activation_192 (Activation)	(None, 6, 6, 192)	0	batch_normalization_192[0][0]
activation_195 (Activation)	(None, 6, 6, 256)	0	batch_normalization_195[0][0]
block8_8_mixed (Concatenate)	(None, 6, 6, 448)	0	activation_192[0][0] activation_195[0][0]
block8_8_conv (Conv2D)	(None, 6, 6, 2080)	933920	block8_8_mixed[0][0]
block8_8 (Lambda)	(None, 6, 6, 2080)	0	block8_7_ac[0][0] block8_8_conv[0][0]
block8_8_ac (Activation)	(None, 6, 6, 2080)	0	block8_8[0][0]
(the end of the non-trainable parameters)			<b>Inception-resnet-C(end) Block#8</b>

Trainable parameters start from Block#9!			Inception-resnet-C(start) Block#9
conv2d_197 (Conv2D)	(None, 6, 6, 192)	399360	block8_8_ac[0][0]
batch_normalization_197 (Batch Normalization)	(None, 6, 6, 192)	576	conv2d_197[0][0]
activation_197 (Activation)	(None, 6, 6, 192)	0	batch_normalization_197[0][0]
conv2d_198 (Conv2D)	(None, 6, 6, 224)	129024	activation_197[0][0]
batch_normalization_198 (Batch Normalization)	(None, 6, 6, 224)	672	conv2d_198[0][0]
activation_198 (Activation)	(None, 6, 6, 224)	0	batch_normalization_198[0][0]
conv2d_196 (Conv2D)	(None, 6, 6, 192)	399360	block8_8_ac[0][0]
conv2d_199 (Conv2D)	(None, 6, 6, 256)	172032	activation_198[0][0]
batch_normalization_196 (Batch Normalization)	(None, 6, 6, 192)	576	conv2d_196[0][0]
batch_normalization_199 (Batch Normalization)	(None, 6, 6, 256)	768	conv2d_199[0][0]
activation_196 (Activation)	(None, 6, 6, 192)	0	batch_normalization_196[0][0]
activation_199 (Activation)	(None, 6, 6, 256)	0	batch_normalization_199[0][0]
block8_9_mixed (Concatenate)	(None, 6, 6, 448)	0	activation_196[0][0] activation_199[0][0]
block8_9_conv (Conv2D)	(None, 6, 6, 2080)	933920	block8_9_mixed[0][0]
block8_9 (Lambda)	(None, 6, 6, 2080)	0	block8_8_ac[0][0] block8_9_conv[0][0]
block8_9_ac (Activation)	(None, 6, 6, 2080)	0	block8_9[0][0]
			Inception-resnet-C(end) Block#9
			Inception-resnet-C(start) Block#10
conv2d_201 (Conv2D)	(None, 6, 6, 192)	399360	block8_9_ac[0][0]
batch_normalization_201 (Batch Normalization)	(None, 6, 6, 192)	576	conv2d_201[0][0]
activation_201 (Activation)	(None, 6, 6, 192)	0	batch_normalization_201[0][0]
conv2d_202 (Conv2D)	(None, 6, 6, 224)	129024	activation_201[0][0]
batch_normalization_202 (Batch Normalization)	(None, 6, 6, 224)	672	conv2d_202[0][0]
activation_202 (Activation)	(None, 6, 6, 224)	0	batch_normalization_202[0][0]
conv2d_200 (Conv2D)	(None, 6, 6, 192)	399360	block8_9_ac[0][0]
conv2d_203 (Conv2D)	(None, 6, 6, 256)	172032	activation_202[0][0]
batch_normalization_200 (Batch Normalization)	(None, 6, 6, 192)	576	conv2d_200[0][0]
batch_normalization_203 (Batch Normalization)	(None, 6, 6, 256)	768	conv2d_203[0][0]
activation_200 (Activation)	(None, 6, 6, 192)	0	batch_normalization_200[0][0]
activation_203 (Activation)	(None, 6, 6, 256)	0	batch_normalization_203[0][0]
block8_10_mixed (Concatenate)	(None, 6, 6, 448)	0	activation_200[0][0] activation_203[0][0]
block8_10_conv (Conv2D)	(None, 6, 6, 2080)	933920	block8_10_mixed[0][0]
block8_10 (Lambda)	(None, 6, 6, 2080)	0	block8_9_ac[0][0] block8_10_conv[0][0]
			Inception-resnet-C(end) Block#10
conv_7b (Conv2D)	(None, 6, 6, 1536)	3194880	block8_10[0][0]
conv_7b_bn (Batch Normalization)	(None, 6, 6, 1536)	4608	conv_7b[0][0]
conv_7b_ac (Activation)	(None, 6, 6, 1536)	0	conv_7b_bn[0][0]
global_average_pooling2d_1 (Global Average Pooling)	(None, 1536)	0	conv_7b_ac[0][0]
dense_1 (Dense)	(None, 768)	1180416	global_average_pooling2d_1[0][0]
dense_2 (Dense)	(None, 2)	1538	dense_1[0][0]
Total params: 55,518,690			
Trainable params: 7,346,850			
Non-trainable params: 48,171,840			



## Appendix C

### Materials for Chapter 6

#### C.1 Visualised prediction maps for the remainder of test slides using the proposed CAD system

In the figure, the test slides are as follows: (1)“11014.svs”, (2)“11013.svs”, (3)“11054.svs”, (4)“13083.svs”, (5)“10586.svs”, (6)“10829.svs”, (7)“10857”, (8)“11035.svs”, (9)“13303.svs”, (10)“13348.svs” and (11)“13239.svs”. All the slides in the figure are correctly classified. Slides in (2) and (3) have LGD labels, slides in (4), (9), (10) and (11) have HGD labels, and the remainder slides have NFD labels.

