D 2022

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

CANCER DIAGNOSIS IN DIGITAL PATHOLOGY: LEARNING FROM LABEL SCARCITY

SARA ISABEL PIRES DE OLIVEIRA
DOCTORAL THESIS PRESENTED TO
FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO IN
ELECTRICAL AND COMPUTER ENGINEERING

Cancer diagnosis in digital pathology: learning from label scarcity

Sara Isabel Pires de Oliveira

Doctoral Program in Electrical and Computer Engineering

Supervisor: Hélder Filipe Pinto de Oliveira, PhD

Co-Supervisor: Jaime dos Santos Cardoso, PhD

Co-Supervisor: Maria João de Viseu Botelho Cardoso Ayres de Campos, MD PhD

December, 2022

Cancer diagnosis in digital pathology: learning from label scarcity

Sara Isabel Pires de Oliveira

Doctoral Program in Electrical and Computer Engineering

Approved in public examination by the Jury:

President:	Professor Luís Miguel Pinho de Almeida
Referee:	Professor José Luis Alba Castro
Referee:	Professor Pétia Georgieva Georgieva
Referee:	Professor Fernando Carlos de Landér Schmitt
Referee:	Professor Miguel Tavares Coimbra
Referee:	Professor Jaime dos Santos Cardoso

Definitive version validated by the Supervisor:

Hélder Filipe Pinto de Oliveira, PhD

December, 2022

Resumo

Com o aumento de novos casos todos os anos, o cancro continuará a ser a maior causa de morte a nível mundial. Apesar de os cancros da mama, colo-rectal e cervical estarem entre os 10 mais prevalentes, a taxa de mortalidade pode ser reduzida através de mais rastreio, detecção precoce, e melhores abordagens de tratamento. Embora as técnicas de imagem, tais como a radiologia, permitam a detecção e gestão de casos de cancro a nível anatómico, devem ser feitos mais testes para avaliar a natureza das lesões. Uma vez que o cancro é uma doença heterogénea, através da análise das propriedades morfológicas dos tecidos das amostras tumorais, feito pela patologia, permite a identificação do subtipo e grau da doença, bem como a resposta expectável à terapêutica.

Hoje em dia, as amostras histológicas antes vistas ao microscópio, são agora convertidas em imagens de alta resolução utilizando um processo chamado *whole-slide imaging*. Assim, à medida que mais laboratórios adoptam um fluxo de trabalho digital, as lâminas digitalizadas estão a tornar-se cada vez mais acessíveis. Para além dos benefícios para a prática clínica, a patologia digital abriu inúmeras oportunidades de investigação na área de visão por computador. A complexidade da avaliação de amostras de histopatologia apresenta novos desafios no desenvolvimento de sistemas de processamento automático de imagem. Contudo, normalmente, estes modelos precisam de ser treinados de uma forma supervisionada, o que torna necessária uma anotação detalhada fornecida por patologistas. Com a actual global falta de patologistas, que têm uma carga de trabalho cada vez maior com o aumento dos programas de rastreio e das taxas de incidência de cancro, estas anotações são ainda mais difíceis de obter, o que motiva o desenvolvimento de modelos de *machine learning* que consigam aprender com pouca supervisão.

O principal objectivo deste projecto de doutoramento é desenvolver sistemas de diagnóstico assistido por computador para patologia computacional, sem a necessidade de dados anotados com muito detalhe, contribuindo para o desenvolvimento de métodos de *weakly-supervised learning*. O trabalho desenvolvido centra-se em modelos de diagnóstico para cancro colo-rectal, cervical e cancro da mama, com contribuições como: um estudo de viabilidade sobre o uso de dados parcialmente anotados para melhorar o desenvolvimento de ferramentas de diagnóstico assistido por computador directamente a partir de lâminas digitalizadas; uma estratégia semi-supervisionada para diagnóstico automático de cancro colo-rectal, com a capacidade de focar a atenção dos patologistas para as áreas mais relevantes dos tecidos; um protótipo de software clínico para classificação e mapeamento de tecidos em amostras colorrectais; uma abordagem *weakly supervised* que segmenta regiões de interesse e classifica displasia cervical; e finalmente, o primeiro trabalho sobre classificação da sobreexpressão de HER2 em amostras de lesões da mama coradas com hematoxilina e eosina, sem necessidade de anotações ao nível do *pixel*.

Por fim, existem ainda muitos desafios a resolver na área da patologia computacional para que os modelos de *machine learning* possam efectiva e extensamente aplicados na clínica. No entanto, espera-se que este trabalho represente um passo em frente no caminho que ainda falta percorrer.

Palavras-chave: Patologia computacional, *whole-slide image*, *weakly-supervised learning*, *deep learning*, visão por computador, cancro colorectal, cancro cervical, cancro da mama

Abstract

With increasing new cases every year, cancer will continue to be the largest cause of death worldwide. Despite breast, colorectal, and cervical cancers ranking among the 10 most prevalent ones, the associated death rate can be greatly reduced by screening, earlier detection, and better treatment approaches. While imaging techniques, such as radiology, enable the detection and management of cancer cases at the anatomical level, further testing should be done to assess the nature of the abnormalities. Since cancer is a heterogeneous illness, by analysing the tissue morphological properties of tumour specimens, the examination of thin histological tissue sections, done by pathology, enables the identification of the disease sub-type, grade, and also therapy responsiveness.

Nowadays, the histological samples mounted on glass slides before seen in the microscope, are now converted into large, high-resolution images using a process called whole slide imaging. Thus, as more laboratories adopt a digital workflow, digitised slides are becoming more and more accessible. Beyond the benefits for clinical practice, digital pathology has opened up numerous prospects for study in the field of computer vision. The complexity of pathology assessment presents fresh difficulties for cutting-edge automatic image processing systems. However, usually, these models need to be trained in a supervised manner, which renders the necessity of having detailed annotation provided by the experts. With the current global lack of pathologists, that have increased workloads with the rise of screening programmes and cancer incidence rates, combined with the huge size of the images, these annotations are even more difficult to obtain, which motivates the development of machine learning models that can learn with little supervision.

The primary objective of this doctoral project is to design computer-aided diagnosis systems for computational pathology, without the requirement for much detailed annotated data, contributing to the development of weakly-supervised learning methods. The developed work is focused on models for colorectal, cervical and breast cancer diagnosis, with contributions to the field as: a feasibility study on leveraging partially annotated datasets to drive the development of computer-aided diagnosis tools for digital pathology, directly from whole-side images; a semi-supervised strategy for colorectal cancer automatic diagnosis, with the capability to guide pathologists' attention towards the most relevant tissue areas; an AI-based clinical software prototype for grading and tissue mapping in colorectal samples; a weakly-supervised approach that segments regions of interest and grades cervical dysplasia form there; and finally, the first published work on the classification of HER2 overexpression status on haematoxylin and eosin stained breast cancer slides, without the need for pixel-level annotations.

In the end, there are still many obstacles to overcome in the field of computational pathology, so machine learning models can effectively get closer to clinical applicability. However, this work is expected to be a step forward in the path that is still left.

Keywords: Computational pathology, whole-slide image, weakly-supervised learning, deep learning, computer vision, colorectal cancer, cervical cancer, breast cancer

Agradecimentos

Em primeiro lugar, gostaria de agradecer aos meu orientadores, pessoas indispensáveis no desenvolvimento desta tese. Ao Hélder Oliveira, por me ter dado a oportunidade de me juntar ao VCMI, por me ter desafiado a fazer o doutoramento e por me ter acompanhado ao longo destes anos. À Dra. Maria João Cardoso por me ensinar tanto sobre o mundo clínico e por ser sempre tão interessante e inspirador falar com ela. Ao Prof. Jaime Cardoso, por ser um exemplo de rigor científico, pelo acompanhamento muito próximo do meu trabalho, por estar sempre disponível para responder às minhas dúvidas e por me ter dado asas para abraçar desafios cada vez maiores. Obrigada pelo apoio, incentivo e por toda a confiança que depositaram em mim!

Agradeço ainda à FCT, pelo financiamento deste trabalho, ao INESC TEC, por ter sido a instituição de acolhimento do meu doutoramento e à FEUP, pela oferta do programa doutoral.

Agradeço a toda a equipa do laboratório IMP Diagnostics, que esteve envolvida no projeto CADpath.AI, por ter sido tão fácil trabalhar convosco e por partilharem do espirito de fazer sempre mais e melhor. Em especial, agradeço à Dra. Isabel Macedo Pinto pelo carinho e pela sua supervisão sempre muito atenta e curiosa. E, claro, à Diana Montezuma e à Ana Monteiro, que rapidamente se tornaram mais do que colegas, com quem gosto tanto de trabalhar (e jantar!) e que alinham sempre nas minhas ideias com o maior entusiasmo. Agradeço também à Dra. Rita Canas-Marques, da Fundação Champalimaud, por todo a simpatia e entusiasmo com que acolheu o meu trabalho, por todo o suporte e pelos desafios que me propôs.

A todos os meus colegas do VCMI, obrigada pelo companheirismo, pela excelente conduta de trabalho e por serem o melhor grupo de colegas que poderia ter tido. Todos, de alguma forma, contribuíram para esta tese: desde os que foram co-autores dos papers que publicámos, como todos os outros que me ajudaram a ver soluções, quer estando sentados ao meu lado a olhar para aquelas imagens gigantes pintadas a cor-de-rosa, quer nos simples "vamos lá a baixo beber um café?!". Aprendi imenso convosco! Obrigada aos que me abriram a porta no primeiro dia e me acolheram tão bem, aos que fizeram esta jornada do PhD em simultâneo comigo e a todos com quem fui partilhando parte dos últimos anos. Obrigada aos que passaram a estar no naipe dos melhores amigos e aos que me nutriram em tantas longas conversas. Obrigada VCMIs por vibrarem com as minhas conquistas e, mais ainda, por serem luz nos dias em que a minha se apagava.

A todos os meus amigos, aos mais recentes e aos de sempre, obrigada por fazerem parte da minha vida e por me terem escolhido para fazer parte da vossa. Obrigada por estarem sempre por perto, mesmo que com alguns kms de distância pelo meio.

Por fim, aos meus pais, João e Elisabete, e às minhas irmãs, Inês e Maria. O tanto que tenho de vos agradecer nunca caberá em palavra alguma. Vocês são os meus pilares! Obrigada pelo vosso amor incondicional! Ao resto da famíilia, obrigada por estarem sempre atentos à minha felicidade. Ao meu avô João, com imensa saudade!

A todos, um sincero obrigado!

*"An investment in knowledge
pays the best interest."*

Benjamin Franklin

Contents

List of Figures

List of Tables

Abbreviations

Acc	Accuracy
ASCO	American Society of Clinical Oncology
AUC	Area Under the Curve
BCa	Breast Cancer
CAD	Computer-Aided Diagnosis
CAP	College of American Pathologists
CCa	Cervical Cancer
CNN	Convolutional Neural Network
CRC	Colorectal Cancer
DCIS	Ductal Carcinoma in Situ
DL	Deep Learning
ER	Estrogen Receptors
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
H&E	Hematoxylin and Eosin
HER2	Human Epidermal growth factor Receptor - type 2
HGD	High-Grade Dysplasia
IDC	Invasive Ductal Carcinoma
IHC	Immunohistochemistry
ILC	Invasive Lobular Carcinoma
ISH	*In situ* Hybridisation
KNN	K-Nearest Neighbour
LCIS	Lobular Carcinoma in Situ
LGD	Low-Grade Dysplasia
MIL	Multiple Instance Learning
ML	Machine Learning
MLP	Multilayer Perceptron
NNeo	Non-Neoplastic
PR	Progesterone Receptors
QWK	Quadratic Weighted Kappa
RNN	Recurrent Neural Network
ROI	Region of Interest
SVM	Support Vector Machine
WSI	Whole Slide Image

Part I

Prologue

Chapter 1

Introduction

1.1 Context & background

Cancer is the leading cause of death worldwide, with almost 9.9 million deaths and around 18.1 million new cases, in 2020, of which breast, colorectal and cervical cancers are among the 10 most common ones. However, despite its increasing incidence trend, the mortality rate can be significantly decreased with screening, earlier detection and better treatment strategies [1].

While imaging techniques, such as radiology, enable the detection and management of cancer at the anatomical level, further testing should be done in order to assess the nature of the abnormalities. Since cancer is a heterogeneous disease, histopathology, the study of thin histological tissue sections of tumour *specimens*, allows the identification of disease sub-type, grade and, in some cases, treatment responsiveness, by assessing tissue morphological characteristics [2]. Thus, pathologists are responsible for the detailed diagnosis of samples collected during biopsies and surgery, determining the precise type and severity of tumours and thus playing a critical role in the management of cancer patients.

With the traditional pathology workflow, tissues are prepared and mounted in glass slides, which requires a staining step to highlight the different tissue structures. The standard staining technique uses a combination of haematoxylin and eosin (H&E) to highlight the nuclei and cytoplasm of cells: haematoxylin binds to DNA, dying the nuclei blue/purple and eosin binds to proteins and dyes other structures pink. A more advanced stain technique, immunohistochemistry (IHC), highlights the presence of specific antigens in the tissue, such as hormone receptors or cell proliferation factors [3]. At the moment, IHC is the standard technique often used to achieve a complete tumour diagnosis, which implies extra time and costs. At the moment, there are no identified morphological features on H&E slides that can be used for such evaluation.

Nowadays, with the transition of pathology labs to the digital era, histological glass slides are digitised into a large high-resolution image, through a technique known as whole slide imaging (WSI) or "virtual microscopy". Thus, WSI are becoming increasingly available, with more laboratories adopting a digital workflow [4–6]. And while this multi-step process requires an additional scanning step (Figure 1.1), the benefits far outweigh the increased initial overhead of these steps.

For example, access to archived cases, collaboration with external laboratories, and data sharing are all made easier. For example, the peer review of a WSI is completed at a quicker pace with a digital pathology workflow. In addition, the ability to easily access images mitigates the risk of errors, making diagnosis more auditable. In fact, due to technological advances, the digitisation of pathology data can support the work conducted by pathologists, enabling fastness, reproducibility and precision in diagnosis [2].

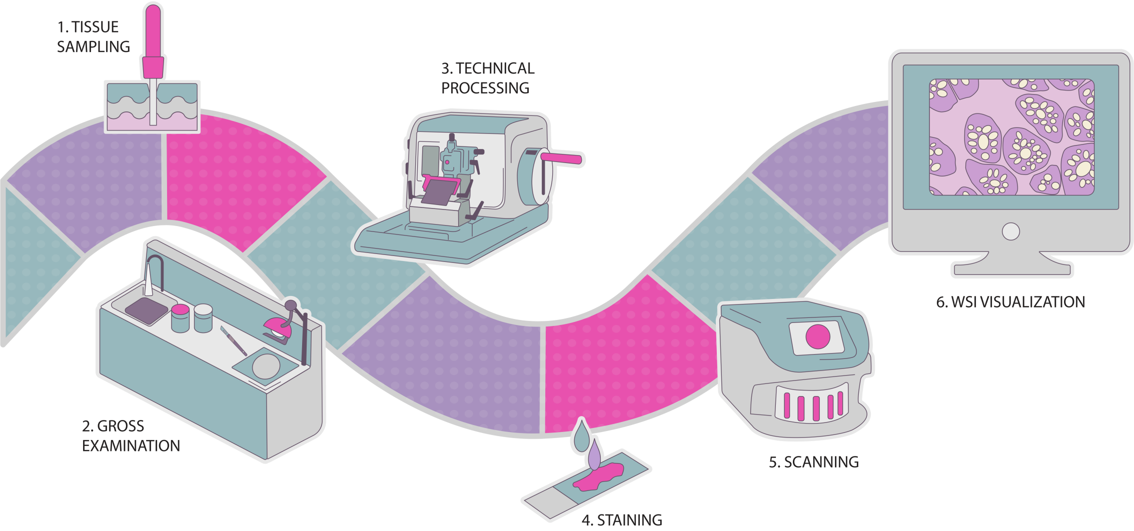


Figure 1.1: Digital pathology workflow, from collecting the biopsy sample to the WSI visualisation.

Over the last decade, the digitisation of histopathology images opened many research opportunities for the field of computer-aided image analysis [2, 3]. In fact, due to the high-resolution and complex nature of whole-slide image (WSI) evaluation, advances in image analysis are required, which provides the opportunity to apply and advance image processing techniques, as well as AI methodologies, such as machine learning (ML) and deep learning (DL) algorithms [2, 3, 7, 8]. Moreover, the integration of AI into healthcare routines is a required milestone for the years to come, and thus, in terms of pathology-focused research, many DL architectures have been applied with many different tasks in mind, either to predict diagnoses or even to identify new biomarkers [8–10].

Regarding the field of AI, and its application in computational pathology, DL models [11], which consist of multiple layers of processing to learn different levels of data representation, are the most common and promising methods nowadays. The networks are composed of multiple layers, each with multiple nodes. The large numbers of hidden layers confer depth to the networks, hence the name. Each node performs a weighted sum of its inputs and then feeds it into a non-linear function, the result of which is passed forward as input to the following layer and so on until the last layer, which provides the network output. In this way, these models have the intrinsic ability to learn features, directly from the input data, useful for the task at hand [11]. In particular, convolutional neural networks (CNN) are applied to images and automatically extract features, which are then used to identify objects/regions of interest or to classify the underlying diagnosis [12]. In digital pathology, this type of model is used, for example, for mitosis detection [13, 14], tissue segmentation [15, 16], cancer grading [17, 18] or histological classification [19, 20]. Additionally, there are also predictive systems that attempt to estimate the patient's probability of survival [21, 22].

Despite the popularity, clear potential, progress and good results of DL in computer vision, and medical imaging, in particular, researchers should carefully consider and manage its pros and cons [7, 23]. Indeed, digital pathology brings some specific challenges that need to be addressed:

- *High dimensionality of data*. Histology images are extremely informative, but at the cost of high dimensionality, usually over $50,000 \times 50,000$ pixels [23]. Hence, these images do not fit in the memory of a Graphics Processing Unit (GPU), which is usually needed to train DL models. Current methods either downsample the original image or extract multiple smaller patches, choosing between the cost of losing pixel information or losing spatial information, respectively;

- *Data variability*, due to the nearly infinite patterns resulting from the basic tissue types, and the lack of standardisation in tissue preparation, staining and scanning;

- *Lack of annotated data*, since extensive annotation is subjective, tedious, expensive and time-consuming;

- *Non-boolean diagnosis*, especially in difficult and rare cases, which makes the diagnosis process more complex;

- *Need for interpretability/explainability*, in order to be reliable, easily debugged, trusted and approved [7, 23, 24].

Therefore, the research community has the opportunity to develop robust algorithms with high performance, transparent and as interpretable as possible, always designed and validated in partnership with pathologists. To this end, one can take advantage of some well-known techniques such as transfer learning (using pre-trained networks instead of training from scratch), weakly/unsupervised learning (analysing images only with slide-level labelling), generative frameworks (by learning to generate images, the algorithm can understand their main distinctive features) or multitask learning (learning interrelated concepts may produce better generalisations) [23].

1.2 Motivation and objectives

Beyond the advantages for clinical practice, digital pathology has created many research opportunities in the computer vision field, with the complex nature of pathology assessment bringing new challenges to advanced automatic image processing systems. However, there is a global shortage of pathologists [25], that have increased workloads with the growth of screening programmes and cancer incidence rates. Therefore, the annotation usually required to train machine learning models represents an extra burden in pathologists' routine, making them even more challenging to obtain. How can this label scarcity be dealt with? How can pathology classification models be developed with robustness from lower supervision? How can models for smaller tissue portions be effectively trained with high-level labels, such as slide diagnoses? These are the questions that will be addressed throughout this thesis, especially for colorectal, cervical and breast tissue sample analysis. The main goal is to develop medical image diagnostic tools that can be used in clinical practice, to aid clinicians and, indirectly, patients towards more personalised precision cancer treatments. With this work, new CAD systems for computational pathology are proposed, with state-of-the-art results, developed without the need for many extensively annotated data, thus contributing to the development of weakly-supervised learning methods.

1.3 List of publications

The contributions of this doctoral research to the computational pathology field have been disseminated as part of fifteen scientific publications. These are (clustered by type and in reverse chronological order):

- Articles in international journals:

  1. S.P. Oliveira*, D. Montezuma*, A. Moreira*, D. Oliveira, P.C. Neto, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, I.M. Pinto and J.S. Cardoso. A CAD system for automatic dysplasia grading on H&E cervical whole-slide images. *Scientific reports*, 2022 [submitted, waiting for decision]
  2. P.C. Neto*, D. Montezuma*, S.P. Oliveira*, D. Oliveira, J. Fraga, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, S. Reinhard, I. Zlobec, I.M. Pinto and J.S. Cardoso. A CAD System for Colorectal Cancer from WSI: A Clinically Validated Interpretable ML-based Prototype. *Nature Communications*, 2022 [submitted, waiting for decision]
  3. Diana Montezuma, S.P. Oliveira, P.C. Neto, D. Oliveira, A. Monteiro, J.S. Cardoso and I.M. Pinto. Annotating for Artificial Intelligence applications in Digital Pathology: a practical guide for pathologists and researchers. *Modern Pathology*, 2022 [accepted, waiting for publication]
  4. P.C. Neto*, S.P. Oliveira*, D. Montezuma*, J. Fraga, L. Ribeiro, S. Gonçalves, I.M. Pinto and J.S. Cardoso. iMIL4PATH: A Semi-Supervised Interpretable Approach for Colorectal Whole-Slide Images. *Cancers*, 14(10):2489, 2022
  5. S.P. Oliveira*, P.C. Neto*, J. Fraga *, D. Montezuma, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, I.M. Pinto and J.S. Cardoso. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Scientific Reports*, 11(1):1-15, 2021
  6. S.P. Oliveira, J.R. Pinto, T. Gonçalves, R. Canas-Marques, M.J. Cardoso, H.P. Oliveira and J.S. Cardoso. Weakly-Supervised Classification of HER2 Expression in Breast Cancer Haematoxylin and Eosin Stained Slides. *Applied Sciences*, 10(14):4728, 2020

[*] Shared co-first authorship

- Articles in international conference proceedings:

    1. T. Albuquerque, A. Moreira, B. Barros, D. Montezuma, S.P. Oliveira, P.C Neto, J. Monteiro, L. Ribeiro, S. Gonçalves, A. Monteiro, I.M. Pinto and J.S. Cardoso. Quality Control in Digital Pathology: Automatic Fragment Detection and Counting. *In 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022 [not covered]

- Abstracts in international conferences:

    1. J. Romão, D. Montezuma, S.P. Oliveira, P.C. Neto, J. Monteiro, L. Ribeiro, S. Gonçalves, A. Monteiro, I.M. Pinto and J.S. Cardoso. Computer-aided tool for CRC diagnosis: from the AI model to the clinical software prototype. *In 18th European Congress on Digital Pathology (ECDP)*, SY12.04, 2022 (presented by the thesis author)
    2. T. Albuquerque, D. Montezuma, S.P. Oliveira, P.C. Neto, J. Monteiro, L. Ribeiro, S. Gonçalves, A. Monteiro, I.M. Pinto and J.S. Cardoso. Quality checkpoint in pathology specimens handling: an AI system to automate fragment detection and count. *In 18th European Congress on Digital Pathology (ECDP)*, SY05.03, 2022 [not covered]
    3. P.C. Neto, S.P. Oliveira, D. Montezuma, J. Fraga, I.M. Pinto and J.S. Cardoso. Colorectal Biopsies Assessment Using Weakly Supervised Classification of Whole-Slide Images. *In 17th European Congress on Digital Pathology (ECDP)*, OP01-9, 2021
    4. D. Montezuma, J. Fraga, S.P. Oliveira, P.C. Neto, A. Monteiro and I.M. Pinto. Annotation in Digital Pathology: how to get started? Our experience in classification tasks in Pathology. *In 33rd European Congress of Pathology–Abstracts, Virchows Archiv, 479(1):S1-S320*, 2021

- Extended abstracts in national conference proceedings:

    1. S.P. Oliveira, P.C. Neto and J.S. Cardoso. A semi-supervised approach for colorectal cancer diagnosis from H&E whole slide images. *In 27th Portuguese Conference in Pattern Recognition (RECPAD)*, 2021
    2. S.P. Oliveira, J.R. Pinto, T. Gonçalves, H.P. Oliveira and J.S. Cardoso. “IHC Classification in Breast Cancer H&E Slides with a Weakly-Supervised Approach. *In 26th Portuguese Conference in Pattern Recognition (RECPAD)*, 2020
    3. S.P. Oliveira, H.P. Oliveira. Automatic Segmentation of Invasive Breast Cancer on Whole-Slide Images. *In 25th Portuguese Conference in Pattern Recognition (RECPAD)*, 2019
    4. S.P. Oliveira, M.J. Cardoso, J.S. Cardoso and H.P. Oliveira. Radio-Pathomics Approach for Breast Tumor Signature: an overview. *In 24th Portuguese Conference in Pattern Recognition (RECPAD)*, 2018

1.4 Collaborations

During this doctoral project, the author established several close collaborations with researchers, not only from the VCMI research group[1], at INESC TEC, but also from national and international institutions, which have added extra value to the work carried out over the last few years.

1.4.1 Research projects

The CADpath.AI project[2], a collaboration between INESC TEC and the national biggest private pathology laboratory, IMP Diagnostics, aimed to develop a tool capable of diagnosing colorectal and cervical cancer through the automatic analysis of histological samples. To this end, the work is focused on learning algorithms, mainly deep learning techniques, to identify and characterise structural and morphological features on WSI. The author of this thesis contributed to the development of the classification algorithms and assisted in the design of a prototype to be integrated into the IMP laboratory's digital workflow.

Following the work developed in this project, particularly in the application for colorectal cancer, the author was invited as one of the speakers at the "Joint Symposium Digestive Diseases Pathology (GI) & IT in Pathology", at the 34th European Congress of Pathology 2022, in Basel.

1.4.2 MSc thesis and internships supervision

The author of this thesis collaborated as co-supervisor, alongside Hélder P. Oliveira (supervisor), in the development of João Alves' work entitled "Prostate Cancer Automatic Grading from Digitized H&E-stained Histopathology Slides" (2021), presented to FEUP as part of the Integrated Master's degree in Informatics and Computing Engineering.

Besides the abovementioned dissertation, the author also collaborated in the supervision of six more students on curricular, extra-curricular and internships and summer internships related to the computational pathology field (presented in reverse chronological order):

- Ana Moreira, extra-curricular internship, 2022, "Epithelium segmentation on cervical tissue samples" (co-supervisor, alongside Jaime S. Cardoso)

- Inês Campos, extra-curricular internship, 2022, "HER2 over-expression classification on H&E breast samples" (supervisor)

- Guilherme Barbosa, summer internship, 2021, "Explainable AI for computational pathology: identify and explain disease" (co-supervisor, alongside Wilson Silva)

- Ana Filipa Ferreira, curricular internship, 2021, "Breast cancer: classification of HER2 status based on H&E slides", (co-supervisor, alongside Jaime S. Cardoso)

[1] vcmi.inesctec.pt

[2] CADpath.AI was funded by the European Regional Development Fund (ERDF), through the Operational Programme for Competitiveness and Internationalisation (COMPETE 2020), within the project POCI-01-0247-FEDER-045413.

- Rui Santos, Leonor Sousa and Ana Filipa Ferreira, summer internship, 2020, "Breast cancer histopathology image segmentation/classification" (supervisor)

1.4.3 Challenges

In 2019, alongside Isabel Rio-Torto, João R. Pinto and Tiago Gonçalves, the author of this thesis participated in the HEROHE challenge[3] (held as part of the *European Congress on Digital Pathology* - ECDP 2020), where the goal was to automatically identify HER2 positive and negative breast cancer specimens, by evaluating only the morphological features present on H&E slides.

In 2020, together with Pedro Costa and Tânia Pereira, the author joined the PANDA challenge[4], the largest histopathology competition to date (with 10,616 prostate biopsy samples), organised by Radboud UMC and Karolinska Institutet, that aimed to develop an AI algorithm for prostate cancer Gleason grading.

1.4.4 Scientific events organisation

This doctoral work and the abovementioned research project, CADpath.AI, motivated the proposal of two workshops on the computational pathology topic, at two renowned international conferences on computer vision:

- In 2021, INESC TEC (VCMI research group) and IMP Diagnostics, together with Google Health (USA), Karolinska Intitutet (Sweden) and Radboud (The Netherlands), organised the workshop on *Computational Challenges in Digital Pathology* (CDpath), hosted at the *International Conference on Computer Vision* (ICCV), where the author of this thesis collaborated as one of the main organisers, program chair and publicity chair.

- In 2022, the same team, jointly with NTUA (Greece) and the University of Lincoln (UK), organised the workshop on *AI-enabled Medical Image Analysis: digital pathology and radiology/COVID19* (AIMIA), hosted at the *European Conference on Computer Vision* (ECCV), where the author collaborated as one of the main organisers, program chair (digital pathology track) and publicity chair.

Additionally, the author helped in the organisation of two other workshops related to the biometrics topic: the 2020 edition of the *International Workshop on Biometrics and Forensics* (IWBF), organised by INESC TEC (VCMI research group) and NTNU; and the three editions of the *Workshop on Explainable & Interpretable Artificial Intelligence for Biometrics* (xAI4Biometrics), hosted yearly at the *IEEE/CVF Winter Conference on Applications of Computer Vision* (WACV), from 2021 to 2023. In all these events, she has collaborated as publicity chair.

Since 2017, the author was also involved in the VISUM summer school on computer vision and machine intelligence, an event organised by the VCMI research group. In the 5th edition (2017), she

[3] https://ecdp2020.grand-challenge.org/

[4] https://www.kaggle.com/c/prostate-cancer-grade-assessment

was part of the staff, during the summer school week; from the 6th to the 9th (2018-2021) editions, she was part of the organising committee, working on the website, social media and logistics; and in the 10th edition (2022), she was a co-chair, alongside Ana Rebelo and Wilson Silva.

1.5 Document structure

The remainder of this document is composed of four parts, besides this introduction, which offers an overview of the fundamental concepts related to computational pathology, and summarises the contributions of this doctoral work.

Part II is centred on colorectal cancer grading, presenting the clinical background and the state-of-the-art on the topic in Chapter 2, and the proposed approaches from Chapter 3 to Chapter 5. It starts with an evaluation of the dataset requirements, followed by the proposal of a weakly-supervised model to grade CRC. Additionally, a framework to integrate this model into clinical practice is presented, together with a clinical evaluation of the results.

Part III covers the work developed on cervical cancer diagnosis. Chapter 6 introduces the clinical aspects of cervical cancer and an overview of the state-of-the-art work on dysplasia grading from H&E cervical tissue samples, and Chapter 7 details the proposed framework, from tissue segmentation to slide classification, with different levels of annotation.

Part IV focuses on breast cancer diagnosis, especially on classifying HER2 expression directly from H&E samples. Similarly to the last ones, this part starts by introducing some clinical insights about breast cancer in Chapter 8, where the state-of-the-art on computational pathology for breast HER2 scoring is also summarised. Finally, in Chapter 9 an approach to this topic is proposed, based on cross-domain adaptation and weakly-supervised learning.

The last Part, Chapter 10, summarises the contributions of this doctoral work and concludes this thesis with some final remarks and ideas for future work.

Part II

Colorectal cancer

Chapter 2

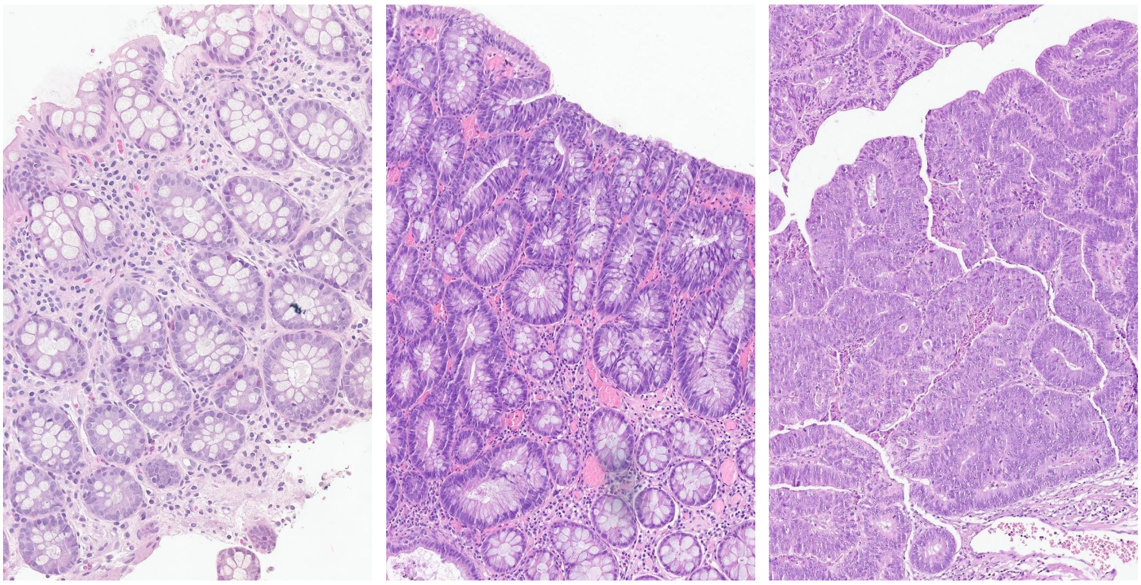
Colorectal cancer insights

2.1 Epidemiology

Colorectal cancer (CRC) represents one of the major public health problems today. Globocan estimated data for 2020 show that CRC is the third most incident cancer (10% of all cancers) and the second most deadly (9.4%; only surpassed by lung cancer, 18%) [26, 27]. CRC is a disease of modern times: the highest rates of incidence happen in developed countries [28]. As the world becomes richer, and people shift to a western lifestyle, the incidence of CRC is expected to increase, since it is a multifactorial disease resulting from lifestyle, genetic, and environmental factors [28, 29]. Population growth and ageing lead to an increased incidence of the disease, as well as better and more numerous screening programs for early detection and prevention. The prevalence of screening among individuals aged 50 years and older increased from 38%, in 2000, to 66%, in 2018, according to data from the National Center for Health Statistics (NHIS) [30]. Importantly, CRC is a preventable and curable cancer if detected early on, and, therefore, screening is an effective tumour prevention measure [31]. Screening determines the decrease in mortality through timely detection and removal of adenomatous polypoid (pre-malignant) lesions, promoting the interruption of progression to cancer. It should begin with colonoscopy in asymptomatic individuals aged 50 years or over (and without personal or family risk factors for CRC) and repeated every ten years if normal [32]. It is worth mentioning that, due to the Covid-19 pandemic, CRC screening programmes have been disrupted worldwide. As such, it is crucial that catch-up screening is provided as soon and effectively as possible, hoping to mitigate the impact on CRC deaths [33, 34]. Computer-aided diagnosis (CAD) solutions in CRC could help in this task, contributing to improving pathology diagnostic capacity.

2.2 CRC dysplasia grading

During the pathological assessment, colorectal biopsies/polyps can be stratified into non-neoplastic, low-grade dysplasia (LGD), high-grade dysplasia (HGD, including intramucosal carcinomas) and invasive carcinomas, regarding their development sequence. Colorectal dysplasia refers to the pre-malignant abnormal development of cells/tissues, which can eventually progress to tumour lesions. It is classified in low- and high-grade, with the last conferring a relatively higher risk of cancer (Figure 2.1).



(**a**) Non-neoplastic (**b**) Low-grade lesion (**c**) High-grade lesion

Figure 2.1: Normal colonic mucosa and dysplastic progression. Examples from CADpath dataset.

It is well-known that grading colorectal dysplasia is a somewhat subjective issue. In a study to evaluate inter-observer variability in HGD diagnosis, five gastrointestinal pathologists conducted a consensus conference in which criteria for colorectal HGD were developed [35]. When grading the same 107 polyps, the inter-observer agreement was found to be poor both before and after the consensus. Other studies have also shown sub-optimal agreement in grading colorectal dysplasia [36, 37]. Despite this, the most recent guidelines from the European Society of Gastrointestinal Endoscopy (ESGE), as well as those from the US multi-society task force on CRC, continue to recommend surveillance for polyps with high-grade dysplasia regardless of their size [32, 38]. Patients requiring surveillance after polypectomy include those with complete removal of:

- at least one adenoma $\geq$ 10 *mm* or with high-grade dysplasia;
- five or more adenomas;
- any serrated polyp $\geq$ 10 *mm* or with dysplasia [32].

As such, it remains a current practice in most countries (although not in every laboratory) to evaluate and grade colorectal dysplasia.

2.2.1 Grading guidelines

To date, there are still no tangible criteria on what distinguishes the high end of LGD from the low end of HGD. Although some reporting guidelines regarding grading dysplasia in colorectal biopsies [39–41], objective criteria are still lacking. It is fairly easy for a pathologist to diagnose a typical low-grade or high-grade adenoma but since in fact, these lesions exist in a continuum, the correct assessment of many intermediate cases is more difficult. Nevertheless, protocols such as the English National Health System (NHS) bowel cancer screening programme guidance, with guidelines from the bowel cancer screening programme pathology group [39] or the Pan-Canadian consensus for colorectal polyps report [41], can aid pathologists in grading colorectal lesions more objectively. Additional information from reference books, such as the World Health Organization (WHO) Classification of Tumours: digestive system tumours [27], can also assist in this task. The most relevant characteristics that differentiate low- and high-grade dysplasia are detailed in Table 2.1.

Table 2.1: Colorectal low- and high-grade dysplasia characterisation.

	Low-grade dysplasia	**High-grade dysplasia**
Extension	–	Changes must involve more than two glands (except in tiny biopsies)
Low power magnification	Lack of architectural complexity suggests low-grade dysplasia throughout	Alterations should be identifiable at low power: complex architectural abnormalities, epithelium looks thick, blue, disorganised and “dirty”
Cytology/ Architecture	Does not combine cytological high-grade dysplasia with architectural high-grade features	Needs to combine high-grade cytological and high-grade architectural alterations
Architectural features*	Gland crowding, showing parallel disposition, with no complexity (no back-to-back or cribriform); Global architecture may vary from tubular to villous	Complex glandular crowding and irregularity; Prominent budding; Cribriform appearance and back-to-back glands; Prominent intra-luminal papillary tufting
Cytological features**	Nucleus are enlarged and hyper-chromatic, many times cigar-shaped; Nucleus maintain basal orientation (only up to the lower half of the height of the epithelium, although in some cases we can see glands with full-thickness nuclear stratification - not HGD if the architecture is bland); There is no loss of cell polarity or pleomorphism; No atypical mitosis; Maintained cytological maturation	Noticeably enlarged nuclei, often with a dispersed chromatin pattern and evident nucleoli; Loss of cell polarity or nuclear stratification to the extent that the nuclei are distributed within all 1/3 of the height of the epithelium; Atypical mitoses; Prominent apoptosis/necrosis, giving the lesion a "dirty" appearance; Lack of cytological maturation

* Architectural features: gland morphology and placement. ** Cytological features: cell level characteristics.

2.3 Computational pathology on CRC

Although the rise of DL and its application in computer vision is critical to computer-aided diagnosis (CAD) research, the development of AI applications for colorectal cancer (CRC) diagnosis on WSI is still limited, as noted by Thakur *et al.* [42]: of the 30 papers reviewed, only 20% have a diagnosis as a final goal. In fact, the majority of the papers deal with a wide variety of tasks, with a particular focus on tissue segmentation, the goal of 62% of the reviewed papers [42]. Last year, Wang [43] *et al.* also published a review on the application of AI to CRC diagnosis and therapy, reflecting the same trend. However, CRC diagnosis is a growing application, with an increasing number of publications in recent years. In the next section, we collect and describe the published works on CRC diagnosis, with a particular focus on slide diagnosis (Table 2.2), but also summarise some works using partial regions of tissue (region crops or tiles) without aggregation for WSI.

2.3.1 CRC diagnosis on WSI

In 2012, Kalkan *et al.* [44] proposed a method for CRC automatic detection from Haematoxylin and Eosin (H&E) slides, combining textural and structural features of smaller patches (1024×1024 pixels). Firstly, the patches are classified into normal, inflamed, adenomatous or cancer with a k-NN classifier, based on local shape and textural features, such as Haralick features, Gabor filters features and colour histograms features. Then, the (up to) 300 patches representing the slide are summarised in the average probabilities for all the four primary classes, and used as a feature vector for a logistic-linear regressor, to obtain a final slide diagnosis: normal or cancer. The proposed method was trained on 120 H&E stained slides and achieved an Area Under the Curve (AUC) of 0.90 and an average accuracy of 87.69%, with accuracies of 79.17% and 92.68% for cancer and normal slides, respectively. Similarly, using traditional computer vision techniques, Yoshida *et al.* [45] presented an approach to classify CRC H&E slides into 4 types: non-neoplastic, adenoma, carcinoma and unclassifiable. For each WSI, all tissue regions are identified, summing 1328 sections from 1068 H&E slides. Then, each section is processed for blur detection and colour normalisation before the analysis in two steps: cytological atypia analysis and structural atypia analysis. In the first step, the method proposed by Cosatto *et al.* [46] is used, based on multiple instance learning (MIL) formulation using a Multilayer Perceptron (MLP), to grade the degree of cytological alteration of the tissue (high or low). Then, the image is classified into low atypia, intermediate atypia, high atypia or unclassifiable, based on structural nuclear features and cytoplasmatic features, extracted from consecutive ROIs, that are summarised by the mean-square of the top 3 ROIs. Finally, each image is classified based on the combination of structural atypia analysis result (high, intermediate or low) and the cytological atypia analysis result (high or low), given that carcinoma presents higher atypia values. The model has an undetected carcinoma rate of 9.3%, an undetected adenoma rate of 0.0% and an overdetection proportion of 27.1%.

The first DL application model was presented in 2017, by Korbar *et al.* [47], to automatically classify colorectal polyps on H&E stained slides into five classes: normal, hyperplastic, sessile serrated, traditional serrated adenoma, tubular adenoma and tubulovillous/villous adenoma. The 697

Table 2.2: Literature overview on colorectal WSI diagnosis (to be continued).

Author	**Year**	**Task**	**Dataset**	**Description**	**Results**
Kalkan *et al.* [44]	2012	CRC detection (normal vs cancer)	120 slides (tile annot.)	1024x1024 px tiles k-NN classifier + Logistic-linear classifier	ACC: 87.69% AUC: 0.90
Yoshida *et al.* [45]	2017	CRC classification (4-class): unclassifiable, non-neoplastic, AD and CA	1068 slides (tissue sections labels)	Tissue sections crop + cytological atypia analysis + structural atypia analysis + overall classification	FNR (CA): 9.3% FNR (adenoma): 0% FPR: 27.1%
Korbar *et al.* [47]	2017	Polyp classification (6-class): normal, HP, SS, TS, T and TV	697 slides (annot.)	$\approx$ 811x984 px ROIs ResNet-152 + argmax of tile class frequency	ACC: 93% Precision: 89.7% Recall: 88.3% F1-score: 88.8%
Iizuka *et al.* [48]	2020	CRC classification (3-class): non-neoplastic, AD and ADC	4536 slides (annot.) + 547 slides (external)	512x512 px tiles at 20X Inception-v3 + RNN	AUC (ADC): 0.962 AUC (ADC, external set): 0.982
Song *et al.* [49]	2020	CRC detection (normal vs adenoma)	411 slides (annot.) + 168 slides (external)	640x640px tiles at 10X Modified DeepLab-v2 + pixel probability selection	AUC: 0.92 ACC (external set): >90%
Wei *et al.* [50]	2020	Polyp classification (5-class): Normal, HP, T, TV, SS	508 slides + 238 slides (external)	224x224 px tiles at 40X ResNet ensemble + hierarchical classification	ACC: 93.5% ACC (external set): 87%

HP: hyperplastic; *SS*: sessile serrated; *TS*: traditional serrated; *T*: tubular; *TV*: tubulovillous/villous; *AD*: adenoma; *CA*: carcinoma; *ADC*: adenocarcinoma; *ROI*: region of interest; *k-NN*: k-nearest neighbours; *SVM*: Support Vector Machine; *MLP*: Multi-Layer Perceptron; *ACC*: Accuracy; *AUC*: Area under the ROC curve; *FNR/FPR*: false negative/positive rate

Table 2.2: Literature overview on colorectal WSI diagnosis.

Author	**Year**	**Task**	**Dataset**	**Description**	**Results**
Xu *et al.* [51]	2020	CRC detection (normal vs cancer)	307 slides (annot.) + 50 slides (external)	768x768 px tiles Inception-v3 + tiles tumour probability thresholding	ACC: >93% ACC (external set): >87%
Wang *et al.* [52, 53]	2021	CRC detection (normal vs cancer)	13,111 slides from 13 centres (62,919 annotated tiles)	300x300 px tiles Inception-v3 + tile-cluster-based aggregation	ACC: >93% AUC: >0.94 Sensitivity: >92% Specificity: >88%
Marini *et al.* [54]	2021	Polyp classification (5-class): Normal, HP, LGD, HGD, CA	2,323 slides (some annotations)	224x224 px tiles ResNet-34 +attention layers + multi-scale framework	ACC: 85.7% F1-score: 0.68 ACC (binary): 87.0% F1-score: 0.89
Ho *et al.* [55]	2022	Patient stratification: low vs. high risk	294 slides (66,191 annotated tiles)	775x522 px tiles Faster-RCNN (with ResNet-101 backbone) + Decision Tree	ACC: 79.3% AUC: 0.9 Sensitivity: 97.4% Specificity: 60.3%

HP: hyperplastic; *LGD*: low-grade dysplasia; *HGD*: high-grade dysplasia; *CA*: carcinoma; *ACC*: Accuracy; *AUC*: Area under the ROC curve

H&E stained slides (annotated by a pathologist) were cropped into ROIs of 811 × 984 pixels (mean size) and then divided into overlapping smaller patches. These patches were classified using the ResNet-152 and the prediction of the slide was obtained as the most common colorectal polyp class among all patches of the slide. However, if no more than five patches are identified with the most common class, with a confidence higher than 70%, the slide is classified as normal. The proposed system achieved 93.0% accuracy, 89.7% precision, 88.3% recall and 8.8% F1-score. Later, the authors proposed a visualisation method [56], based on this approach, to identify highly-specific ROIs for each type of colorectal polyps within a slide, using the Guided Grad-CAM method [57] and a subset of data (176 H&E colorectal slides).

In 2020, several authors presented solutions for CRC diagnosis, with varying degrees of detail. Iizuka *et al.* [48] proposed the combination of an Inception-v3 network with a recurrent neural network (RNN) to classify H&E colorectal WSI into non-neoplastic, adenoma and adenocarcinoma. Each slide was divided into patches of 512 × 512 pixels (at 20X magnification, with a sliding window of 256 pixels) and assigned to one of the three diagnostic classes. Then, all tiles are aggregated using a RNN, trained to combine the features outputted by the CNN. The dataset consists of subsets from two different institutions, summing 4536 WSIs. Moreover, the model was also evaluated on a subset of 547 colon surgical resection cases from The Cancer Genome Atlas (TCGA) repository [58], containing adenocarcinoma and normal samples (TCGA-COAD collection). On the private dataset, the proposed approach measured AUCs of 0.962 and 0.992 for colorectal adenocarcinomas and adenomas, respectively. On the public dataset, the model achieved an 0.982 AUC for adenocarcinomas. It is noteworthy that the authors report that, since the samples from the external subset are much larger than the biopsies used for training, the RNN aggregation was replaced by a max-pooling aggregation. Meanwhile, Wei *et al.* [50] aimed to identify five types of polyps in H&E stained colorectal slides: normal, tubular adenoma, tubulovillous or villous adenoma, hyperplastic polyp, and sessile serrated adenoma. To train the model, the authors used 509 slides (with annotations of relevant areas by five specialised pathologists). For further testing, they used an external set of 238 slides, obtained from different institutions. The model consists of an ensemble of the five versions of ResNet (namely, networks with 34, 50, 101, and 152 layers) to classify tiles of 224 × 224 pixels (at 40X magnification). Then, the patches are combined with a hierarchical classifier to predict a slide diagnosis. Based on the predicted tile classes, the model first classifies a polyp as adenomatous or serrated, by comparing the frequency of tiles classes (tubular, tubulovillous, or villous vs. hyperplastic or sessile serrated). Adenomatous polyps with more than 30% tubulovillous or villous adenoma tiles are classified within this class and the remaining are classified as tubular adenoma. Serrated polyps with more than 1.5% of sessile serrated tiles are classified within this class and the remaining are classified as hyperplastic. The thresholds were set with a grid search over the training set, reaching an accuracy of 93.5%, on the internal test set, and 87.0% on the external test set.

Also in 2020, two other authors proposed segmenting colorectal tissue simultaneously with the diagnosis. Song *et al.* [49] presented an approach based on a modified DeepLab-v2 network on 640 × 640 pixel tiles, at a 10X magnification. The dataset consists of 411 annotated slides, labelled as colorectal adenomas or normal mucosa (which includes chronic inflammation), and a subset of 168 slides collected from two other institutions, to serve as an external test. The authors modified the DeepLab-v2 network by introducing a skip layer that combines the upsampled lower layers with the higher layers, to retain semantic details of the tiles. Then, the 15th largest pixel-level probability is used for the slide-level prediction. In the inference phase, the slide is decomposed into tiles of 2200×2200 pixels. The proposed approach achieved an AUC of 0.92 and when tested on the independent dataset, an accuracy over 90%. In turn, the model of Xu *et al.* [51] was trained on a set of 307 slides (normal and CRC), with tissue boundaries manually annotated by a pathologist, achieving a mean accuracy of 99.9% for normal slides and 93.6% for cancer slides, and a mean dice coefficient of 88.5%. For further testing, the model was also evaluated on an external set of 50 CRC slides and achieved a mean accuracy of 87.8% and a mean Dice coefficient of 87.2%. The method uses the Inception-v3 architecture, pre-trained on the ImageNet dataset, to classify patches of 768×768 pixels, resized to 299×299 pixels. The final tumour regions and slide diagnosis are obtained by thresholding the tile predictions: tiles with tumour probability above 0.65 are considered cancer.

In addition, in 2021, using the Inception-v3 architecture, Wang *et al.* [52, 53] developed a framework to detect tumours which retrieves the final classification of a slide and also a map of tumour regions, using 13,111 slides from 13 independent centres. From the tile classifier, which distinguishes normal and cancer tiles, slide prediction is obtained with a tile-cluster-based aggregation: a WSI is positive if several positive patches are topologically connected as a cluster, e.g., four patches as a square, and negative otherwise. This approach was tested on several WSI sets, achieving accuracies higher than 93%, an AUC higher than 0.94, sensitivities higher than 92% and specificities higher than 88%. Marini *et al.* [54] proposed a multi-scale task multiple instance learning (MuSTMIL) method to classify five colon-tissue findings: normal glands, hyperplastic polyps, low-grade dysplasia, high-grade dysplasia and carcinomas. Using multiple scale branches, in a multi-task network, the model combines features from several magnification levels of a slide in a global prediction. Developed with more than 2,000 WSI, this method reached an ACC of 87.0% and a 0.893 F1-score, in the binary setup, and an ACC of 85.7% and 0.682 F1-score, in the multi-class setup.

In 2022, Ho *et al.* [55] presented an algorithm that simultaneously segments glands, detects tumour areas and sorts the slides into low-risk (benign, inflammation or reactive changes) and high-risk (adenocarcinoma or dysplasia) categories. The authors proposed a Faster-RCNN architecture, with a ResNet-101 backbone network, for glandular segmentation of tiles, followed by a gradient-boosted decision tree for slide classification, using features such as the total area classified as adenocarcinoma or dysplasia, and the average prediction certainty for these areas. The dataset comprises 294 slides, combining samples of a private set and samples from the TCGA collection. The model achieved an ACC of 79.3% with an AUC of 0.917, sensitivity of 97.4% and specificity of 60.3%.

While some of the reported results are impressive and show high potential, there are still some obvious shortcomings that need to be addressed. One of the issues is model evaluation: most of the papers analysed have not used any form of external evaluation on public benchmark datasets, as can be seen by the dataset descriptions in Table 2.2. This validation is necessary to understand and compare the performance of models that, otherwise, cannot be directly compared to each other due to the use of distinct datasets. It also limits the study of the robustness of the model when it is exposed to data from sources other than those used for training. On the other hand, as with any DL problem, the size of the dataset is crucial. Although, as mentioned earlier, it is expensive to collect the necessary amount of data to develop a robust model, it is noticeable that the reviewed articles could greatly benefit from an increase in the volume of data since most of the works are trained on only a few hundred slides. Describing and sharing how the data collection and annotation processes were performed is also crucial to assess the quality of the dataset and the quality of the annotations. For example, the number of annotators, their experience in the field, and how their discrepancies were resolved. However, this description was not a common practice in the articles reviewed. Moreover, comparing models becomes more complicated when one realises that the number of classes used for the classification tasks is not standardised across published work. Therefore, together with the difference in the kind of metrics presented, direct comparisons should be made with caution.

2.3.2 CRC classification on tiles/crops

Despite the small number of published works on colorectal WSI diagnosis, there is a myriad of other articles also working on CRC classification using information from smaller tissue regions, that can be exploited as a basis for general diagnostic systems. Despite the different tasks, these works that use image crops, or even small patches [59–63], can be leveraged for slide diagnosis, in combination with aggregation methods that combine all the extracted information in a single prediction.

As for WSI classification, there are also approaches for crop image classification based on traditional computer vision methods or DL models, and even a combination of both. In 2017, Xu *et al.* [64] proposed the combination of an Alexnet (pre-trained on ImageNet) as a feature extractor and an SVM classifier to develop both a binary (normal vs. cancer) and a multiclass (CRC type) classification approach for cropped images (variable size, 40X magnification) from CRC H&E slides. The latter goal is to distinguish between 6 classes: normal, adenocarcinoma, mucinous carcinoma, serrated carcinoma, papillary carcinoma cribriform adenocarcinoma. Each image is divided into overlapping patches of 672×672 pixels (then resized to 224×224 pixels), from which 4096-dimensional feature vectors are extracted. For cancer detection, features are selected based on the differences between positive and negative labels: the top 100 feature components (ranked from the largest differences to the smallest) are kept. Then the final prediction is obtained with a linear SVM (one-vs-rest classification for CRC type). The CRC detection model has an accuracy of 98% and the CRC type classification model has an accuracy of 87.2%, trained on 717 image crops. Already in 2019, Yang *et al.* [65] and Ribeiro *et al.* [66] proposed works based on colour and geometric features, and classical ML methods, to classify CRC. With colour pictures (350×350 pixels, 20X magnification) from H&E stained colorectal tissue sections (labelled and marked by professional pathologists), Yang *et al.* [65] proposed a method based on sub-patch weight colour histogram features, the RelicfF based forward selection algorithm and a Morlet wavelet kernel-based least squares SVM classifier. The method was developed using a total of 180 images and obtained an AUC and accuracy of 0.85 and 83.13%, respectively. Ribeiro *et al.* [66] associated multidimensional fractal geometry, curvelet transforms and Haralick features and tested several classifiers on 151 cropped images (775 × 522 pixels, 20X magnification) from 16 H&E adenocarcinoma samples. The best result, an AUC of 0.994, was achieved with multiscale and multidimensional percolation features (from curvelet sub-images with scales 1 and 4), quantifications performed with multiscale and multidimensional lacunarity (from input images and their curvelet sub-images with scale 1) and a polynomial classifier.

Regarding DL models, there are also several proposed approaches for several CRC classification tasks. In 2017, Haj-Hassan *et al.* [67] proposed a method based on multispectral images and a custom CNN to predict 3 CRC types: benign hyperplasia, intraepithelial neoplasia and carcinoma. From the H&E stained tissue samples of 30 patients, 16 multispectral images of 512 × 512 pixels are acquired, in a wavelength range of 500-600nm. After a CRC tissue segmentation with an Active Contour algorithm, images are cropped in smaller tiles of 60 × 60 pixels (with the same slide label) and fed to a custom CNN (input size of 60 × 60 × 16), reaching an accuracy of 99.17%. In 2018, Ponzio *et al.* [68] adapted a pre-trained VGG16 net for CRC classification into adenocarcinoma, tubulovillous adenoma and healthy tissue. They used tissue subtype large ROIs, identified by a skilled pathologist from 27 H&E stained slides of colorectal tissue from a public repository [69], that were then cropped into 1089 × 1089 patches, at a magnification level of 40x. By freezing the weights up to the most discriminative pooling layer (determined by t-SNE) and training only the final layers of the network, the solution provided a classification accuracy over 90%. The system was evaluated at two levels: the patch score (fraction of patches that were correctly classified) and patient score (per-patient patch score, averaged over all cases), which reached 96.82% and 96.78%, respectively. In 2019, Sena *et al.* [70] proposed a custom CNN to classify four stages of CRC tissue development: normal mucosa, early pre-neoplastic lesion, adenoma and carcinoma. The dataset consists of 393 images from H&E colorectal slides (20X magnification), cropped into nine sub-images of 864 × 548 pixels. For further validation on significantly different images, the authors also used the GLaS challenge dataset [71, 72], with 151 cropped images. Since both datasets differ in resolution, the GLaS images were resized with bi-cubic interpolation and centrally cropped. The proposed method obtained an overall accuracy of 95.3% and the external validation returned an accuracy of 81.7%. Meanwhile, Zhou *et al.* [73] proposed a pipeline to classify colorectal adenocarcinomas, based on the recent graph neural networks, converting each histopathological image into a graph, with nucleus and cellular interactions being represented by nodes and edges, respectively. The authors also propose a new graph convolution module, Adaptive GraphSage, to combine multilevel features. With 139 images (4548 × 7520 pixels, 20x magnification), cropped from WSI labelled as normal, low grade and high grade, the method achieved an accuracy of 97%. For the same classification task, in 2020, Shaban *et al.* [74] proposed a context-aware convolution neural network to incorporate contextual information in the training phase. Firstly, tissue regions (1792 × 1792 pixels) are decomposed in local representations by a CNN (224 × 224 pixels input), and the final prediction is obtained by combining all contextual information with a representation aggregation network, considering the spatial organisation of smaller tiles. This method was developed on 439 images ($\approx 5000 \times 7300$ pixels, 20X magnification) and achieved an average accuracy of 99.28% and 95.70% for a binary and three-class setup, respectively.

2.4 Summary

Colorectal cancer (CRC) diagnosis is based on samples obtained from biopsies, assessed in pathology laboratories. Due to population growth and ageing, as well as better screening programs, the CRC incidence rate has been increasing, leading to a higher workload for pathologists. In this sense, the application of AI for automatic CRC diagnosis, particularly on WSI, is of utmost relevance, in order to assist professionals in case triage and case review.

Despite the ever-growing number of publications of ML methods applied to CAD systems, there is a dearth of published work for the task of joint detection and classification of colorectal lesions from WSI, lagging CRC behind pathologies such as breast cancer and prostate cancer. Furthermore, a significant amount of the work developed does not use the entire WSI but instead uses crops and regions of interest extracted from these images. While these latter works show significant results, the applicability of such works in clinical practice is limited. Similarly, publicly available datasets often consist of crops instead of the original image. Others include only abnormal tissue, limiting the development of CRC diagnostic systems and the detection task.

Chapter 3

Feasibility study on a weakly-annotated dataset for CRC grading

Author Contributions

The research work described in this chapter was conducted in collaboration with Pedro C. Neto and the IMP Diagnostics team, under the clinical supervision of Isabel M. Pinto, and the technical supervision of Jaime S. Cardoso. The author of this thesis contributed to this work on the problem conceptualisation, data curation, the preparation and conduction of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- S.P. Oliveira*, P.C. Neto*, J. Fraga *, D. Montezuma, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, I.M. Pinto and J.S. Cardoso. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Scientific Reports*, 11(1):1-15, 2021
- S.P. Oliveira, P.C. Neto and J.S. Cardoso. A semi-supervised approach for colorectal cancer diagnosis from H&E whole slide images. *In 27th Portuguese Conference in Pattern Recognition (RECPAD)*, 2021

*Shared co-first authorship

Collecting and labelling data for computational pathology problems is a lengthy and expensive process. As seen in Chapter 2, research is often conducted on small datasets containing a high granularity of annotations per sample. Despite the benefits of detailed annotations, researchers have recently turned their attention to weakly-supervised approaches. These approaches, notwithstanding the simplified annotation, can leverage larger datasets for learning. More importantly, weakly-supervised learning techniques are less prone to bias in data collection. Performance is also evaluated on a more extensive test set, and thus, the behaviour of the model in the real world can be generalised much more accurately. In this chapter, we conduct a feasibility study on the use of efficiently annotated datasets to drive the development of computer-aided diagnosis (CAD) systems for colorectal cancer (CRC) from whole-slide images (WSI). We attempt to answer the question of the required dimension of the dataset, as well as the extension of annotations, to enable the robust learning of predictive models. We also analyse the advantage of using a loss function adapted to the ordinal nature of the classes corresponding to the CRC scores.

3.1 Methodology

Traditional supervised learning techniques would require all the tiles extracted from the original image to be labelled. However, cancer grading (in clinical practice) aims to classify the WSI, not individual tiles. Moreover, labelling the tiles represents a significant effort with regard to the workload of the pathologists. Therefore, techniques such as multiple instance learning (MIL), have been adapted to computational pathology problems [75–77]. MIL only requires slide-level labels and the original supervised problem is converted to a weakly-supervised problem. The nature of the problem allows the implementation of this technique knowing that, if a WSI is classified with a label Y, no tile belongs to a more severe class than Y and at least one tile belongs to the label Y. Therefore, using the MIL concept, we propose a workflow (Figure 3.1) based on the work of Campanella *et al.* [76] with several adaptations:

a) *Ordinal labels:* First, the problem at hand has a multiclass formulation, whereas the original had only two labels. In order to contextualise the premises of the MIL method and the clinical information, the labels must not be seen as independent and their relation must be modelled. For instance, normal tissue is closer to low-grade lesions than to high-grade dysplasias. Thus, there is an order regarding labels.

b) *Removal of recurrent aggregation:* The original approach leveraged a Recurrent Neural Network (RNN) to aggregate predictions of individual tiles into a final prediction. All the tests conducted for the feasibility results did not show any benefit of having this RNN aggregation, in fact, the performance degraded. Thus, it was removed from the pipeline.

c) *Tile ranking using the expected value:* Using a single tile for the prediction requires a ranking rule in order to select the most representative of potentially thousands of tiles. Since the problem is non-binary, the original rule is not applicable [76]. Therefore, to create a ranking of tiles that are meaningful for the final prediction, the backbone network is used to compute the outputs of each tile and the expected value is then computed from these outputs:

$$tile\_score = \sum_{i=1}^{n\_classes} x_i \times p_i$$

with n_classes the number of classes, x_i the class, p_i the correspondent probability;

d) *Loss function:* The problem includes ordinal labels, so the minimisation of the cross-entropy fails to fully capture the model's behaviour. As mentioned before, the distance between labels is different and cross-entropy treats them as if they are equally distant. Thus, in an attempt to increase the performance of the initial baseline experiments the model is now optimised to minimise an approximation to the Quadratic Weighted Kappa (QWK) [78].

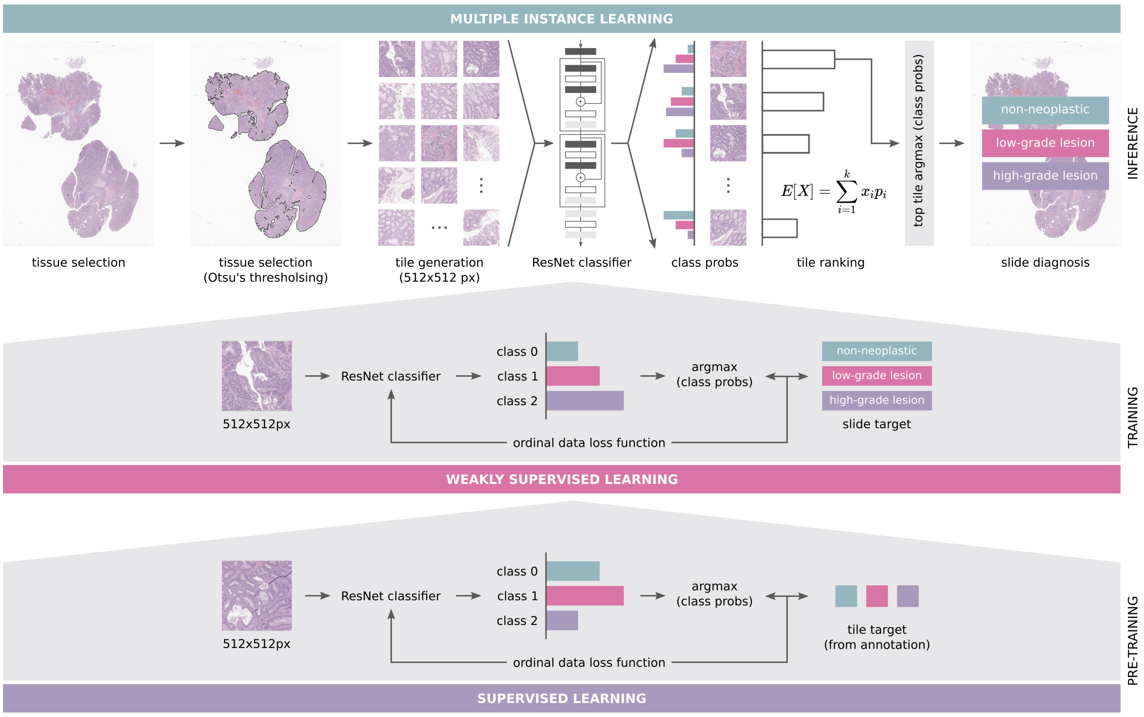


Figure 3.1: Proposed workflow for colorectal cancer diagnosis on whole-slide images.

3.1.1 Training details

The setup of the experiments was similar across datasets: ResNet-34 as the backbone, batch-size of 32, the Adaptive Moment Estimation (Adam) algorithm with a learning rate of 1×10^{-4} as the optimiser, tiles of 512×512 pixels that include 100% of tissue and mixed-precision from the Pytorch available package. Only one tile was used for predicting the label of the slide (MIL formulation), thus the training set was only regarding the selected tile. As for hardware, all the experiments were conducted using an Nvidia Tesla V100 (32 GB) GPU.

3.2 Datasets

This feasibility study was conducted on two datasets: the first contains colorectal haematoxylin & eosin (H&E) stained slides (CRS1k dataset), whereas the second includes prostate cancer (PCa) H&E-stained biopsy slides (PANDA dataset). As mentioned in section 2.3.1, there is a shortage of large public datasets containing colorectal WSIs and most of the existing ones are based on cropped regions instead of entire slides. Hence, we relied on a PCa dataset that, while not fully transferring to colorectal use case, is one of the largest WSI datasets publicly available. This amount of data allowed us to study the data requirements of a weakly supervised approach and how the performance evolved with the growth of the dataset.

3.2.1 CRS1k dataset

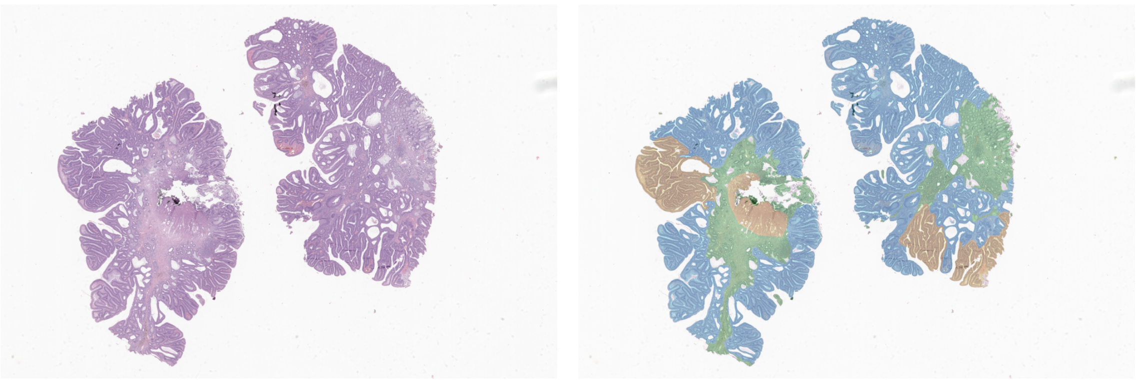
The CRS1k dataset contains 1,133 colorectal biopsy and polypectomy slides and is the result of our ongoing efforts to contribute to CRC diagnosis with a reference dataset. We aim to detect high-grade lesions with high sensitivity. High-grade lesions encompass conventional adenomas with high-grade dysplasia (including intra-mucosal carcinomas) and invasive adenocarcinomas. In addition, we also intend to identify low-grade lesions (corresponding to conventional adenomas with low-grade dysplasia). Accordingly, we created three diagnostic categories for the algorithm, labelled as non-neoplastic (NNeo), low-grade (LG) and high-grade (HG) lesions (Table 3.1). In the NNeo category, cases with suspicion/known history of inflammatory bowel disease/infection were omitted. We selected conventional adenomas as they were the largest group on daily routine (serrated lesions, and other polyp types, were omitted).

Table 3.1: CRS1k dataset class definition.

Algorithm data classes	**Pathological diagnosis**
Non-neoplastic	Normal CR mucosa, non-specific inflammation, hyperplasia
Low-grade lesion	Low-grade conventional adenoma
High-grade lesion	High-grade conventional adenoma and invasive adenocarcinoma

All cases were retrieved from the data archive of IMP Diagnostics laboratory, Portugal, and were digitised by 2 Leica GT450 WSI scanners, and evaluated by one of two pathologists (Figure 3.2a). Data collection and usage were performed in accordance with national legal and ethical standards applicable to this type of data. Since the study is retrospectively designed, no protected health information was used and patient informed consent is exempted from being requested.

Diagnostics were made using a medical grade monitor LG 27HJ712C-W and Aperio eSlide Manager software. When reviewing the cases, most diagnoses were coincident with the initial pathology report and no further assessment was made. In case of difference, the case was rechecked



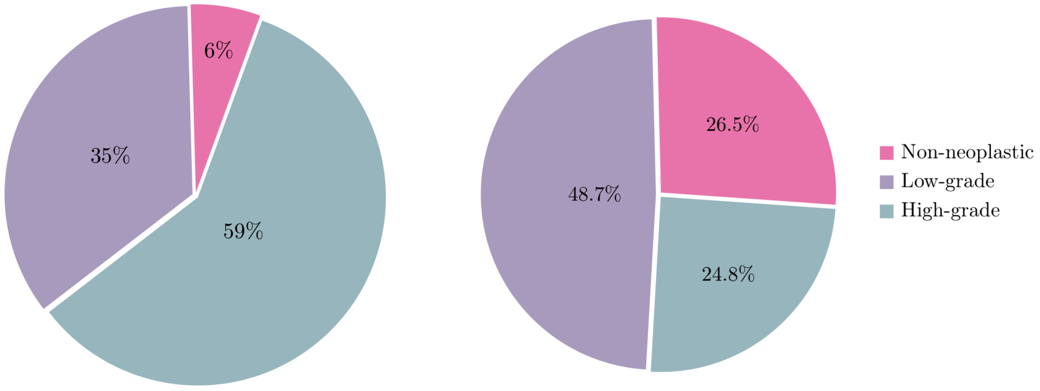
(**a**) (**b**)

Figure 3.2: Example of a colorectal WSI (**a**), with manual segmentations overlayed (**b**). Tissue regions are annotated as non-neoplastic (green), low-grade (blue) or high-grade lesions (yellow).

and decided between the two pathologists. A small number of cases (n=100) were further annotated with region marks (Figure 3.2b) by one of the pathologists and then rechecked by the other, using the Sedeen Viewer software [79]. Corrections were made when considered necessary by both.

For complex cases, or when an agreement could not be reached, both the label and/or annotation were reevaluated by a third pathologist. Case classification followed the criteria previously described in section 2.2.1. Accordingly, cases with only minimal high-grade dysplasia areas (only one or two glands), or with areas of florid high-grade cytological features but without associated worrisome architecture, were kept in the low-grade dysplasia class, as well as cases with cytological high-grade dysplasia only seen on the surface. It is worth noting that some cases may be more difficult to grade and have to be decided on a case-by-case basis, preferentially by consensus. Additionally, as recommended by the World Health Organization (WHO), intramucosal carcinomas were included in the high-grade lesions set [39, 80].

Regarding the distribution of slide labels, while the annotated samples are considerably imbalanced, as seen in Figure 3.3(a) when combined with the non-annotated samples, the distribution of the labels is significantly more even. Figure 3.3(b) shows this final distribution and it is closer to what is seen in clinical practice. Moreover, it was important to fully annotate cases that are especially difficult or high-grade, so the model can learn more about these critical cases. The CRS1k dataset was used not only to develop the proposed methodology but also to evaluate the relevance of annotations in a model pretraining step: can a small set of annotated images leverage the overall performance of the weakly supervised model?



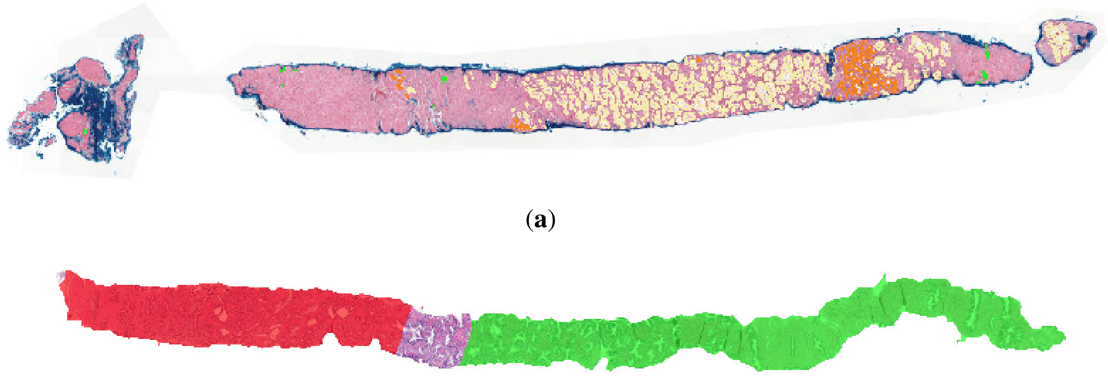
(**a**) Annotated samples

(**b**) All samples

Figure 3.3: Slide classes distribution on CRS1k dataset.

3.2.2 PANDA dataset

Besides the influence of the level of annotations, we also aimed to evaluate the proposed classification methodology on a larger dataset, also with a multiclass formulation, to investigate the impact of the dataset size on the performance of the algorithm. In this sense, we used 9,825 PCa biopsy slides from the dataset of the Prostate cANcer graDe Assessment (PANDA) challenge [81]. The



(**a**)

(**b**)

Figure 3.4: Example of WSI from the PANDA dataset, with manual segmentations overlayed: (a) sample from the Radboud UMC, with normal tissue in green, and tumour tissue in yellow and orange, accordingly its Gleason score; (b) sample from the Karolinska Institutet, with normal and tumour tissue in green and red, respectively.

available full training set consists of 10,616 WSI of digitized H&E stained PCa biopsies (we excluded cases with some type of error) obtained from two centres: the Radboud University Medical Centre (Figure 3.4a) and the Karolinska Institutet (Figure 3.4b), and includes both labelling and tissue segmentation. Each image is labelled with the corresponding ISUP grade and includes tissue annotation, differentiating tumour areas from normal tissue. The International Society of Urological Pathology (ISUP) grading system is the current score to grade PCa, which is based on the modified Gleason system (a score based on glandular architecture within the tumour), providing accurate stratification of PCa [82].

The PANDA dataset contains six different labels, corresponding to the five ISUP grades and the normal label, whereas the CRS1k dataset has only three different labels. Histopathological slides are quite different for different types of cancer, for instance, the quantity of tissue varies significantly. The images require some preprocessing that creates the tiles from the WSI. Such processing removes the background, and thus, tissue variations deeply affect the number of tiles on one slide. Table 3.2 displays an illustrative example of this, by comparing the number of tiles and the mean number of tiles per slide included in both datasets.

Table 3.2: Comparison between the number of tiles extracted from the slides of the PANDA and the CRS1k datasets.

Dataset	**# Slides**	**# Tiles**	**Mean # tiles per slide**
PANDA	9,825	253,291	25.78
CRS1k all	1,133	1,322,596	1,167.34
CRS1k annotated	100	211,235	2,112.35

The average number of tiles per slide is approximately 82x and 45x higher, respectively on the CRS1k annotated subset and on the complete dataset, when compared to the PANDA dataset.

Because of this variation in tissue proportion, despite having 8.6x more slides, the PANDA dataset still has around 5x fewer tiles.

3.3 Experimental results & discussion

In deep learning problems, it is not always trivial to determine the required dataset size to achieve the expected performance. Usually, it is expected that increasing the size of the dataset increases the model performance. However, this is not always true. Hence, to fully understand the impact of the dataset size in the computational pathology domain, the developed approach was trained on several subsets of the original PANDA training set with different sizes: 80, 160, 500, 1000, 2500, 5000 and 8348 (complete training set). For a fair comparison, all the experiments were evaluated on the same test set, which included 1477 slides (15% of the total dataset) independent from the training set. As can be seen in Table 3.3, the model is able to leverage more data in order to achieve better performance. Moreover, in line with these results, Campanella *et al.* [76] stated that for the MIL approach to work properly, the dataset must contain at least 10,000 samples. In our experiments, the performance with 5000 slides was already close to the best performance.

Table 3.3: Evolution of the model performance when trained on subsets of PANDA dataset with different sizes, keeping the test set size constant (n=1,477).

# Slides	# Tiles	QWK score	Accuracy
80	1,919	0.497	32.36%
160	3,851	0.586	37.71%
500	38,175	0.628	41.28%
1,000	25,757	0.692	47.66%
2,500	64,697	0.738	50.03%
5,000	129,734	0.771	58.43%
8,348	215,116	**0.789**	**59.40%**

To further infer the generalisation capability, an extra model was trained on 80 slides and evaluated on 20 slides randomly sampled from the 1477 test set. As seen in Table 3.4, as expected, when the size of the test set increases, the performance rapidly degrades, nursing the concerns and requirements for larger datasets. It is also worth noting that the performance of the model is considered poor in terms of accuracy scores. The QWK, on the other hand, records reasonable values. This

Table 3.4: Performance comparison of the model trained on a subset of PANDA dataset, when evaluated on test sets with different sizes.

Dataset	# Train slides	# Test slides	# Train tiles	# Test tiles	QWK score
PANDA	80	1,477	1,919	38,175	0.497
PANDA	80	20	1,919	579	**0.591**

difference in performance means that, while the model misclassifies about 40% of the slides, it classifies them with neighbour classes of the ground truth. One possible reason for this could be the noise present in the labels of this specific dataset.

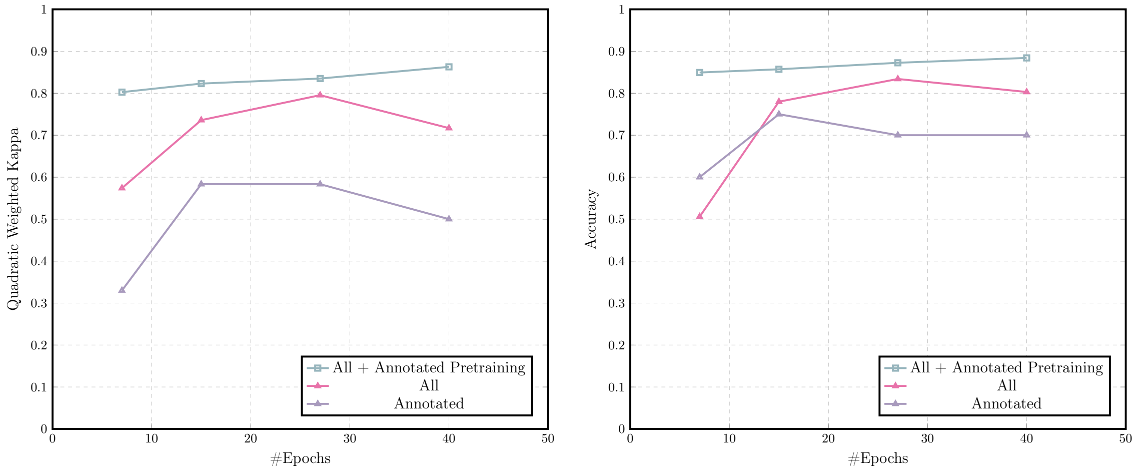
The third set of experiments explores the potential to leverage the annotations of a subset of data in order to improve the performance of the overall MIL method. Table 3.5 shows the results of the best epoch of each of the experiments.

Table 3.5: Performance of the model on the different experiments on the CRS1k dataset.

Dataset	pretrain	QWK	Accuracy	Convergence Time (Epoch)
CRS1k Annotated (n=100)	No	0.583	75.00%	6.5 hours (13)
CRS1k All (n=1,133)	No	0.795	84.17%	2 days and 19 hours (27)
CRS1k All (n=1,133)	Yes	**0.863**	**88.42%**	4 days (40)

There are notable performance gains in both the accuracy and the QWK score as the number of training samples increases. However, perhaps the most exciting performance gain is related to the pretraining of the backbone network on the 100 annotated samples for only two epochs before the start of the MIL training. This experiment is able to outperform the best epoch of the experiment without pretraining in only 7 epochs, in other words, 12 hours of training, with 84.94% accuracy and 0.803 QWK score. Moreover, these values kept increasing until the last training epoch, reaching an accuracy and QWK score of 88.42% and 0.863, respectively. The final results presented in Table 3.5 can be extended with sensitivity to lesions of 93.33% and 95.74% for the last two entries respectively. The training set comprises 874 samples (100 annotated and 774 non-annotated), whereas the test set has 259 WSI.

The results shown in Figures 3.5a and 3.5b, respectively for the QWK and the accuracy, are representative of the gains that both the number of samples and the use of annotations bring to the



(**a**) Quadratic Weighted Kappa score evolution

(**b**) Accuracy evolution

Figure 3.5: Performance evaluated on CRS1k dataset.

model. gains that both the number of samples and the use of annotations bring to the model. Moreover, the use of annotations appears not only to speed up convergence at high values but also to increase the model's ability to learn at further epochs.

The finding in these feasibility results supports the need for larger datasets. Not only that, but it also increases the confidence in the performance of weakly-supervised learning techniques, especially if it is possible to include at some point some supervised training to propel the performance even more. It is expected that these novel techniques and larger datasets converge to models that are closer to being deployed for clinical practice.

3.4 Summary

As studied in this chapter, increasing the number of WSI in the training data leads to an increase in performance, as does detailed annotation of, at least, part of the dataset. Therefore, the first and perhaps most crucial step for the further development of computer-aided diagnosis (CAD) systems for CRC is to establish a large and meaningful dataset.

Nonetheless, the construction of larger datasets with extensive annotations is not an easy and expeditious task. Hence, there is still a plethora of techniques to be explored with weakly labelled datasets. One of these tasks is known as multiple instance learning (MIL) and while it has been employed several times on these types of problems, it can still be improved to achieve more accurate results. As shown in Section 3.3, the performance of MIL systems is greatly improved with a pretraining on the 10% of the dataset that is annotated.

Since the main goal of deep learning (DL) in computational pathology is to develop a solution that can be deployed in a clinical environment, it is important to develop it in a similar fashion to the clinical practice, in other words, to handle the same type of data given to pathologists, WSI. In that sense, this work is considerably more in line with the end goal of CAD systems for computational pathology: our proposal can be directly applied to a lab workflow. However, in order for these approaches to be used in practice, it is important that researchers develop techniques to inform pathologists about the spatial location that was most responsible for the diagnosis and to explain the reasons for the prediction. Interpretability and explainability have been explored in medical applications of DL [83], and so they should be present in Computational Pathology use cases [84], such as CRC diagnosis. The ultimate goal is to create transparent systems that medical professionals can trust and rely on. The work presented in the next chapter tries to solve this requisite.

Chapter 4

Semi-supervised & interpretable approach for CRC grading

Author Contributions

The research work described in this chapter was conducted in collaboration with Pedro C. Neto and the IMP Diagnostics team, under the clinical supervision of Isabel M. Pinto, and the technical supervision of Jaime S. Cardoso. The author of this thesis contributed to this work on problem conceptualisation, data curation, the preparation of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- P.C. Neto*, S.P. Oliveira*, D. Montezuma*, J. Fraga, L. Ribeiro, S. Gonçalves, I.M. Pinto and J.S. Cardoso. iMIL4PATH: A Semi-Supervised Interpretable Approach for Colorectal Whole-Slide Images. *Cancers*, 14(10):2489, 2022

*Shared co-first authorship

The main goal of this work was to develop a system that is one step closer to being used by pathologists in their daily routine, which includes the following contributions: (1) an improved method that combines weakly and supervised learning methods to construct a novel system to diagnose colorectal cancer (CRC) from digitised Haematoxylin-Eosin (H&E) stained slides, with high ACC and sensitivity; (2) a thorough comparison of several aggregation methods to increase the number of tiles used for predictions, which can reduce the number of false positives; (3) extensive experiments on an extended version of the publicly available CRS1k dataset; (4) a study of the model's interpretability and capability to self-explain the diagnosis areas through the reconstruction of the slide with individual tile predictions without requiring added training. This latter contribution can be especially useful to guide pathologists' attention towards the most relevant tissue areas within each WSI; and (5) evaluation of domain generalisation on two public colorectal WSI datasets.

4.1 Problem definition

Automated diagnosis of CRC histological samples requires the use of images with large dimensions. In addition, the labelling of these images is difficult, expensive, and tedious. Therefore, the availability of WSIs is limited, and, when available, they often lack meaningful labelling: while slide-level diagnoses are generally available, detailed spatial annotations are almost always lacking. A prototypical example is the CRS1k dataset, presented in Chapter 3, containing 1133 colorectal H&E samples with slide-level diagnoses.

Thus, following previous work on CRC diagnosis, and on automatic diagnostic systems in general, we assumed a semi-supervised learning procedure. A slide $\mathscr{S}$ can be viewed as a set of tiles $\mathscr{T}_{s,n}$, where s is the index of the slide and $n \in \{1, \cdots, n_s\}$ is the tile number. We assumed that there were individual labels $C_{s,n} \in \{C^{(1)}, \cdots, C^{(K)}\}$ for the tiles within the slide. The classes $C^{(k)}$ were considered ordered and correspond to the different diagnostic grades. For a strongly annotated slide, each corresponding tile label $C_{s,n}$ is known. In a weakly annotated slide, there is no access to those labels and they remain unknown during training. A weakly annotated slide has only a single label for the entire set (bag) of tiles, see Figure 4.1.

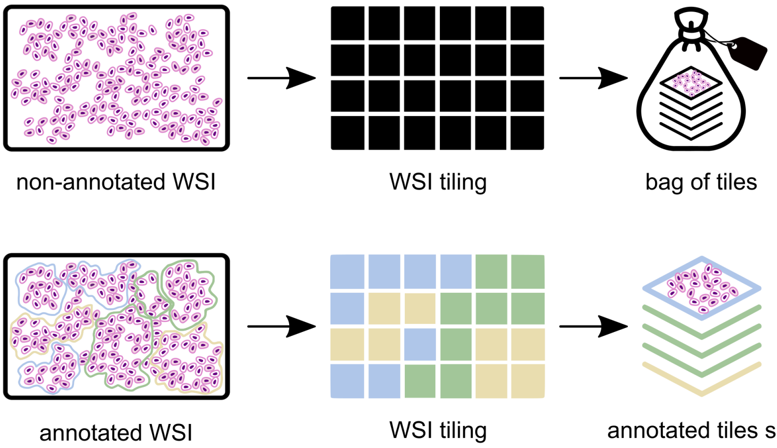


Figure 4.1: Labelling scheme: weakly annotated slides (above) have only a global label, from the pathology report, whereas a strongly annotated slide (below) has labels for each individual tile, retrieved directly from the pathologists' spatial annotations.

Finally, we assumed that the slide label C_s is the worst-case of the tile labels:

$$C_s = \max_n \{C_{s,n}\}.$$

If there is one tile in the set of tiles extracted from a slide that is classified as high-grade dysplasia, then the slide label will be the same. Second, if there is no dysplasia in any of the tiles, then the slide label is non-neoplastic. This learning setting corresponds to a simple generalisation of multiple-instance learning (MIL), from the binary problem to the ordinal classification problem.

4.2 Methodology

4.2.1 Data pre-processing

The H&E slide pre-processing includes an automatic tissue segmentation with Otsu's thresholding [85] on the saturation (S) channel of the HSV colour space, obtaining the tissue regions clearly separated from the whitish background. This step, performed on the 32× downsampled slide, returned the mask used for tile extraction. Tiles with a size of 512 × 512 pixels (Figure 4.2) were then extracted from the slide with original dimensions (without downsampling) at the maximum magnification (40×), provided they were completely within Otsu's mask.

The tile size was chosen by empirical experiments, which showed that 512 × 512 is the best trade-off between memory and performance. Larger sizes represent more context and tissue per tile, at the expense of memory and computation time. Using tiles with a full area of tissue reduces the number of instances by not including the tissue at the edges, which drastically decreases the computational cost, without hurting the performance of the model. Since the original size of each WSI and the amount of tissue per slide varies greatly, the number of tiles extracted also varies a lot.

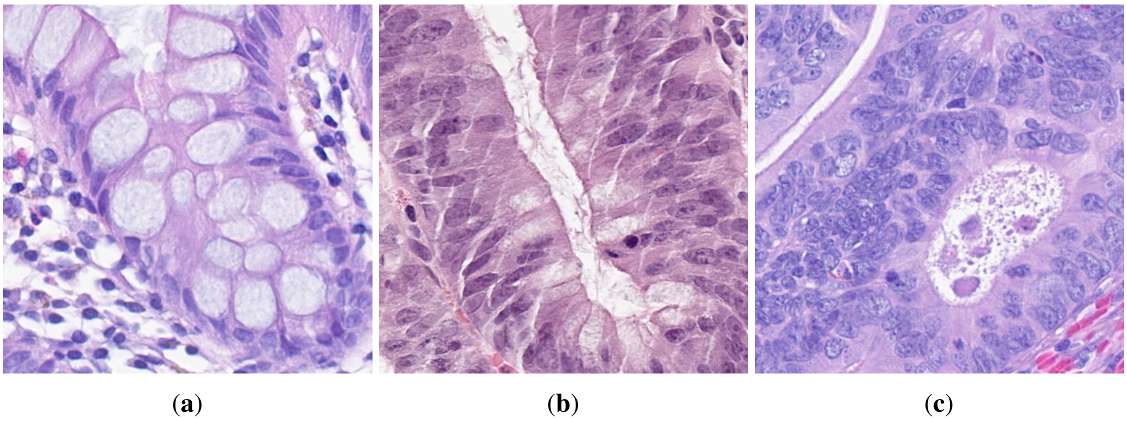


Figure 4.2: Examples of tiles with 512 × 512 pixels (40× magnification), representing each class: non-neoplastic (a), low-grade dysplasia (b) and high-grade dysplasia (c).

4.2.2 Model architecture

In Chapter 3, we presented an approach that has already introduced some modifications to the MIL method proposed by Campanella et al. [76]. Here, we further extended those modifications and adjusted them to better fit the requirements of an automatic CRC diagnosis system. In Figure 4.3, the architecture of our system is introduced, which is mainly composed of a supervised pre-training phase, to leverage the samples that include annotations (≈9% in the adopted dataset), a weakly supervised training phase, where all the dataset is used, and a final stage with feature extraction and training of an aggregation method. While supervised learning requires extensive use of annotations, we used an approach that merges weakly and supervised learning, needing less than one annotated per ten non-annotated samples, while performing on par with the state-of-the-art methods.

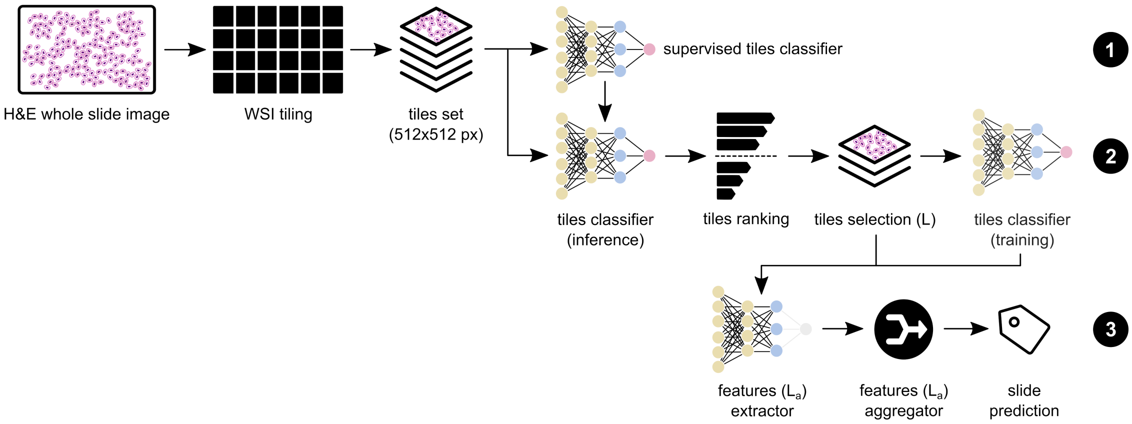


Figure 4.3: Proposed workflow for colorectal cancer diagnosis on whole-slide images, as a three-step method: (1) supervised tile classifier, using the annotated slides set; (2) weakly supervised tile classifier (initialised with the supervised weights), selecting the most relevant tiles by ranking by the expected values; and (3) whole-slide prediction by aggregating the features (obtained with the previous CNN model) extracted from the most relevant tiles.

(1) Supervised pre-training

The supervised training phase leverages the annotations of all tiles in the strongly annotated WSIs to train a ResNet-34 [86], which classifies into the three diagnostic classes by minimising a loss function based on the quadratic weighted kappa (QWK). The QWK loss is appropriate for ordinal data because it weights misclassifications differently, according to the equation:

$$\kappa = 1 - \frac{\sum_{i,j=1}^{n} w_{ij} x_{ij}}{\sum_{i,j=1}^{n} w_{ij} m_{ij}} \tag{4.1}$$

where K is the number of classes, w_{ij} belongs to the weight matrix, x_{ij} belongs to the observed matrix and m_{ij} are elements in the expected matrices. The $n \times n$ matrix of weights w is computed based on the difference between the actual and predicted class, as follows:

$$w_{i,j} = \frac{(i-j)^2}{(n-1)^2} \tag{4.2}$$

As shown in Chapter 3, pre-training on a small set of data with supervised learning leads to faster convergence and also better results on all metrics.

During our studies, we found that the approach presented in Chapter 3, used as the baseline, could be improved with increased pre-training. Compared to the weakly supervised training phase, the supervised training was significantly faster to complete an epoch. In addition, thus, with a trivial computational cost, it was possible to increase the number of pre-training epochs from two to five. This change positively impacted the algorithm's performance on the test set.

(2) Weakly-supervised training

The weakly-supervised training phase uses all the available training slides and only slide-level labels to complete the training of the deep network. The model is used to infer all tiles classes and then, based on those predictions, the tiles are ranked. We followed the approach of Chapter 3, which performed a tile ranking based on the expected value of the predictions.

For tile $\mathscr{T}_{s,n}$, the expected value of the score is defined as

$$\mathbb{E}(\hat{C}_{s,n}) = \sum_{i=1}^{K} i \times p\left(\hat{C}_{s,n} = C^{(i)}\right) \tag{4.3}$$

where $\hat{C}_{s,n}$ is a random variable on the set of possible class labels $\{C^{(1)}, \cdots, C^{(K)}\}$ and $p\left(\hat{C}_{s,n} = C^{(i)}\right)$ are the K output values of the neural network.

Despite ranking all tiles, only the worst tile (from a clinical point of view), i.e., the one with the highest expected value, was used to optimise the network weights. From the perspective of MIL, this corresponds to an aggregation function based on the maximum of the observations of the bag. This can slow down the training and even make it more unstable, especially in the first epochs, when the tile predictions are still very noisy. Therefore, instead of using only the tile with the highest expected value, we considered the generalisation of max function, $top_L(.)$, which keeps the first L tiles with the highest score.

By changing the number of tiles used to optimise the network, we also increased the variability and possible changes between epochs. For example, it is more likely that none of the selected tiles will change if only one is selected. However, by selecting $L > 1$, we increased the probability that the tiles will change in successive epochs while maintaining the stability of the training. Similar to the previous change, this one also resulted in a more robust model than the baseline.

After the model's performance with the one tile MIL aggregation ($L = 1$), and also after an in-depth analysis with pathologists, we decided that the WSI on the adopted dataset contained, on average, enough information to use at least five tiles. The definition of sufficient information was determined by the number of tiles that contained information related to the slide diagnosis. For instance, if a WSI label was from a high-grade dysplasia, only tiles with information of a possible high-grade dysplasia were considered to be useful, and, thus, tiles with only normal tissue should not be used to optimise the network weights. The value of L was then set to $L = 5$, since this value represents a significant increase from $L = 1$ and it does not use (in the majority of the slides) tiles with a potentially distinct diagnosis from the slide diagnosis.

There is growing concern regarding semi-supervised methods' overconfident behaviour. There have also been works that aimed at addressing this problem on other tasks through network calibration [87]. However, in this specific scenario, it is believed that an over-confident model in severe cases leads to fewer false negatives and higher sensitivity. In addition, thus, it is not seen as a potential problem of the model. On the other hand, the proposed aggregation approaches in the following section show properties that mitigate the risk of overconfidence.

(3) Feature extraction and aggregation

Regarding the max-pooling aggregation on multiple-instance learning approaches, one can argue about its robustness, such as the discussion presented, for example, by Campanella et al. [76], since it is a biased aggregation towards positive labels, and one small change can impact the entire tile classification. Hence, we studied the incorporation of shallow aggregation structures into our model to improve the results given by max-pooling.

It was found that the use of only one tile leads to a bias of the network towards more aggressive predictions. For this reason, we followed a strategy that has been adopted in other domains: the CNN was trained end-to-end as a classification model (using a combination of supervised pre-training and weakly supervised learning) and, after training, the fully-connected layer was removed. The network then outputs a feature vector for each tile, which were aggregated and used to train a supervised method at the slide level to improve the classification capabilities of the system. For this problem, we chose to use L_a feature vectors, corresponding to the L_a tiles with the highest expected value for the score (Equation (4.3)). In our experimental study, L_a was empirically set to 7, representing a good trade-off between additional information and the introduction of noise.

To compare different classifiers, we selected six aggregation models to test within the proposed framework:

- A support-vector machine (SVM) with a radial basis function kernel and a C of 1.0;
- A K-nearest neighbour (KNN) with a K equal to 5;
- A random forest (RF) with a max. depth of 4 and the Gini criterion;
- AdaBoost and XGBoost with 3000 and 5000 estimators, respectively;
- Two distinct multi-layer perceptrons (MLP) with two layers; the first MLP with layers of 75 and 5 nodes - MLP(75;5) - and a second one with layers of 300 and 50 nodes - MLP(300;50).

Besides these individual models, we also combined the previous ones into voting schemes, following a soft voting technique based on the probabilities of each model: SVM and KNN; and SVM, RF and KNN.

4.2.3 Interpretability assessment

Nowadays, deep learning models are becoming more complex and opaque. This is alarming, especially when we look at the potential applications of these models in the medical domain. If they are designed to work all by themselves, we need to ensure they are completely transparent. In addition, if they are to be used as a tool to help pathologists make a particular diagnosis and improve their confidence, then they must at least direct their focus to the areas relevant to the diagnosis. It is necessary to understand the behaviour of the model to extend the validity of the typical analysis supported by metrics such as ACC, QWK and sensitivity.

Therefore, a method was developed to generate visual explanations of model predictions. This method was constructed with the following ideas in mind:

(a) for large WSI images, it is helpful to direct the pathologist to specific areas of high relevance;

(b) since the model was not trained on tile ACC, it is sufficient if it is able to highlight a subset of the relevant tiles in a given area of interest;

(c) and since, for the slide prediction, the model requires the processing of all tiles, creating a map of tile predictions does not require additional computational cost or idle time for the pathologist.

Given these ideas, the proposed method leverages the evaluation of the MIL method, which processes every tile in the WSI. Even if the tile is not selected for aggregation, it will be processed by the backbone network, which results in a tile-score prediction (c). These tile-level predictions are converted into colours based on the result of the *Argmax* function applied to their scores. Afterwards, these colours can be spatially allocated based on a remapping strategy from the tile at the original slide magnification to a 32× reduced WSI (a). In addition, while some of the predictions might be misclassified, the entire reconstruction of the WSI shall be sufficient to redirect the attention of pathologists towards the areas of interest (b).

4.2.4 Training details

We trained the convolutional neural network using Pytorch with the Adaptive Moment Estimation (Adam) optimiser, a learning rate of 6×10^{-6}, a weight decay of 3×10^{-4} and a batch size of 32, for both the strongly and weakly supervised training steps. For the inference step in the weakly supervised approach, we used a batch size of 256 and the network was set to evaluation mode. The method's performance was evaluated at the end of each epoch to select the best model based on the QWK. The training was conducted on a single Nvidia Tesla V100 (32 GB) GPU for 5 strongly supervised epochs and 30 weakly supervised epochs.

Seven feature vectors from the worst tiles were concatenated to train the aggregation methods. This led to a feature vector of size 3584. Afterwards, these feature vectors were used as input to train the aggregators developed with the help of the scikit-Learn library. In addition to this, the MLP aggregator required additional training parameters. It was optimised with stochastic gradient descent, mini-batches of 32 samples and an initial learning rate of 10^{-3}.

4.3 Datasets

4.3.1 CRS1k & CRS4k datasets

This work was developed with the CRS1k dataset, described in Chapter 3, and with an extended version (CRS4k) that includes approximately 4× more samples (4433 colorectal H&E slides), of which a subset ($n = 400$) is also annotated according to the guidelines followed on Chapter 3 [88].

The CRS1k dataset was used for the selection and comparison of aggregation methods. The CRS4k was used to create a more robust test set and a larger training set to train the methods previously selected. Table 4.1 summarises the class distribution of annotated and non-annotated data, including the number of tiles obtained after the pre-processing described in Section 4.2.1.

Table 4.1: Colorectal dataset summary, with the number of slides (annotated samples are detailed in parenthesis) and tiles distributed by class: non-neoplastic (NNeo), low-grade (LG) and high-grade (HG) lesions.

		NNeo	**LG**	**HG**	**Total**
CRS1k dataset	*#slides*	300 (6)	552 (35)	281 (59)	1133 (100)
	# annotated tiles	49,640	77,946	83,649	211,235
	# non-annotated tiles	-	-	-	1,111,361
CRS4k dataset	*#slides*	663 (12)	2394 (207)	1376 (181)	4433 (400)
	# annotated tiles	145,898	196,116	163,603	505,617
	# non-annotated tiles	-	-	-	5,265,362

The CRS4k dataset represented an increase in the approximate average number of non-annotated tiles per slide from 1075 to 1305. However, the approximate average number of tiles per annotated slide decreased from 2112 to 1264. This might represent a tougher task to solve on this dataset.

4.3.2 TCGA & PAIP datasets

Two external datasets were also included for a domain generalisation evaluation. The first is composed of samples of the TCGA-COAD [89] and TCGA-READ [90] collections from The Cancer Imaging Archive [91], containing mostly surgical resection samples (Figure 4.4a), excluding slides with pen markers, large air bubbles over tissue, tissue folds and other artefacts in large areas of the slide. We ended up with 232 samples reviewed and validated by the pathologist team, from which 230 of them were diagnosed as high-grade lesions, one as a low-grade lesion and one as non-

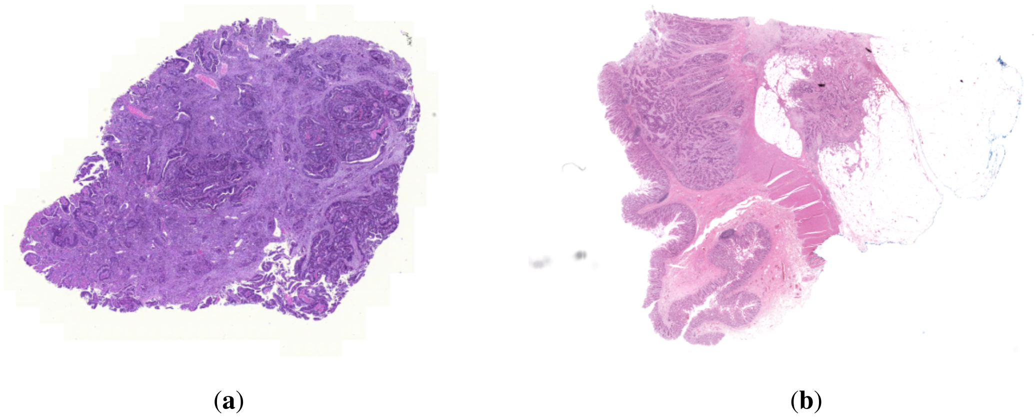

(a) (b)

Figure 4.4: Example of slides from the TCGA (a) and the PAIP (b) datasets.

neoplastic. The second external validation set is composed of 100 H&E slides from the Pathology AI Platform [92] colorectal cohort, which includes all the cases with more superficial sampling of the lesion (Figure 4.4b), to better compare with our CRS4k dataset. All samples were also reviewed and validated as high-grade lesions by the pathologists team.

4.4 Experimental results & discussion

4.4.1 CRS1k dataset evaluation

We evaluated our approach with the MIL-aggregation (max-pooling) and with eight different types of aggregators, as seen in Table 4.2. Approaches with tile aggregation at inference are, in general, better than the baseline method for CRC diagnosis from WSI. From those, the MLP aggregator using seven feature vectors outperformed the baseline and all the other aggregation schemes. Other approaches with aggregation showed overall good results, but are not on par with the MLP approach. In addition, the MLP has an increased specificity (by reducing the number of false positives) while avoiding a significant degradation of the sensitivity. The SVM and the KNN have the best results from the remaining approaches, with the KNN achieving the same specificity as MLP. Finally, the two voting approaches show notable improvements over the stand-alone aggregation methods, with the combination of SVM and KNN beating all the previous approaches on nearly every metric and achieving the same specificity of the MLP.

Table 4.2: Comparison of feature aggregation methods against the approach of Chapter 3, on the same test set. Both the ACC and the QWK score were computed for a three-classes problem, whereas the sensitivity and the specificity were computed for a binary problem by considering the LG and HG classes as a unique class.

Method	**Annotated Samples**	**Training Tiles (L)**	**Aggregation Tiles (L_a)**	**QWK**	**ACC**	**Sensitivity**	**Specificity**
Model of Chapter 3	100	1	1	0.863	88.42%	0.957	-
Supervised baseline	100	-	1	0.027	29.73%	0.449	0.796
Max-pooling	100	5	1	0.881	91.12%	**0.990**	0.852
MLP (75;5)	100	5	7	**0.906**	**91.89%**	0.980	**0.981**
SVM				0.887	90.35%	0.971	0.944
KNN				0.890	90.35%	0.971	**0.981**
RF				0.878	89.57%	0.966	0.963
AdaBoost				0.862	88.03%	0.961	0.907
XGBoost				0.879	89.58%	0.961	0.963
SVM + KNN	100	5	7	0.898	91.12%	0.971	**0.981**
SVM + RF + KNN	100	5	7	0.893	90.73%	0.971	**0.981**

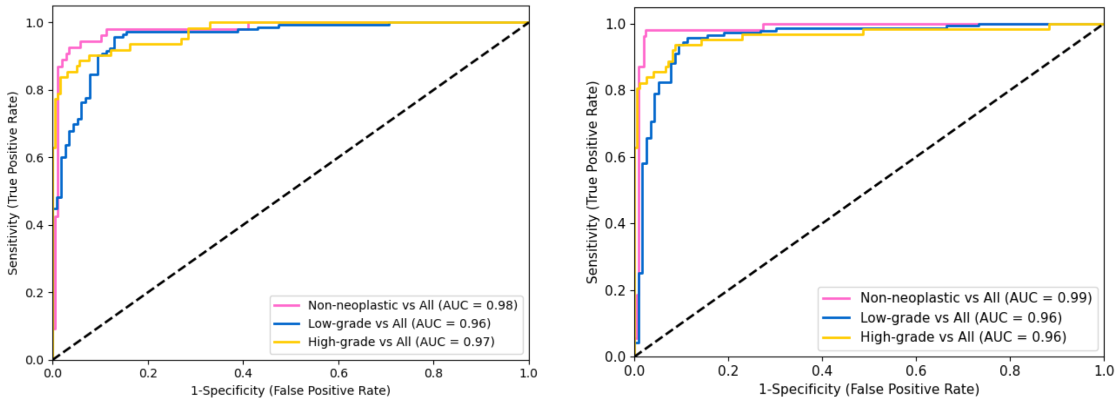
In Table 4.3 we can see the confusion matrix for the MLP (75;5), which was the best-performing method. It is worth noting that MLP (75;5) did not fail any prediction by more than one consecutive class (for instance, predicting HG as NNeo or vice-versa). This ensures that HG lesions are at least classified as LG or HG, which can be seen as a desired feature of the model. When analysed

Table 4.3: Confusion matrix of the MLP (75;5) in the multiclass setup, using the CRS1k test set (259 samples), with non-neoplastic (NNeo), low-grade (LG) and high-grade (HG) classes.

		Actual class		
		NNeo	*LG*	*HG*
Predicted	*NNeo*	**53**	4	0
	LG	1	**137**	14
	HG	0	2	**48**

as a binary classification problem, it is possible to observe that only 5 samples out of 259 are misclassified. This means that the proposed model shows a binary ACC of 98.1%.

We also plotted the receiver operating characteristic (ROC) curve of the baseline and the best aggregation method. It was intended to verify not only their area under the curve (AUC), but the performance of the model per class. Once more, as seen in Figure 4.5, the MLP method outperformed the other approach in almost every class. Moreover, as expected, it is easier to distinguish non-neoplastic cases from the rest, than to decide between low- and high-grade lesions.



(**a**) Max-pooling

(**b**) MLP (75;5)

Figure 4.5: ROC curves for max-pooling and MLP (75;5) aggregator.

4.4.2 CRS4k dataset evaluation

We aimed to understand the relevance of adding additional annotated and non-annotated data to the performance of the algorithm. Hence, the results in Table 4.4 show the performance of the model with the CRS1k dataset, with increased annotated samples and with an increased number of non-annotated samples. In addition, we further introduced another version of the MLP aggregator, which comprises different layer dimensions, to test if the increased number of samples required more complex models. Surprisingly, the results did not evolve as expected, since the performance was negatively affected by the increase in the size of the dataset. This is likely caused by the

Table 4.4: Model performance evaluation with increasing training sets and/or annotated samples (in parenthesis). Both the ACC and the QWK score were computed for a three-classes problem, whereas the sensitivity and the specificity were computed for a binary classification problem by considering the LG and HG classes as one unique class.

Method	**Training Samples**	**Test Samples**	**Aggregation Tiles (L_a)**	**QWK Score**	**ACC**	**Sensitivity**	**Specificity**
Max-pooling	874 (100)	259	1	0.881	91.12%	**0.990**	0.852
MLP (75;5)			7	**0.906**	**91.89%**	0.980	**0.981**
MLP (300;50)			7	0.885	91.12%	0.966	**0.981**
Max-pooling	1174 (400)	259	1	**0.874**	**91.12%**	**0.985**	0.907
MLP (75;5)			7	0.838	86.49%	0.946	0.926
MLP (300;50)			7	0.850	87.26%	0.941	**0.944**
Max-pooling	4174 (400)	259	1	**0.834**	**89.96%**	**0.980**	0.870
MLP (75;5)			7	0.810	83.78%	0.922	0.889
MLP (300;50)			7	0.816	83.01%	0.927	**0.926**
Max-pooling	3424 (400)	1009	1	0.884	89.89%	**0.992**	0.815
MLP (75;5)			7	0.871	88.89%	0.982	0.839
MLP (300;50)			7	**0.888**	**90.19%**	0.988	**0.857**

overfitting of the aggregation method to the training data, which leads to a poor generalisation capability on test data.

In an attempt to fully understand the reason behind the performance drop, we created new training and test sets, with the latter being roughly 3.89 times larger than its previous version. The results of this new experiment are presented in the last three rows of Table 4.4. The improvements shown by training and evaluating on these larger training and test sets indicate that the smaller test set used for evaluation in Table 4.4 might have noisy labels or not be representative enough. Hence, the proposed model seems to be robust when given more training data and a larger test set. Finally, the superior performance of the aggregators on the new dataset split shows its relevance to the construction of well-balanced and accurate algorithms.

4.4.3 Domain generalisation evaluation

The development of medical-oriented deep neural networks is usually strongly influenced by the data source. Colour, saturation and image quality are important factors for the performance of these networks. Moreover, the type of sample is also important; for instance, despite the shared similarities, biopsies and surgical resection samples are quite distinct from each other. Hence, to evaluate the domain generalisation, the proposed method trained on CRS4k biopsies samples was evaluated on two external public datasets. The results are presented in Tables 4.5 and 4.6.

Table 4.5: Model performance evaluation on the TCGA test set.

Method	ACC	Binary ACC	Sensitivity
Max-pooling	**71.55%**	**80.60%**	**0.805**
MLP (75;5)	61.20%	75.43%	0.753
MLP (300;50)	58.62%	74.13%	0.740

Table 4.6: Model performance evaluation on the PAIP test set

Method	ACC	Binary ACC	Sensitivity
Max-pooling	**99.00%**	**100.00%**	**1.000**
MLP (75;5)	77.00%	98.00%	0.980
MLP (300;50)	77.00%	98.00%	0.980

As expected, due to its high sensitivity, and since almost all cases evaluated are high-grade cases, the max-pooling approach achieves the best results in terms of multiclass ACC, binary ACC and sensitivity. Regarding the TCGA dataset, these results can be explained by the fact that these samples are mostly from surgical resections, with bigger portions of tissue, whereas ours are from biopsies/polipectomies. Moreover, the datasets are somewhat different regarding the represented classes, with TCGA containing more poorly differentiated and mucinous adenocarcinomas, which are underrepresented in the CRS4k training set. Finally, the lower tissue image quality, when compared to the CRS4k dataset, may also explain this performance drop.

Regarding the better results on PAIP dataset, it can be explained by the better quality of the WSIs and a H&E staining colour being closer to CRS4k dataset. Moreover, although all PAIP slides seem to derive from surgical specimens, the sampling of the neoplasias was more superficial in most of the cases used (representing mostly mucosa and submucosa layers) as opposed to TCGA samples, in which many samples showcased all colonic layers (mucosa, submucosa, muscular and adipose tissue), differing greatly from the biopsies and polipectomies of the CRS4k dataset.

Domain generalisation is a complex topic that derives from several variables. In our scenario, the model displays a good capability to comprehend the content of a WSI collected on another lab, as seen in Table 4.6. However, there is still work to be done on the generalisation capability between strong colour differences and the capability of also assessing surgical specimen samples.

4.4.4 Interpretability Assessment

In order to assess how the model classified each tile and to better understand the class distribution within each case, we retrieved the single tile predictions and assigned them to their respective position on the slide, creating a predictions map. For each case, we also retrieved the worst tile, in clinical terms (Figure 4.6(d)). This experiment was conducted with slides from the annotated data subset (Figure 4.6(a)), using the model trained on the full dataset and further analysed by pathologists. By constructing these maps, we allowed pathologists to understand the reasoning of the model behind a slide prediction. Moreover, if necessary, it can guide and direct the focus of the pathologist to relevant areas in order to improve the overall workflow in clinical environments. As can be seen in Figure 4.6(c), although the model was not trained for segmentation, nor focused on individual tile-label prediction, the results are quite accurate in terms of lesion localisation, when compared to the ground truth (Figure 4.6(b)). On slides classified as NNeo (top) and LG (middle), the precision of the tile classification compared to the pathologists' masks is rather impressive. For the HG slide (bottom), despite the lower density of tiles predicted as HG, the model was capable of capturing the majority of the fragments affected, as we verified on the maps generated for all the annotated slides.

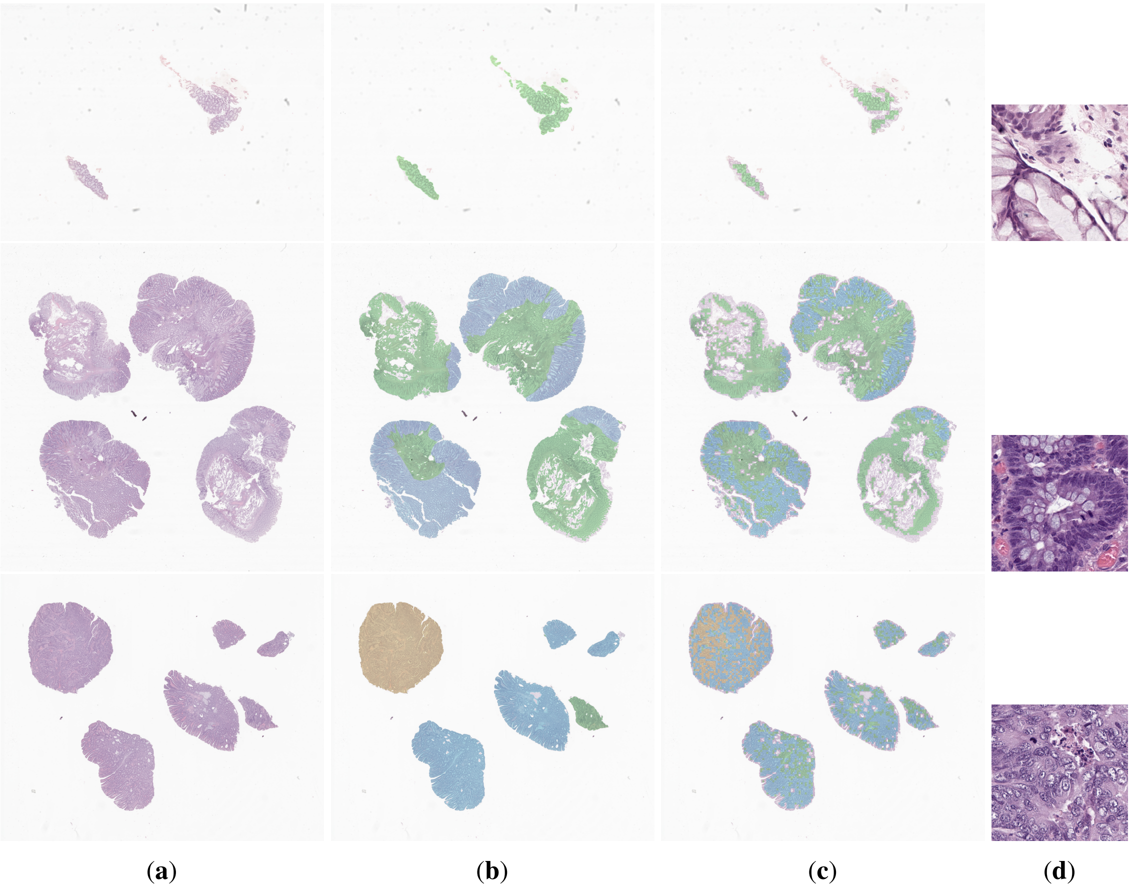


Figure 4.6: Examples of a model prediction map for each class, from the annotated data subset: a non-neoplastic case (top), a low-grade lesion (middle) and a high-grade lesion (bottom). Each column has the slides examples (a), the ground-truth annotation (b), the map with the tile predictions (c) and the most relevant tile (512 × 512 px), with the worst clinical class (d). The non-neoplastic, low-grade and high-grade regions are represented in green, blue and yellow, respectively.

4.5 Summary

In this chapter, we presented an improved framework for CRC diagnosis. Not only are metrics such as ACC and the QWK better, but the sensitivity achieves values close to the maximum. Furthermore, the method was trained and tested on an extended version of one of the largest datasets of colorectal histological samples publicly available, which increases the robustness of the results. Finally, the model was validated on external datasets for domain generalisation. Despite the performance drop in the TCGA dataset, when compared to CRS4k dataset, and some misclassifications in the PAIP dataset, it is worth noting that the model can detect high-grade lesions reasonably well, even in sets with many distinct properties compared to the one used for training.

Although achieving remarkable performance, medical applications of DL-based methods have been severely criticised due to their natural black-box structure. Here, we presented a model that attempts to support slide decision reasoning in terms of the spatial distribution of lesions. However, questions such as model robustness and how can we include AI models in the digital workflow of the pathology lab should be also addressed. In the next chapter, we attempt to answer these requisites to integrate models into the clinical practice to assist and ease the workload of pathologists.

Chapter 5

From the deep learning model to a clinical software prototype

Author Contributions

The research work described in this chapter was conducted in collaboration with Pedro C. Neto, João Romão and the IMP Diagnostics team, under the clinical supervision of Isabel M. Pinto, and the technical supervision of Jaime S. Cardoso. The author of this thesis contributed to this work on problem conceptualisation, software design and requirements analysis, data curation and pre-processing, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- P.C. Neto[*], D. Montezuma[*], S.P. Oliveira[*], D. Oliveira, J. Fraga, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, S. Reinhard, I. Zlobec, I.M. Pinto and J.S. Cardoso. A CAD System for Colorectal Cancer from WSI: A Clinically Validated Interpretable ML-based Prototype. *Nature Communications*, 2022 [submitted, waiting for decision]
- J. Romão, D. Montezuma, S.P. Oliveira, P.C. Neto, J. Monteiro, L. Ribeiro, S. Gonçalves, A. Monteiro, I.M. Pinto and J.S. Cardoso. Computer-aided tool for CRC diagnosis: from the AI model to the clinical software prototype. *In 18th European Congress on Digital Pathology (ECDP)*, SY12.04, 2022

[*]Shared co-first authorship

The current state-of-the-art CAD systems are based on deep learning approaches. These systems rely on large volumes of data to learn how to perform a given task. Particularly in computational pathology, this high volume of data, in addition to the massive resolution of the images, creates a significant bottleneck of DL approaches that the WSI decomposed into tiles. Hence, in this chapter, we introduce a bigger colorectal WSI dataset and an efficient tile sampling strategy that is performed one single time without any sacrifice in the predictive performance. Moreover, to bring the proposed CAD system closer to clinical practice, and to infer its capabilities to aid pathologists, we developed a prototype based on a server-side web application. We further collected information on the misdiagnoses and the pathologists' feedback.

5.1 Methodology

As concluded in Chapter 3, increasing the number of WSI in training data leads to increased performance, as does detailed annotation of, at least, part of the dataset. Also, to translate ML models to the clinic, they should be robust and trained on as many different cases as possible, to account for the huge data variability. Thus, in this work, we propose a new dataset of colorectal samples with more than 10,000 WSI (the CRS10k dataset), with about 9% of samples annotated by pathologists, which adds even more data compared to the extended dataset of Chapter 4. Following the latter work, we used the learning framework with better performance across the validation datasets (Figure 5.1). Briefly, it takes advantage of a supervised learning step on annotated data, that is then used to infer the class of non-annotated tiles. Then, all tiles are ranked based on the expected value and the top one is used to predict the slide label. With such a bigger dataset, we propose a tile sampling strategy, to leverage training efficiency.

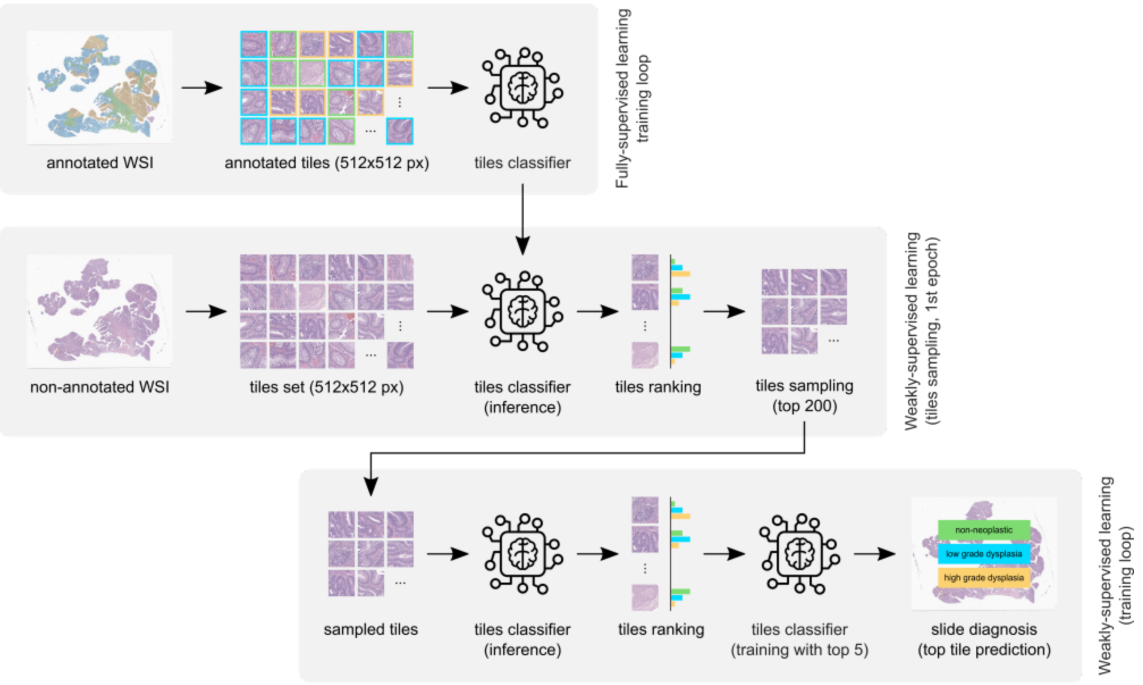

annotated WSI
annotated tiles (512x512 px)
tiles classifier
Fully-supervised learning training loop
non-annotated WSI
tiles set (512x512 px)
tiles classifier (inference)
tiles ranking
tiles sampling (top 200)
Weakly-supervised learning (tiles sampling, 1st epoch)
sampled tiles
tiles classifier (inference)
tiles ranking
tiles classifier (training with top 5)
non-neoplastic
low grade dysplasia
high grade dysplasia
slide diagnosis (top tile prediction)
Weakly-supervised learning (training loop)

Figure 5.1: Overall scheme of the proposed methodology containing the mix-supervision framework used for diagnosing colorectal samples from WSI.

5.1.1 Tile Sampling

Adding more slides to the train set implies increasing the number of tiles seen by the model (which can amount to more than 1000 per slide) and, consequently, the training time. In fact, in the weakly-supervised step, the non-annotated tiles are fed to the network and only then sorted to select the worst one for training. With this setup, all the tiles within a slide have to be seen by the model, at each epoch. However, in each slide, only the tiles corresponding to lesions are relevant to the decision. Furthermore, as the model converges, it is expected to select the important tiles better, settling on a relatively stable set across epochs. Thus, if the original set of non-annotated tiles is reduced to a smaller set at the end of the first inference loop in the weakly-supervised phase, training will become more efficient.

In this sense, we first assess the average number of relevant tiles for the decision, by evaluating the heterogeneity of tiles on the top rank across the first weakly-supervised epochs. The number of selected tiles considers a trade-off between computational cost and information potentially lost, and for that reason, it resulted in empirical experiments. Then, we set the sampling threshold to a maximum and evaluated its impact on training and validation sets.

This sampling approach is just applied to the non-annotated data for training, since we want to keep all the annotated data for model supervision. It is also not used for testing, since at that phase we need to assess all tiles within each slide.

5.1.1.1 Confidence Interval

In order to quantify the uncertainty of a result, it is common to compute the 95 percent confidence interval. In this way, two different models can be easily understood and compared based on the overlap of their confidence intervals. The standard approach to calculating these intervals requires several runs of a single experiment. As we increase the number of runs, our interval becomes narrower. However, this procedure is impractical for the computationally intensive experiments presented in this document. Hence, we use an independent test set to approximate the confidence interval as a Gaussian function [93]. To do so, we compute the standard error (*SE*) of an evaluation metric *m*, which is dependent on the number of samples (*n*), as seen in Equation 5.1.

$$SE = \sqrt{\frac{1}{n} \times m \times (1-m)} \tag{5.1}$$

For the SE computation to be mathematically correct, the metric *m* must originate from a set of Bernoulli trials. In other words, if each prediction is considered a Bernoulli trial, then the metric should classify them as correct or incorrect. The number of correct samples is then given by a Binomial distribution $X \sim (n, p)$, where p is the probability of correctly predicting a label, and n is the number of samples. For instance, accuracy is a metric that fits all these constraints.

Following the definition and the properties of a Normal distribution, we compute the number of standard deviations (z), known as a standard score, that can be translated to the desired confidence (c) set to 95% of the area under a normal distribution. This is a well-studied value, which is approximately $z \approx 1.96$. This value z is then used to calculate the confidence interval, calculated as the product of z and SE as seen in Equation 5.2.

$$M \pm z * \sqrt{\frac{1}{n} \times m \times (1-m)} \tag{5.2}$$

5.1.2 Datasets

Following the CRS1k and CRS4k datasets, presented in Chapters 3 and 4, with about 1,000 and 4,000 samples, respectively, we further extended our data to 10,000 WSI (CRS10k dataset). This new set contains 9.26x and 2.36x more slides than CRS1k and CRS4k, respectively. Similarly, the number of tiles is multiplied by a factor of 12.2 and 2.58 (Table 4.1). This volume of slides is translated into an increase in the flexibility to design experiments and infer the robustness of the model, as well as the inclusion of an independent test set. Roughly 9% of the CRS4k dataset (967 WSI) was manually annotated by a team of pathologists, as described in Chapter 3.

The CRS4k dataset was divided into train, validation and test sets, with 8587, 1009 and 900 non-overlapping samples each, respectively. The first includes all the strongly annotated slides and other slides randomly selected. The validation set was also randomly selected. This partition kept the slides of the CRS1k and CRS4k in the same sets, ensuring that slides previously used for validation were not used for training with the new dataset. Finally, the test set was selected from the newly added data.

Domain generalisation was also tested with the same two external datasets used in Chapter 4: TCGA and PAIP datasets.

Table 5.1: Datasets summary, with the number of slides (annotated samples detailed in parenthesis) and tiles, distributed by class.

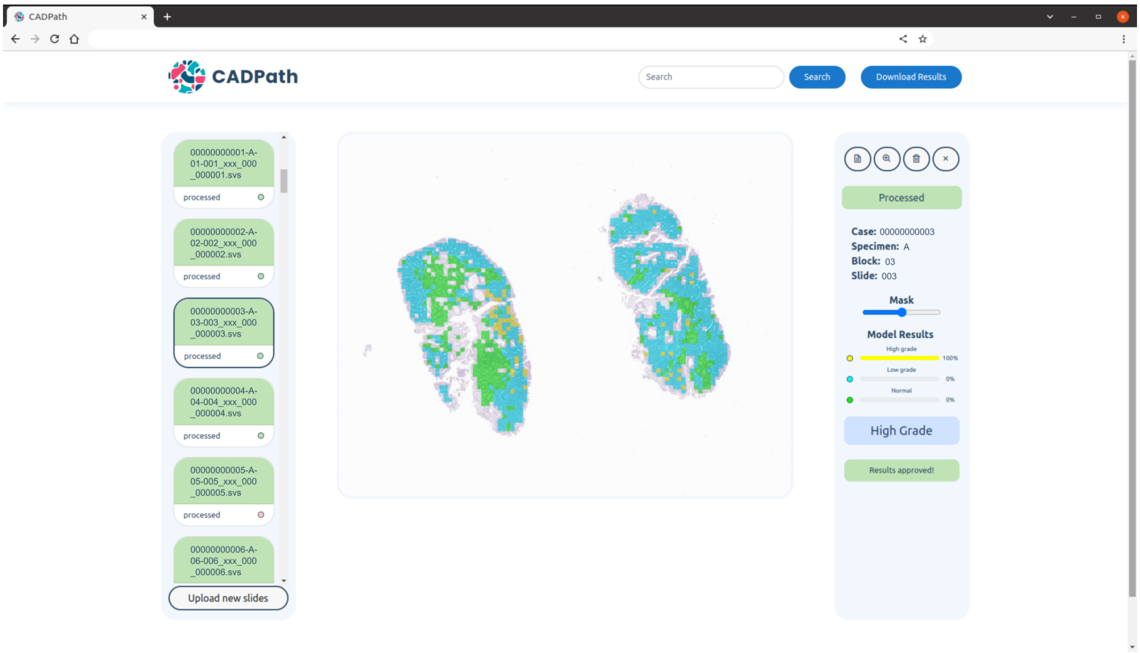
		NNeo	**LG**	**HG**	**Total**
CRS1k dataset	*# slides*	300 (6)	552 (35)	281 (59)	1133 (100)
	# annotated tiles	49,640	77,946	83,649	211,235
	# non-annotated tiles	-	-	-	1,111,361
CRS4k dataset	*# slides*	663 (12)	2394 (207)	1376 (181)	4433 (400)
	# annotated tiles	145,898	196,116	163,603	505,617
	# non-annotated tiles	-	-	-	5,265,362
CRS10k dataset	*# slides*	1740 (12)	5387 (534)	3369 (421)	10,496 (967)
	# annotated tiles	338,979	371,587	341,268	1,051,834
	# tiles	-	-	-	13,571,871
TCGA	*# slides*	1	1	230	232
	# tiles	-	-	-	1,568,584
PAIP	*# slides*	-	-	100	100
	# tiles	-	-	-	97,392

5.1.3 Experimental setup

The backbone network used was ResNet-34 [86], trained on Pytorch with the Adaptive Moment Estimation (Adam) [94] optimiser, a learning rate of 6×10^{-6} and a weight decay of 3×10^{-4}. The training batch size was set to 32, for both fully and weakly supervised training, whereas the test and inference batch size was 256. The model's performance was evaluated on the validation set, and it was used to select the best model based on the accuracy and QWK. The training was accelerated by an Nvidia Tesla V100 (32GB) GPU.

5.2 Prototype

The proposed algorithm was integrated into a fully functional prototype to enable its use and validation in a real clinical workflow. This system was developed as a server-side web application that can be accessed by any pathologist in the lab (Figure 5.2).



00000000001-A-01-001_xxx_000_000001.svs processed

00000000002-A-02-002_xxx_000_000002.svs processed

00000000003-A-03-003_xxx_000_000003.svs processed

00000000004-A-04-004_xxx_000_000004.svs processed

00000000005-A-05-005_xxx_000_000005.svs processed

00000000006-A-06-006_xxx_000_000006.svs

Processed

Case: 00000000003
Specimen: A
Block: 03
Slide: 003

High grade 100%
Low grade 0%
Normal 0%

High Grade

Results approved!

Figure 5.2: Main view of the CAD system prototype: Slide identification, confidence per class, diagnostic, mask overlay controller, results download as csv and slide search are some of the features visible.

The system supports the evaluation of either a single slide or a batch of slides simultaneously and in real-time. It also caches the most recent results, allowing re-evaluation without the need to re-upload slides. In addition to displaying the slide diagnosis, and confidence level for each class, a visual explanation map is also retrieved, to draw the pathologist's attention to key tissue areas within each slide (all seen in Figure 5.2). The opaqueness of the map can be set to different thresholds, allowing the pathologist to control its overlay over the tissue. An example of the zoomed version of a slide with a lower overlay of the map is shown in Figure 5.3.

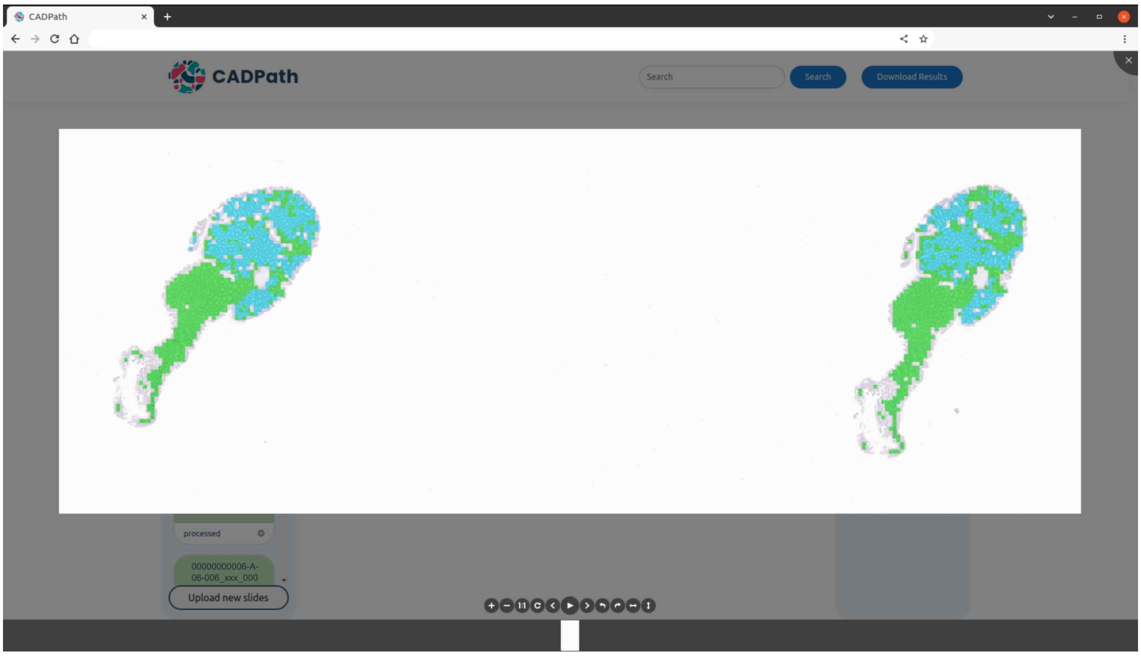
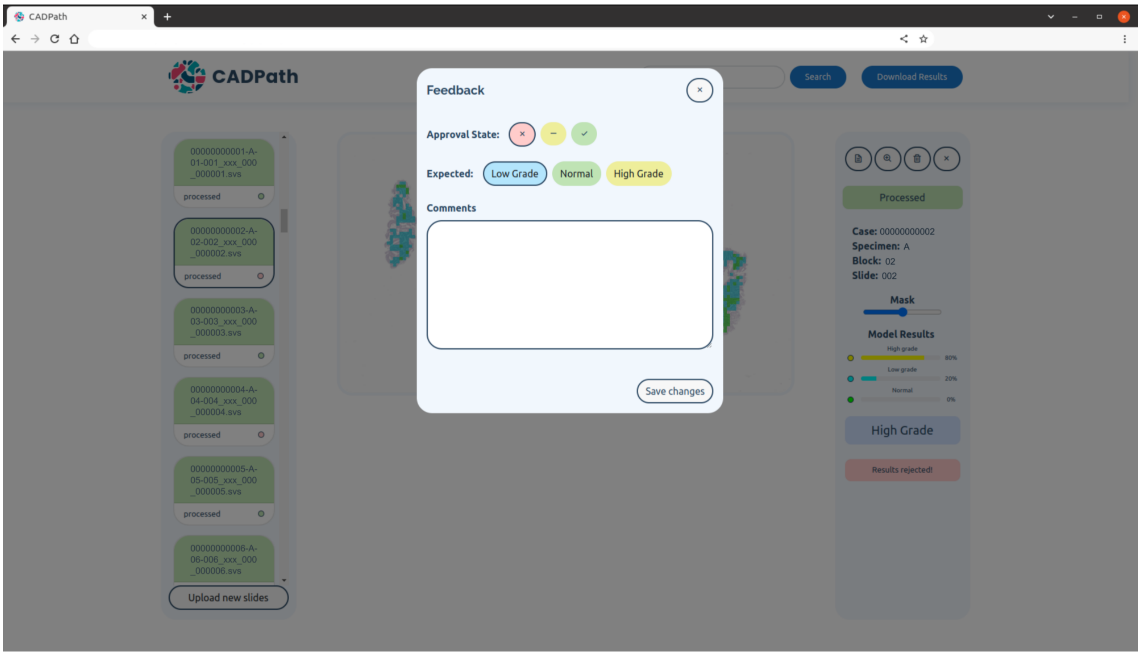


Figure 5.3: Zoomed view of a slide from the CAD system prototype, with the predictions map with a small overlay threshold.

Furthermore, the prototype also allows user feedback where the user can accept/reject a result and provide a justification (Figure 5.4), an important feature for software updates, research development and possible active learning frameworks that can be developed in the future. These results can be downloaded with the corrected labels to allow for further retraining of the model.



Case: 00000000002
Specimen: A
Block: 02
Slide: 002

Figure 5.4: CAD system prototype report tool: the user can report if the result is either correct, wrong or inconclusive and leave a comment for each case individually.

There are several advantages to developing such a system as a server-side web application. First, it does not require any specific installation or dedicated local storage in the user's device. Secondly, it can be accessed at the same time by several pathologists from different locations, allowing for a quick review of a case by multiple pathologists without data transference. Moreover, the lack of local storage of clinical data increases the privacy of patient data, which can only be accessed through a highly encrypted virtual private network (VPN). Finally, all the processing can be moved to an efficient graphics processing unit (GPU), thus reducing the processing time by several orders of magnitude. Similar behaviour on a local machine would require the installation of dedicated GPUs in the pathologists' personal devices. This platform is the first Pathology prototype for colorectal diagnosis developed in Portugal, and, as far as we know, one of the pioneers in the world. Its design was also carefully thought to be aligned with the needs of the pathologists.

The proposed solution was tested in the clinical environment, with 100 extra slides: 28 NNeo, 44 LG and 28 HG. These differ from the CRS10k dataset, in the sense, that they were not selected from the archive. Instead, these cases were actively collected from routine exams, and the prototype was used to support the pathologists in their tasks.

5.3 Results and discussion

5.3.1 Effectiveness of tile sampling

To find the most suitable threshold for sampling the tiles used in the weakly supervised training, we evaluated the percentage of relevant tiles that would be left out of the selection, if the original set was reduced to 75, 100, 150 or 200 tiles, over the first five inference epochs. A tile is considered relevant if it shares the same label as the slide, or if it would take part in the learning process in the weakly-supervised stage. As it is possible to see in Figure 5.5, if we set the maximum number of tiles to 200 after the second loop of inference, we would discard only 3.5% of the potentially informative tiles, in the worst-case scenario. On the other side of the spectrum, a more radical sampling of only 50 tiles would lead to a cut of up to 8%.

Moreover, to assess the impact of this sampling on the model's performance, we also evaluated the accuracy and the QWK with and without sampling the top 200 tiles after the first inference iteration (Table 5.2). This evaluation considered sampling applied only to the training tile set, and to both the training and validation tile sets. As can be noticed, the performance is not degraded and the model is trained in a much faster way. In fact, using the setup previously mentioned, the first epoch of inference, with the full set of tiles takes 28h to be completed, while from the second loop the training time decreases to only 5h per epoch. Without sampling, training the model for 50 epochs would take around 50 days, whereas with sampling it takes around 10.

5.3.2 CRS10k test set

After model inference in the CRS10k test set ($n = 900$), 81 cases had a different prediction from the ground truth. These cases were all blindly reviewed by two pathologists, and discordant cases

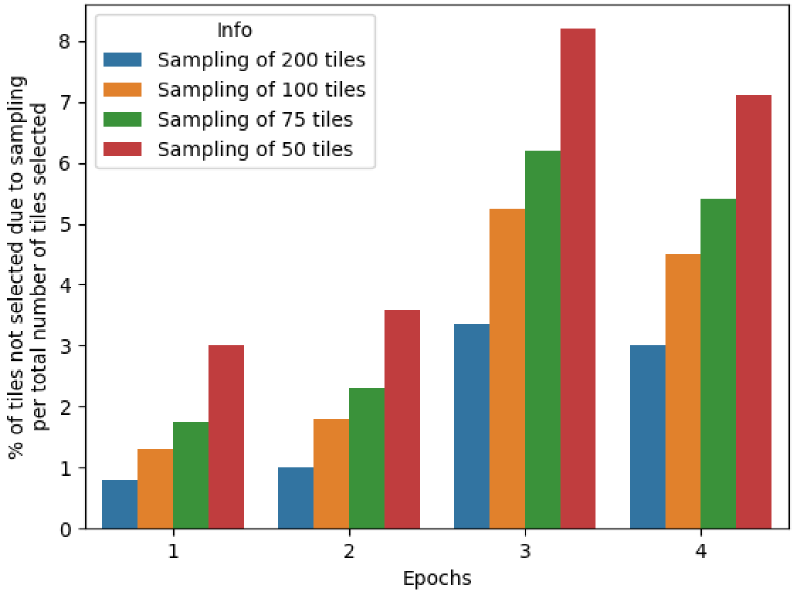


Figure 5.5: Tile sampling impact on information loss: percentage of tiles not selected due to sampling with different thresholds, over the first four inference epochs.

Table 5.2: Model performance comparison with and without tile sampling of the top 200 tiles from the first inference iteration. Compared the best epoch of the initial five and ten epochs.

	Best Accuracy at		Best QWK at	
Sampling	5th epoch	10th epoch	5th epoch	10th epoch
No	84.94% ± 2.20	86.42% ± 2.11	0.809 ± 0.024	**0.829 ± 0.023**
Train	85.43% ± 2.18	86.82% ± 2.08	0.817 ± 0.024	0.828 ± 0.023
Train and Validation	**86.12% ± 2.13**	**86.92% ± 2.08**	**0.824 ± 0.023**	**0.829 ± 0.023**

from the initial ground truth were discussed and classified by both pathologists (and in case of doubt/complexity, a third pathologist was also consulted). Of these, 22 cases were found to be actually correctly classified by the algorithm, and only 59 cases have true misclassification.

Globally, most of the remaining misclassifications were over-calls (n=37), when the model upstaged grading by one category. In 22 cases the algorithm has predicted a lower grade than expected. It is also interesting to note that in no case did the algorithm miss a prediction on a 2-order scale, *i.e.* classifying a normal case as high risk or the opposite. The final test results are presented in Table 5.3, where we also present the comparison with the model trained with fewer data (CRS4k dataset), which helps to conclude that adding more data to training is in fact beneficial for model performance.

Moreover, we also inspected the distribution of the confidence values of the model when predicting the class for each slide of the test set. From Figure 5.6, it is interesting to see that the model is much more confident when predicting right (blue lines), using both the CRS4k and the

Table 5.3: Model performance evaluation on the CRS10k test set.

Training data	**ACC**	**Binary ACC**	**Sensitivity**
CRS4k	93.44% ± 1.61	97.78% ± 0.96	99.60% ± 0.50
CRS10k	89.44% ± 2.01	96.11% ± 1.26	99.72% ± 0.38

CRS10k datasets. Moreover, adding more data increases the robustness of the model, with wrong cases having generally lower confidence in the CRS10k model.

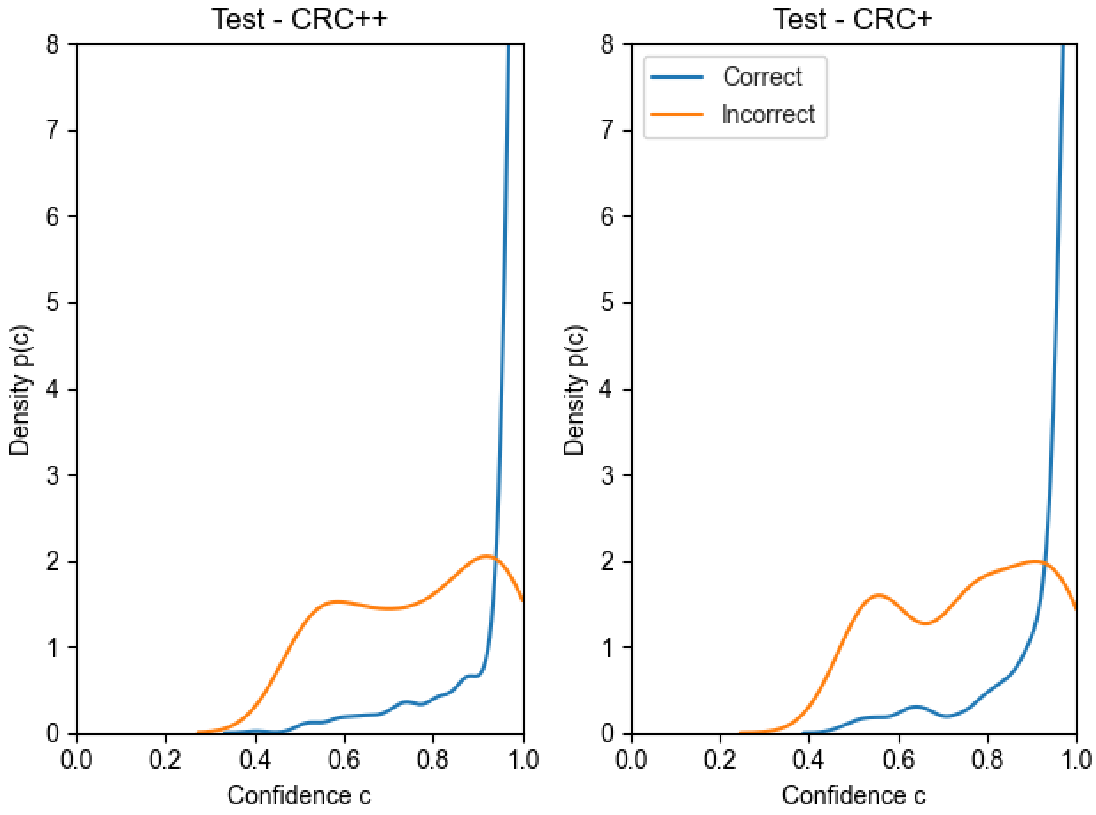


Figure 5.6: Kernel density estimation of the confidences of correct and incorrect predictions performed by the model trained with CRS10k training data (left) and the model trained with the CRS4k training data (right).

5.3.3 Domain generalisation

As in Chapter 4, we also evaluated the domain generalisation capability of the model, using the TCGA (Table 5.4) and PAIP datasets (Table 5.5). Here too one can conclude that more training data leads to an improved model, with the CRS10k version showing much better overall metrics in both external datasets. In fact, particularly on the PAIP dataset, the model trained with fewer data (CRS4k) misled some samples as low-grade dysplasia, which were all corrected in the new version.

Table 5.4: Model performance evaluation on the TCGA test set.

Training data	ACC	Binary ACC	Sensitivity
CRS4k	70.69% ± 5.86	98.71% ± 1.45	0.991 ± 0.012
CRS10k	84.91% ± 4.61	99.13% ± 1.19	0.996 ± 0.008

Table 5.5: Model performance evaluation on the PAIP test set.

Training data	ACC	Binary ACC	Sensitivity
CRS4k	69.00% ± 9.06	100.00% ± 0.00	1.000 ± 0.000
CRS10k	100.00% ± 0.00	100.00% ± 0.00	1.000 ± 0.000

5.3.4 Prototype usability

The first test of the prototype in the "real world" scenario revealed an accuracy of 89.00% ± 6.13, a binary accuracy: 98.00% ± 2.74 (by joining both low and high-grade cases) and sensitivity of 98.60 ± 0.26, in line with the results obtained in the previous test sets.

As it is currently designed, the prototype works preferentially as a "second opinion", allowing to assess difficult and troublesome cases, without the immediate need for the intervention of a second pathologist. Due to its “user-friendly” nature and very practical interface, the cases can be easily introduced into the system and results are rapidly shown and easily accessed. Also, by not only providing results but presenting visualisation maps (corresponding to each diagnostic class), the pathologists can compare their own remarks to those of the algorithm, towards a future “AI-assisted diagnosis”.

Another relevant aspect is that the prototype allows for user feedback (agreeing or not with the model’s proposed result), which can be further integrated into further software updates. Also interesting, is the possibility of using the prototype as a triage system on a pathologist´s daily workflow (by running upfront, before the pathologist checks the cases). By signalling the cases that would need to be more urgently observed (namely high-risk lesions), it would allow the pathologists to prioritise their workflow. Further, by providing a previous assessment of the cases, it would also contribute to enhancing the pathologists’ efficiency. Although it is possible to use the model as it is upfront, it would still classify many samples incorrectly (since it was not trained on the full spectrum of colorectal pathology). Thus, we believe it should also have (at least) a "others" category, to better serve as a triage system.

Presently, there is no recommendation for dual independent diagnosis of colorectal biopsies (contrary to gastric biopsies, where, in cases in which surgical treatment is considered, it is recommended to obtain a pre-treatment diagnostic second opinion [27]). However, if in the future this also becomes a requirement, a tool such as this prototype could assist in this task. In fact, such double revision guidelines, together with the worldwide shortage of pathologists, make the need for CAD systems even more relevant for better patient care. Lastly, we also anticipate that this prototype, and similar tools, can be used in a teaching environment since its easy use and interpretable capability (through the visualisation maps) allows for easy understanding of the given classifications. Moreover, the web-based interface allows for easy use and sharing.

5.4 Summary

In this chapter we propose a novel dataset, with more than 10,000 colorectal WSI, a sampling approach, to reduce the difficulty of using large datasets, and a prototype that provides a simple integration in clinical practice and visual explanations of the model's predictions. Furthermore, the mistakes committed by the model were blindly revised by the pathologists, which enabled the correction of some mislabelling errors and a deeper understanding of the model's strengths and weaknesses.

The proposed prototype was evaluated in the daily routine of the pathology lab, from which we conclude that it really aids in the analysis of CRC slides, by detecting high-grade lesions in colorectal biopsies with high sensitivity. Moreover, by visually displaying tissue classification, it helps to focus the pathologist's attention on key areas. Thus, it can be used as a second opinion and even as a flag for details that may have been missed at first glance.

Part III

Cervical cancer

Chapter 6

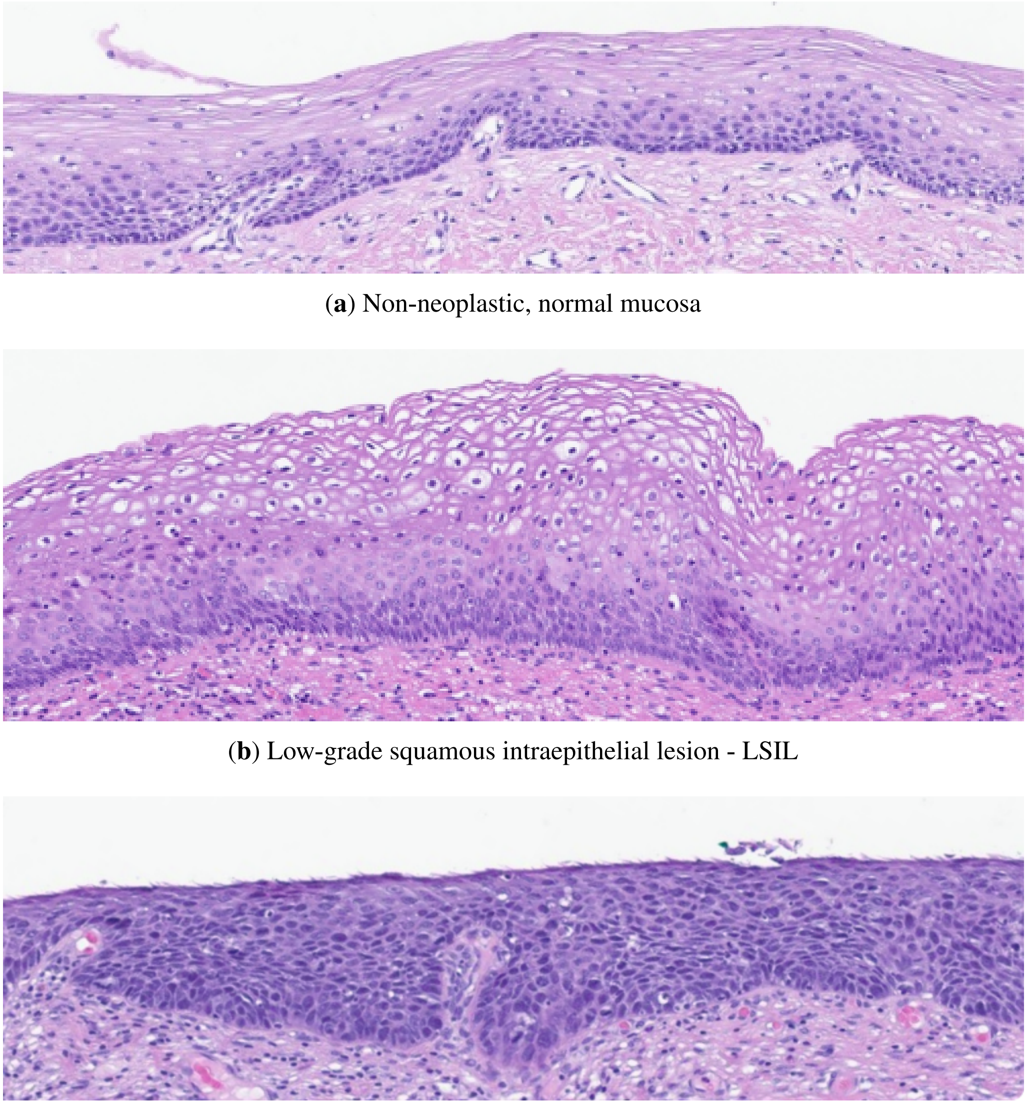
Cervical cancer insights

6.1 Epidemiology

Cervical cancer (CCa) is the fourth most prevalent cancer and also the fourth cause of cancer-related deaths in women, worldwide [95]. Around 80-90% of CCa are squamous cell carcinomas (SCC) and the large majority of these are caused by human papillomavirus (HPV) infection, although HPV-indepen- dent forms of cancer also exist [96]. Fortunately, CCa is one of the most successfully preventable and treatable forms of cancer, as long as it is early detected and effectively managed [97]. As such, screening pre-cancerous lesions and vaccination are key to preventing the disease. Squamous intraepithelial lesions (SILs) of the uterine cervix, or cervical intraepithelial neoplasia (CIN), are pre-malignant HPV-driven proliferations of the squamous epithelium, showing viral cytopathic changes and/or maturation alterations that do not extend beyond the basement membrane [96].

6.2 Cervical dysplasia grading

Grading cervical dysplasia is currently based on a two-tiered system (LSIL/HSIL), as it shows enhanced reproducibility between pathologists and higher biologic significance when compared with the previously used system (CIN 1/2/3) [96]. Grading is mostly based on the proportion and location of immature cells within the squamous epithelium and on the cytopathic changes caused by HPV. According to WHO classification of tumours, low-grade squamous intraepithelial lesions (LSIL) are characterised by the proliferation of basal/parabasal-like (immature looking) cells within the lower third of the epithelium (Figure 6.1(b)), along with the so-called koilocytic atypia (*i.e.* recognisable nuclear and cytoplasmic changes caused by HPV, namely characteristic perinuclear vacuolisation and nuclear enlargement, irregularity and hyperchromasia) [96]. Mitotic activity can be observed, but atypical mitoses should be absent or rare. In high-grade squamous intraepithelial lesions (HSIL), on the other hand, the proliferation of basal/parabasal-like cells extends to the middle and upper third of the epithelium (Figure 6.1(c)). Nuclear abnormalities are seen throughout the thickness of the epithelium as well as mitotic activity (including atypical mitosis) [96]. The term HSIL encompasses the formerly known CIN2 and CIN3 lesions, as well as carcinoma *in situ* lesions. It is known that similarly to other grading tasks in pathology, grading cervical dysplasia is hurdled by significant inter- and intra-observer variability [98]. Nonetheless, distinguishing normal mucosa, LSIL and HSIL has major clinical implications and remains an important task in gynaecological pathology. The main interpretative problems in this context are the distinction of normal mucosa (Figure 6.1(a)), which can show nonspecific inflammatory/reactive changes, from dysplasia (LSIL) and also to distinguish irrelevant, mostly transient, dysplasia (LSIL) from pre-cancer (HSIL) [99]. Using ancillary markers, namely p16 staining, can help in some cases, but its overuse should be avoided, since it might cause over-diagnosis of high-grade lesions, and it shouldn't exceed 20-40% (or even less) of cervical biopsies [96].



(**a**) Non-neoplastic, normal mucosa

(**b**) Low-grade squamous intraepithelial lesion - LSIL

(**c**) High-grade squamous intraepithelial lesion - HSIL

Figure 6.1: Cervical squamous epithelium dysplastic progression: (a-c) from normal mucosa to HSIL. Examples from CADpath dataset, sampled at 10X magnification.

6.3 Computational Pathology in CCa

Computational pathology studies applied to cervical cancer have been mostly focused on cytology [100–102], since the initial screening of lesions is performed using cytology specimens (smears or liquid-based preparations), either by HPV molecular testing or co-testing HPV in combination with cytology evaluation [103–105]. Nonetheless, it is the histologic assessment of cervical biopsies, loop electrosurgical excision procedure (LEEP) samples and surgical specimens that constitutes the gold standard for the diagnosis of cervical lesions. As such, cervical histopathology image analysis is also an active research field [106]. Noteworthy, in the last five years, most classification studies using deep-learning approaches in cervical cancer have only used cropped images as opposed to full WSI [107–115].

6.3.1 Automatic cervical dysplasia grading

In 2019, Li *et al.* [107], proposed a transfer learning framework of Inception-V3 network to classify well, moderately and poorly differentiated cervical carcinomas. They used 307 IHC-stained images and reported a 77.3% accuracy. Further, also in 2019, they presented an improved method for the same task, with 88% global accuracy. In this study [108], a weakly supervised framework based on multilayer-hidden conditional random fields was developed on a dataset of more than 100 IHC-stained images. Also in 2019, a study by Xue *et al.* [109] analysed the 4-class (normal, CIN 1 to 3) classification problem, using conditional generative adversarial networks (cGANs) to expand the training dataset, by synthesising realistic cervical images. They report using 1,112 normal, 181 CIN1, 463 CIN2 and 454 CIN3 patches of 256×128 pixels, although not specifying from how many patients/samples the patches derive. Their results show a significant improvement in classification accuracy from 66.3% to 71.7%, using the same ResNet-18 baseline classifier, after leveraging with the cGAN-generated images. More recently, the same group proposed also a synthetic augmentation framework that selectively adds new synthetic image patches, generated by their GAN model (HistoGAN), rather than expanding directly the training set with synthetic images [112]. This experiment, using a similar cervical dataset, and an additional dataset of metastasised lymph node images, resulted in significant and consistent improvement of the classification performance: 6.7% and 2.8% higher accuracy, compared with their previous work, for cervical histopathology and metastatic cancer datasets, respectively.

In 2020, Huang *et al.* [111] proposed a classification method based on the least absolute shrinkage and selection operator (LASSO) and ensemble learning support vector machine (EL-SVM). Images from 468 cervical biopsies were used, and an 86.84% average accuracy was shown, classifying normal, LSIL and HSIL lesions and carcinoma. Alternately, Sornapudi *et al.* [110] have built a network pipeline (DeepCIN), with two classifier networks, to analyse high-resolution images ($n = 453$, manually extracted from 146 WSIs), with an accuracy of 88.4% (normal, CIN 1 to 3).

In 2021, Albayrak *et al.* [113] reported a classification accuracy of 65.4% also grading cervical precursor lesions (normal, CIN 1 to 3), using a morphological-based feature extraction method. Their dataset consists of 128 images from 54 patients.

In 2022, Cho *et al.* [114], aimed to develop and validate deep learning (DL) models to classify cervical intraepithelial neoplasia (normal, CIN1 to 3) automatically. The models were developed on a dataset comprising 1106 images from 588 patients, and the mean accuracies for the four-class classification were 88.5% by DenseNet-161 and 89.5% by EfficientNet-B7, which were similar to the reported performance of two pathologists (93.2% and 89.7%). Further, they also calculated the performance for a three-class classification (correspondent to normal, LSIL, and joining CIN2 and CIN3 as HSIL), and the mean accuracies of DenseNet-161 and EfficientNet-B7 increased to 91.4% and 92.6%, respectively (whereas pathologists' performances were 95.7% and 92.3%). Lastly, Habtemariam *et al.* [115] have also proposed a cervical cancer classification system using DL techniques. In this study, the authors resorted to 915 histopathology images (and also included 4005 colposcopy images). They reported a test accuracy of 94.5% for cervical cancer classification into normal, pre-cancer, squamous cell carcinoma and adenocarcinoma, using the Efficientnet-B0 model. Regarding the colposcopy images, the model achieved an accuracy of 96.84% for cervix-type classification.

To the best of our knowledge, regarding classification tasks on cervical pathology, only the work of Sornapudi *et al.* [116], in 2021, was developed directly on WSI. Working on such images has the increased difficulty of high dimensionality and resolution, which can not be easily fitted in graphics process units (GPU), usually used to train DL models. However, it is the most significant as it is the only way for models to be effectively translated to clinical practice. They have developed a novel image analysis toolbox to automate CIN diagnosis of cervical biopsies, with an 85% exact-class accuracy. This work shows the potential of the used methodology but it was trained with only 150 WSI and using the CIN grading system, which is not usually used nowadays.

6.4 Summary

Cervical cancer is the fourth most common female cancer worldwide and the fourth leading cause of cancer-related death in women. Nonetheless, it is also among the most successfully preventable and treatable types of cancer, provided it is early identified and properly managed. As such, the detection of pre-cancerous lesions is crucial. These lesions are detected in the squamous epithelium of the uterine cervix and are graded as LSIL and HSIL, respectively. Being located in a usually small portion of the tissue sample, these malignant areas can sometimes be not so evident and, because of their complex nature, are subjective to grade. Therefore, developing machine learning models, particularly directly on WSI, can assist pathologists in this task.

Most of the work already developed for the automatic assessment of cervical dysplasia focuses on a CIN classification system, prior to the currently recommended system, which is not as reproducible among pathologists and has lower biological significance. On the other hand, only one author presents a methodology to use WSI and not just crops/images of manually identified areas of epithelium. Despite the potential reported in the literature, these works cannot be directly applied in clinical practice, and more effort is needed in their development into a diagnostic tool.

Chapter 7

Weakly-supervised learning for cervical cancer grading

Author Contributions

The research work described in this chapter was conducted in collaboration with Ana Moreira, Pedro C. Neto and the IMP Diagnostics team, under the clinical supervision of Isabel M. Pinto, and the technical supervision of Jaime S. Cardoso. The author of this thesis contributed to this work on problem conceptualisation, data curation, the preparation and conduction of experiments (namely, the classification model and the final framework), the results discussion, and publication writing. This chapter was adapted from the article originally submitted to the *Scientific Reports* journal, as:

- S.P. Oliveira[a], D. Montezuma[*], A. Moreira[*], D. Oliveira, P.C. Neto, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, I.M. Pinto and J.S. Cardoso. A CAD system for automatic dysplasia grading on H&E cervical whole-slide images. *Scientific reports*, 2022 [submitted, waiting for decision]

[*]Shared co-first authorship

[a]Shared co-first authorship

The World Health Assembly adopted a global strategy for cervical cancer elimination in August 2020, which is set on 3 pillars: vaccination, screening and treatment [117, 118]. Each country should meet the 90-70-90 targets, by 2030: 90% of girls fully vaccinated against HPV by the age of 15; 70% of women screened using a high-performance test by the age of 35, and again by the age of 45; 90% of women with pre-cancer treated and 90% of women with invasive cancer managed [117, 118]. Thus, as stated by Rodriguez *et al.* [119], technological innovation is needed to achieve cervical cancer screening and management targets, as well as the strategic implementation of automatic diagnosis tools in clinical practice. To this end, we propose an approach to grade dysplasia directly from cervical whole-slide images (WSI), that first identifies squamous epithelium tissue and then classifies it, helping pathologists assess these cases.

7.1 Methodology

7.1.1 Problem definition

As seen before, automatic cervical dysplasia grading should be done on H&E WSI and focused on the cells of the squamous epithelium, that on LEEP samples and surgical specimens is usually a thin area within the sample tissue. Thus, despite the usual big dimension of digitised slides, which easily become hard and tedious to annotate, fully-supervised models for cervical dysplasia may require annotation and labelling of epithelium, besides the usual annotation of relevant areas for diagnosis/classification. However, in general, WSI are not usually publicly available, and, when they are, they have only the associated diagnosis and no detailed pixel-level annotations.

In this sense, and following a common approach in computational pathology, we propose a weakly-supervised methodology for cervical dysplasia grading, based on tiles (small areas within the slide) using different levels of training supervision, in an attempt to leverage a big dataset only partially annotated with full details. In this particular case, we define three different levels of data for training, as represented in Figure 7.1:

1. Labelled slides (LS): each slide labelled only with the slide diagnosis can be abstracted as a set (or bag) of tiles from the epithelial regions (tiles from non-epithelial regions are not used in the development of the DL model); The labelling of each tile is unknown, but it is assumed that the worst of the unknown tile labels corresponds to the bag labelling (slide diagnosis); Epithelial areas are automatically identified using a (deep) segmentation model;

2. Annotated epithelium (AE): epithelium was delineated and labelled on a subset of the slides; For model development, each labelled epithelial region is considered a set/bag of tiles where the labelling of the set is known but the labelling of the individual tiles is not; The slides with

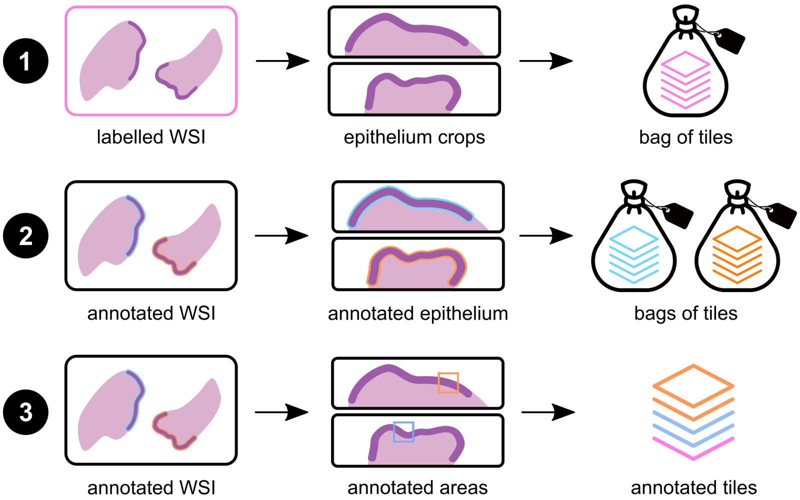


Figure 7.1: Levels of annotation used for model training: labelled slides (top), annotated epithelium (middle) and annotated tiles (bottom).

labelled *epithelia* yield smaller bags than the slides with diagnosis only, which improves the quality of training;

3. Annotated tiles (AT): within the annotated epithelium, smaller regions of interest were delineated, indicating unequivocal tissue areas of non-neoplastic tissue, LSIL and HSIL, from where tiles with known labels were retrieved.

Thus, besides alleviating the task of annotation, such a scheme can serve two other purposes: to gather details to train a supervised segmentation model, focusing the evaluation of dysplasia degree on the regions of interest (ROIs), and add more information to the tiles bags, to facilitate the classifier learning.

7.1.2 Learning framework

The proposed framework for cervical dysplasia grading on H&E digitised slides (Figure 7.2) consists of two main parts: an epithelium segmentation model, to locate the region of interest (ROI) to focus on, and a tile classifier, to distinguish normal epithelium, low and high-grade lesions.

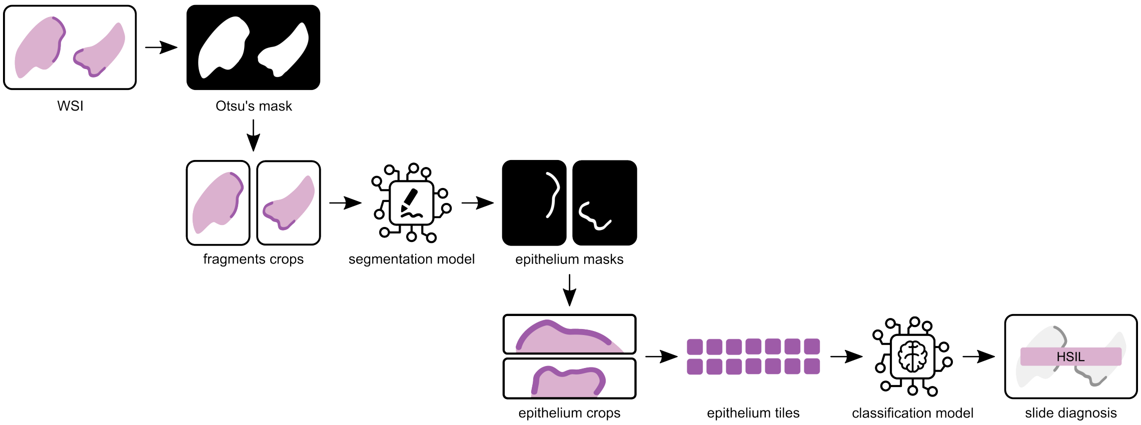
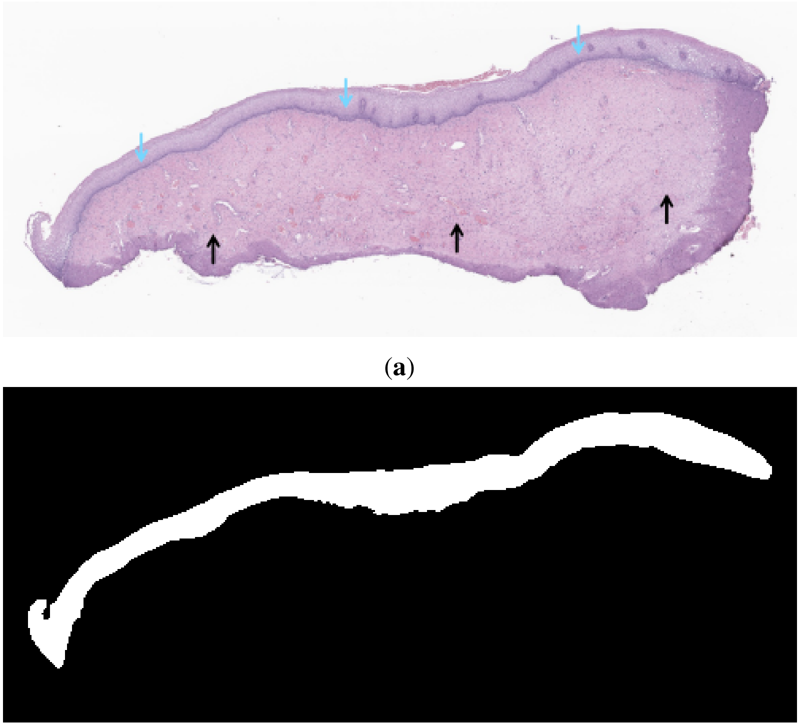


Figure 7.2: Proposed framework for cervical dysplasia grading on H&E digitised slides, with an epithelium segmentation model and a tiles classification model.

7.1.2.1 Epithelium segmentation

As mentioned before, relevant features of cervical dysplasia are located in the epithelium, a small portion of the tissue area (marked with blue arrows in the example of Figure 7.3(a)). In this way, assessing the whole tissue area would not only be more time-consuming but would also introduce a lot of noise into the learning process, as the features of the submucosa areas (beneath the epithelium band, as the black arrows in Figure 7.3(a)) are not directly related to the degree of dysplasia.

To this end, we propose the U-NET architecture [120], a standard biomedical image segmentation model, trained with fragment crops of the slide, resized to 1024 × 1024 pixels, to account for the huge variability of slide/tissue/epithelium dimensions between slides. In fact, the amount of



(**a**)

(**b**)

Figure 7.3: Fragment crop (a), with the epithelium annotation mask (b).

tissue in each slide may vary a lot, as well as the ratio of epithelium/submucosa tissue. Moreover, WSIs have huge dimensions (commonly several hundreds of pixels in both width and height) and can not be used in their original size. By using each tissue fragment individually, we can reduce the loss of information of image resize and also the sparsity of images to be segmented. Each fragment is cropped from the slide using an Otsu's thresholding [85] mask as a reference to locate tissue and define its limits (Figure 7.2). The end goal is to get a mask, such as the annotation one (example in Figure 7.3(b)), to guide the image pre-processing step for the classification model.

For learning the segmentation model, we used the BCE-Dice loss (L_s), a standard function used on image segmentation, based on the combination of the binary cross-entropy (BCE) and the Dice loss, defined as:

$$L_s = BCE + Dice\ Loss \tag{7.1}$$

where the BCE term is defined as,

$$BCE\ (y,\hat{y}) = -(y \times log(\hat{y}) + (1-y) \times log(1-\hat{y})) \tag{7.2}$$

and the Dice loss term is defined as,

$$Dice\ Loss\ (y,\hat{y}) = 1 - \frac{2y\hat{y}+\lambda}{y+\hat{y}+\lambda} \tag{7.3}$$

with $y \in \{0,1\}$ being the target value at each individual pixel, $\hat{y} \in [0,1]$ the predicted value retrieved by the model and λ a smoothing factor (set to 1).

7.1.2.2 Tissue classification

Following the segmentation step, with the output mask (or epithelium annotations in case of annotated samples), we crop smaller regions containing the identified ROIs. However, these images are still too big to be fitted within a GPU and thus, need to be decomposed, to avoid the loss of key tissue details if the image is resized. Moreover, ideally, the smaller areas to be extracted should include the entire epithelium thickness, so the model can learn the proliferation of basal/parabasal-like cells, as described in Section 6.2. Hence, we use the centre line of the epithelium mask to sample tiles of 512 × 512 pixels, from the WSI level corresponding to 20X, along the extension of the epithelium (Figure 7.4).

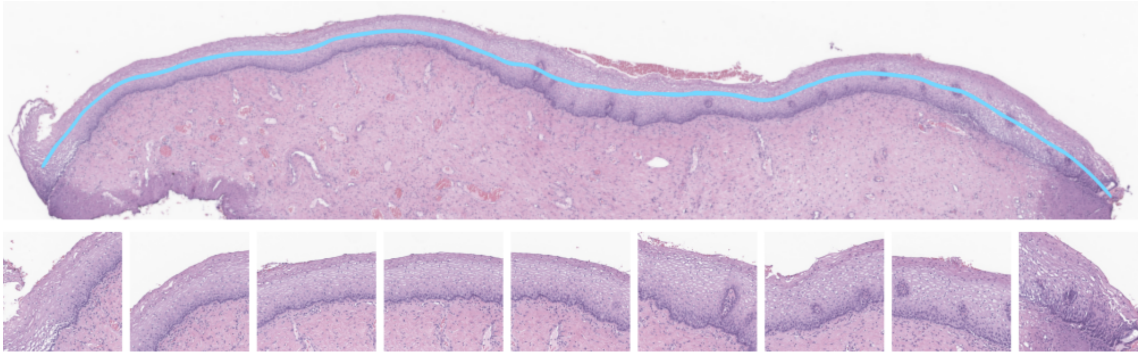


Figure 7.4: Example of an epithelium crop (top) with some tiles of $512x512$ pixels (bottom), sampled along the centre line (in blue) of the epithelium area.

Following a common approach in computational pathology, we propose a semi-supervised learning scheme, based on multiple instance learning (MIL), using instance bags that can include the tiles of all epithelium areas within a slide, labelled with the slide diagnosis, or tiles of each epithelium area, in the case of annotated samples. Also, taking advantage of some areas specifically annotated for each class, we include some individually labelled tiles, for extra supervision.

With such a setup, to train the classification model, we assume that each bag is represented by its worst tile (the tile with poorer diagnosis), following a generalisation of the MIL assumption. If there is one HSIL tile in the set, then the bag (*i.e.* the epithelium area or the slide) is diagnosed accordingly. On the other hand, if no tile has dysplasia, the bag is classified as non-neoplastic. It is worth mentioning that for the classification model we only use three classes (NNeo, LSIL and HSIL) since the "others" categories correspond to slides without epithelium areas. To find the most representative tile, the model is firstly used to infer the class of all tiles in the set, which are then ranked. Tile ranking is based on the expected value of the predictions, following the approach of [121, 122]. For a tile $\mathscr{T}_{s,t}$, where s is the index of the slide and $n \in \{1,\cdots,n_s\}$ is the tile number within the slide, the expected value of the score is defined as:

$$\mathbb{E}(\hat{C}_{s,t}) = \sum_{i=1}^{n} i \times p\left(\hat{C}_{s,t} = C^{(i)}\right) \tag{7.4}$$

where $\hat{C}_{s,t}$ is a random variable on the set of possible class labels $\{C^{(1)}, \cdots, C^{(n)}\}$ and $p\left(\hat{C}_{s,t} = C^{(i)}\right)$ are the n output values of the neural network.

For this task, we used the weighted κ loss function (L_c), based on the quadratic weighted kappa (QWK), defined as:

$$L_c = 1 - \kappa, \quad with\ \kappa = 1 - \frac{\sum_{y,\hat{y}=1}^{n} w_{y,\hat{y}}\, x_{y,\hat{y}}}{\sum_{y,\hat{y}=1}^{n} w_{y,\hat{y}}\, m_{y,\hat{y}}} \tag{7.5}$$

where n is the number of classes, $w_{y\hat{y}}$ belongs to the weight matrix, $x_{y\hat{y}}$ belongs to the observed matrix and $m_{y\hat{y}}$ are elements in the expected matrix. The $n \times n$ matrix of weights w is computed based on the difference between the actual (y) and the predicted ($\hat{y}$) class, as follows:

$$w_{y,\hat{y}} = \frac{(y - \hat{y})^2}{(n-1)^2} \tag{7.6}$$

In this way, class ordinality is taken into account during model training by weighting misclassifications: model predictions closer to the target label are less penalised than more distant ones. In fact, the classification targets are defined upon dysplasia grading, meaning that misclassifying an HSIL as NNeo is worse than identifying it as LSIL.

7.1.3 Dataset

For this work, we gather an in-house dataset that contains 2000 WSIs of LEEP samples and surgical specimens (cervical biopsies were not included). Since the main goal is to grade cervical dysplasia, labelling follows the two-tiered diagnostic system, thus dividing the slides into four categories: non-neoplastic (NNeo), LSIL, HSIL or non-representative ("others").

The slides were assessed and labelled by one of two pathologists, and the diagnosis was compared with the original report (which served as a second grader). If both were coincident, no further assessment was performed. In case of difference, or case complexity, the case was rechecked and decided by a third pathologist. Further, a subset of slides ($n = 186$, approximately 10% of the complete dataset) was also manually annotated (like the example in Figure 7.5), using the Sedeen Viewer software [79], delineating epithelium areas and characteristic areas correspondent to the different classification categories. Table 7.1 summarises the class distribution of annotated and non- annotated data, including the number of fragment crops, epithelium crops (obtained from annotations or from the segmentation model for the non-annotated set), and the tiles obtained after the pre-processing described in Section 7.1.2.2.

All cases were retrieved from the data archive of the IMP Diagnostics laboratory, Portugal, and were digitised with 2 Leica GT450 WSI scanners, at 40X magnification (pixel size of $0.26\mu m^2$).

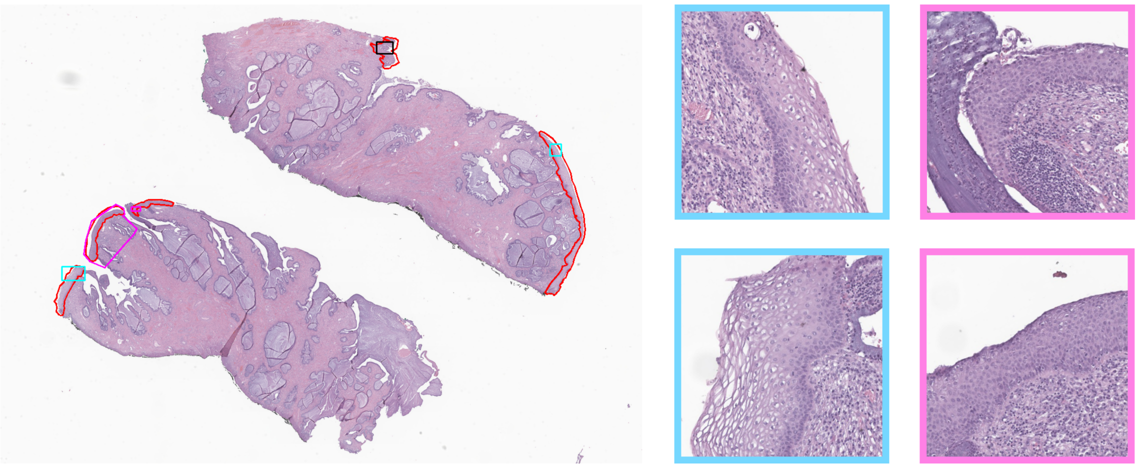


Figure 7.5: Data annotation example: (red) epithelium areas, (blue) LSIL and respective tiles below, (pink) HSIL and respective tiles below, (black) uncertain label areas not used for training.

Table 7.1: Dataset summary: number of samples per class, with the annotated ones indicated in parenthesis. *Fragment crops are divided into positive (that include NNeo, LSIL and HSIL classes) and negative samples ("others" class), if they contain or do not contain epithelium areas, respectively.

Classes	**Slides**	**Fragment crops**	**Epithelium areas**	**Tiles**
NNeo	702 (34)	3,496 (224)*	2,082 (52)	35,038 (1,413)
LSIL	885 (67)		5,620 (240)	58,419 (4,868)
HSIL	323 (61)		1,885 (91)	17,154 (1,087)
Others	90 (24)	184 (88)*	–	–
Total	2,000 (186)	3,680 (312)	9,587 (383)	110,611 (7,368)

Data collection and usage were performed following national legal and ethical standards applicable to this type of data. Since the study is retrospectively designed, and no protected health information was used, patient informed consent is exempted from being requested. Diagnostics were made using a medical grade monitor LG 27HJ712C-W and the Aperio eSlide Manager software.

To test the entire framework in an independent set of slides, the dataset was divided into a training and a test set, containing 1400 (70%) and 600 (30%) samples, respectively. All the annotated slides were kept to train the models and, from the non-annotated set, all the "others" cases, that don't have epithelium areas to be segmented, were used for testing.

7.1.4 Training details

The UNet model was randomly initialised and trained using the Adaptive Moment Estimation (Adam) [123] optimiser (learning rate of 1×10^{-4}), during 250 epochs, with mini-batches of 4 images, resised for 1024×1024. The classification model was initialised with the ImageNet weights and trained with the Adam optimiser, a learning rate of 1×10^{-5} and a batch size of 16, for 300 epochs. The models' performance was evaluated at the end of each epoch to select the best model based on the validation loss and the validation accuracy for the segmentation and classification tasks, respectively. All experiments were conducted using Pytorch and on a single Nvidia Tesla V100 (32 GB) GPU.

7.2 Results & discussion

7.2.1 Segmentation model

Considering the lack of literature methods that use the entire slide and the same grading system, to perform a benchmark, we performed several ablation studies to confirm the capabilities of the proposed methodology. In this section, the results of experiments conducted for the segmentation and the classification models are presented individually, as well as the results of the complete framework.

To train the segmentation model we used all of the annotated slides (186), from which we cropped 312 tissue fragments, divided into training and validation sets, in a $\approx 70/30$ ratio respectively, taking into account class (with or without epithelium) and patient stratification (Table 7.2). All experiments were evaluated in 88 fragments crops, based on 5 metrics at the pixel level: Dice score, intersection over union (IoU), sensitivity, precision and accuracy (Table 7.3).

Table 7.2: Class distribution of the fragments crops used to train the segmentation model.

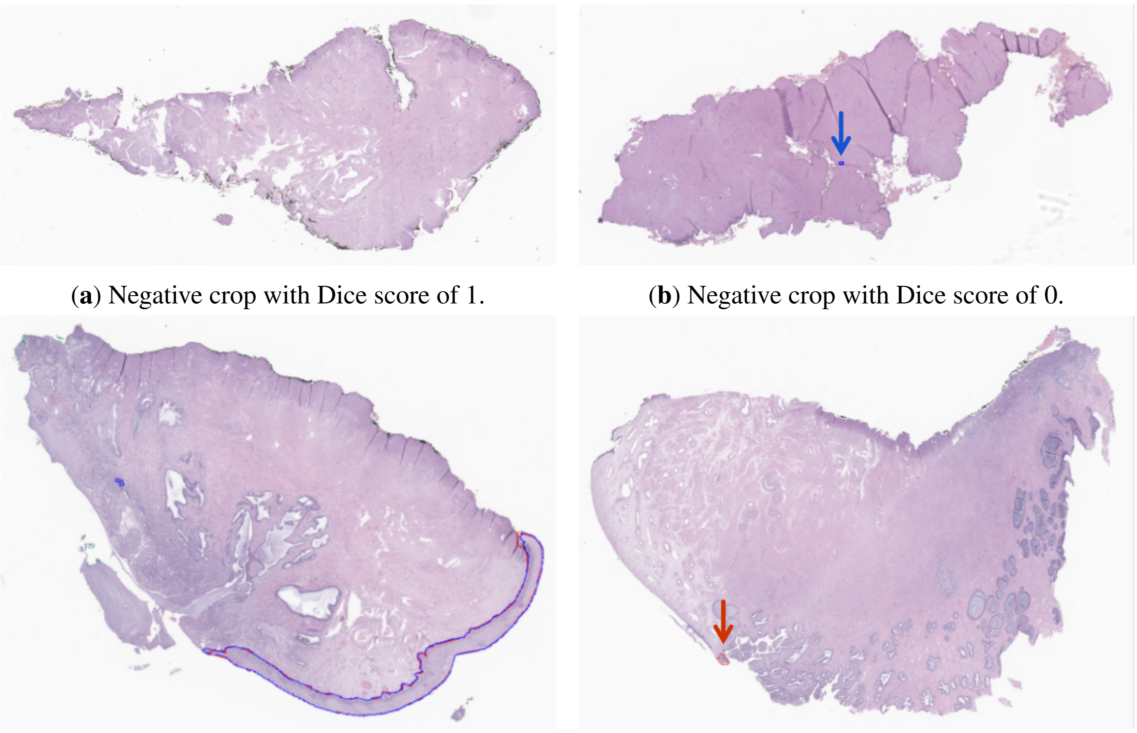
Classes	**Training Set**	**Validation Set**	**Total**
Positives	183 (81.70%)	72 (81.82%)	255 (81.73%)
Negatives	41 (18.30%)	16 (18.18%)	57 (18.27%)
Total	224	88	312

Table 7.3: Performance of the UNet model for epithelium segmentation, trained with different input images (3-channel *vs* 1-channel) and different loss functions (BCE vs. BCE-Dice Loss).

Model version	**Loss function**	**Dice score**	**IOU**	**Sensitivity**	**Precision**
UNet w/ RGB channels	BCE	**74.46%**	**60.67%**	**75.90%**	**74.41%**
UNet w/ grayscale		70.54%	55.63%	73.93%	68.13%
UNet w/ saturation channel		53.24%	37.47%	54.69%	55.24%
UNet w/ RGB channels	BCE-Dice	**80.64%**	**68.17%**	82.80%	**79.68%**
UNet w/ grayscale		77.81%	64.44%	**83.27%**	74.73%
UNet w/ saturation channel		66.86%	51.75%	72.49%	63.45%

The first experiment used the images' RGB channels as input, which achieved a Dice score of 80.64%. Next, in an attempt to understand if the difference in the colour of the epithelium could be confusing the model, since some epithelium areas are darker than others, we trained the model with grayscale images, and also using the saturation channel of the HSV colour space. However, as reported in Table 7.3, the models trained with only one channel did not perform better, so we can conclude that colour information is relevant for the correct distinction between epithelium and submucosa regions, even though the variability mentioned above. Thus, the RGB version was selected as the segmentation model to use for the complete framework. Additionally, we also trained the segmentation model with the standard pixel-wise binary cross-entropy loss (BCE), which showed to be less adequate for the task at hand with any type of input (Table 7.3).

When analysing the predicted masks, as some examples in Figure 7.6, it is possible to see that, in most cases, the errors should not significantly affect the next stage of classification, since most errors are a few extra or missing pixels at the edges of the epithelium. Only very rarely does the model fail to recognise a large part of the epithelium or misidentify a significant area.



(**a**) Negative crop with Dice score of 1.

(**b**) Negative crop with Dice score of 0.

(**c**) Positive crop with Dice score of 0.97.

(**d**) Positive crop with Dice score of 0.

Figure 7.6: Examples of outputs of the segmentation model (in blue) in comparison with the ground truth from annotations (in red).

7.2.2 Classification model

To train the classification model we used 383 annotated epithelial regions, divided into training and validation sets, also taking into account patient and class stratification, as for the segmentation task (Table 7.4), resulting in 111 (29%) examples to validate and choose the best classification model. Since we only want to classify epithelium areas, the data classes are NNeo, LSIL and HSIL.

Table 7.4: Class distribution of the annotated epithelium areas used to train the classification model.

Classes	**Training Set**	**Validation Set**	**Total**
NNeo	37 (13.60%)	15 (13.51%)	52 (13.58%)
LSIL	177 (65.08%)	63 (56.76%)	240 (62.66%)
HSIL	58 (21.32%)	33 (29.73%)	91 (23.76%)
Total	272	111	383

For this task experiments, we started by testing two loss functions (the standard cross-entropy, and an ordinal one, the weighted κ) and three versions of the ResNet network (with 18, 34 and 50 layers), evaluating the accuracy, quadratic weighted kappa (QWK), sensitivity, precision, F1-score and the mean AUC (Table 7.5). The best performing model was the ResNet-34, trained with the weighted κ loss function, which achieved an accuracy of 69.64% and a sensitivity of 72.97%.

Table 7.5: Performance of the classification model: architectures and loss functions comparison.

Model	**Loss function**	**Accuracy**	**QWK**	**Sensitivity**	**Precision**	**F1-score**	**AUC**
ResNet-18	CE	67.47%	**0.56**	68.47%	70.86%	69.09%	0.76
ResNet-34		**67.90%**	0.55	**72.07%**	**73.23%**	**72.50%**	**0.80**
ResNet-50		66.36%	0.50	70.27%	71.45%	70.66%	0.79
ResNet-18	Weighted κ	69.59%	**0.58**	72.07%	73.21%	72.43%	0.78
ResNet-34		**69.64%**	0.51	**72.97%**	**74.86%**	**73.65%**	**0.81**
ResNet-50		67.14%	**0.58**	71.17%	71.47%	71.20%	0.78

In an attempt to leverage the classification learning task, after choosing the best model, we re-trained this version by adding some individual labelled tiles ($n = 263$) to the training set, to guide model training with direct supervision of some tiles. In Table 7.6 (middle row) we can see that, with this addition, all metrics increased, ending with an accuracy of 74.31% and a sensitivity of 74.77%. As expected, by combining the selected tile of each epithelium area, that only has the label of the correspondent bag, with tiles that have a particular labelled associated, the tile selection process was improved.

Table 7.6: Performance of the classification model (ResNet-34) trained with the weighted κ loss function and different supervision levels: tiles from annotated epitheliums (AE), annotated tiles (AT) and tiles from labelled slides (LS).

Training data	**Accuracy**	**QWK**	**Sensitivity**	**Precision**	**F1-score**	**AUC**
AE	69.64%	0.51	72.97%	74.86%	73.65%	0.81
AE + AT	**74.31%**	0.65	74.77%	76.44%	74.98%	0.84
AE + AT + LS	73.78%	**0.66**	**78.27%**	**78.38%**	**78.31%**	**0.85**

Lastly, to take advantage of the complete dataset, we re-trained the ResNet-34 also adding bags of tiles ($n = 1198$) from the non-annotated slides, using the best epithelium segmentation model (Table 7.6, last row). It is worth mentioning that, despite adding more data, we also add more noise with the automatic epithelium segmentation and bags of tiles per slide. Nonetheless, the model achieved improved results across all metrics, except the balanced accuracy. In particular, sensitivity had a gain of 3.5%, with the model predicting right more cases from LSIL and HSIL classes combined. When comparing the confusion matrices of both versions (Table 7.7), we can conclude that the version trained with the non-annotated data (b) misidentified two more HSIL cases and one more normal case. On the other hand, it got seven more LSIL right. However, since HSIL and normal classes are less represented, such misclassifications are more penalised when computing the balanced accuracy.

Table 7.7: Validation set confusion matrices.

(a) AE+AT

		Actual class		
		NNeo	*LSIL*	*HSIL*
Predicted	*NNeo*	**10**	7	0
	LSIL	3	**45**	5
	HSIL	2	11	**28**

(b) AE+AT+LS

		Actual class		
		NNeo	*LSIL*	*HSIL*
Predicted	*NNeo*	**9**	5	0
	LSIL	4	**52**	7
	HSIL	2	6	**26**

7.2.3 Complete framework

Finally, we tested the complete framework with an independent set of slides ($n = 600$), using the best epithelium segmentation model (UNet with RGB channels) and the overall best classification model (ResNet-34 trained with the complete dataset). Here, the output of the segmentation model is used, not only for ROI identification but also to classify slides as "others". Since these samples don't have epithelium areas, if the segmentation model result is empty, then the slide is automatically classified accordingly.

From Table 7.8, it is possible to verify that the segmentation model only misses ROIs in two LSIL cases. However, in the non-representative cases, the model has identified 28 cases correctly out of 66, meaning a balanced accuracy of 71.03% for identifying negative and positive samples. From the over-segmented cases, the classification model does not recognise any as HSIL and classifies most of them as non-neoplastic. Thus, the overall framework achieves a balanced accuracy of 63.75%, precision of 71.02%, sensitivity 68.67% and an F1-score of 68.18%.

When excluding the "others" class, the classification model achieved a balanced accuracy of 71.07%, a QWK of 0.67, precision of 74.15%, sensitivity of 72.18%, F1-score of 72.11% and a mean AUC of 0.85%, being in line with the validation performance reported in Section 7.2.2 Therefore, we can conclude that the errors of the segmentation model still have some impact

Table 7.8: Test set ($n = 600$) confusion matrix using the complete framework, with all classes.

		Actual class			
		NNeo	*LSIL*	*HSIL*	*"others"*
Predicted	*NNeo*	**126**	21	0	22
	LSIL	73	**202**	24	16
	HSIL	5	25	**56**	0
	"others"	0	2	0	**28**

on the overall performance of the model, especially on precision and sensitivity. In fact, if the segmentation model misses some relevant area, the classifier would be misled.

7.3 Summary

In this chapter, we propose a weakly-supervised methodology for grading cervical dysplasia (non-neoplastic, LSIL, HSIL and non-representative cases), using different levels of training supervision, in an effort to gather a bigger dataset without the need of having all samples fully annotated. With the first step of segmentation, we can identify ROI to focus on for the classification, allowing the use of non-annotated WSI for training, and the automatic diagnosis of unseen cases. Then, the classifier is capable of diagnosing the grade of dysplasia in tiles from those areas.

Nonetheless, despite the overall acceptable performance of the complete framework on the test set, further efforts should focus on the improvement of both parts individually, but also on how to better link them. In fact, from the reported results we can conclude that there is some noise being propagated from the segmentation model to the classifier, weakening it. In that sense, an end-to-end training framework can possibly improve the results of the segmentation model by penalising them based on classification quality. Moreover, more information on the heterogeneity of epithelium types within a WSI could be used as an extra layer of weak supervision to guide tile selection.

Part IV

Breast cancer

Chapter 8

Breast cancer insights

8.1 Epidemiology

Breast Cancer (BCa) is the most commonly diagnosed cancer among women worldwide, affecting over 2 million women every year, representing about 25% of all oncological diagnoses in this gender. Moreover, BCa is the second leading cause (15%) of cancer death among women [124, 125]. Although BCa is more prevalent among women, it can also occur in men. During the most recent years, despite its incidence trends having increased, the mortality rate has significantly decreased, due to earlier detection and better treatment strategies [126].

8.2 BCa diagnosis & sub-typing

Breast cancer is considered a heterogeneous disease since its different types are characterised by variable histopathological, biological and genetic features that, in fact, result in different clinical outcomes and prognosis, as well as different responses to therapy [127]. Thus, concerning the therapeutic decision-making process, accurate classification of BCa into relevant sub-types comes out as an important task [128].

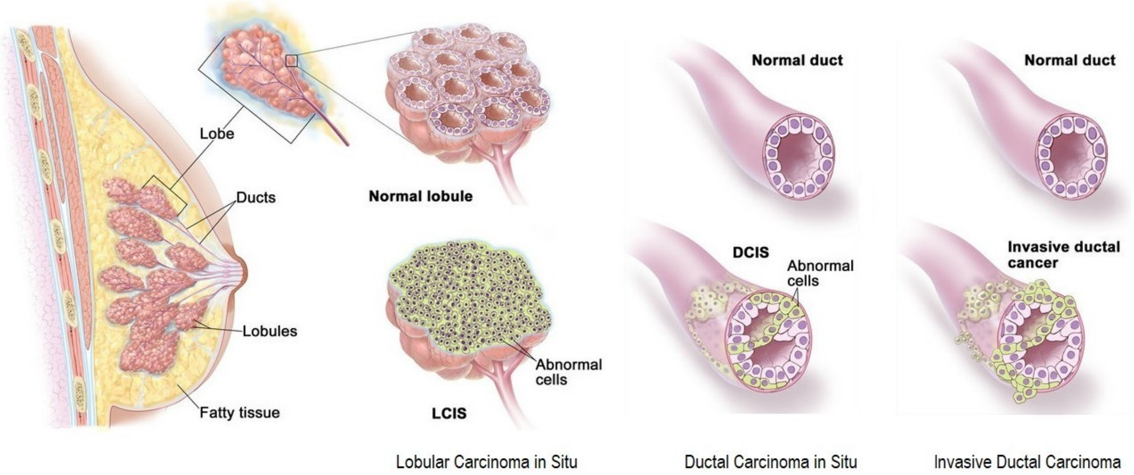
Histologically, BCa is classified accordingly to the tissue type that is affected. The most common histological BCa type is carcinoma, which starts in the cells of the breast lobules or ducts. Breast carcinomas are then classified as lobular/ductal carcinoma *in situ* or invasive lobular/ductal carcinoma (Figure 8.1):

- Lobular Carcinoma in Situ (LCIS): cancer grows in the milk-producing glands of the breast (lobules) but does not grow through their walls. Typically, it does not become invasive, but increases the risk of developing an invasive BCa;

- Ductal Carcinoma in Situ (DCIS): non-invasive or pre-invasive disease in which the cancerous cells grow among the duct but do not spread through its walls into the nearby breast tissue. Over time, DCIS may spread out of the duct and result in possible metastases;

- Invasive Lobular Carcinoma (ILC): starts in the lobules but, unlike LCIS, it can spread to other organs. It may be harder to detect and, when compared to other types of invasive carcinomas, it might increase the probability of developing cancer in both breasts;

- Invasive Ductal Carcinoma (IDC): starts in the milk ducts, breaks through their walls and grows into the nearby breast tissues. It also may be able to spread to other parts of the body, via the lymphatic and blood vessels. It is the most common type of invasive breast carcinomas [129].



Lobular Carcinoma in Situ

Ductal Carcinoma in Situ

Invasive Ductal Carcinoma

Figure 8.1: Breast cancer histological types (adapted from Terese Winslow LLC [1]).

Despite more than 80% of the diagnosed BCa are histologically classified as IDC, these cancers are biologically diverse and distinct, which implies a refined classification based on immunohistochemistry (IHC) markers, such as Estrogen Receptors (ER), Progesterone Receptors (PR), Human Epidermal growth factor Receptor - type 2 (HER2), protein Ki67 and basal markers [130]. Thus, breast tumours are classified into 5 intrinsic molecular subtypes (Table 8.1):

- Luminal A: are ER/PR positive, HER2 negative and have low levels of Ki67. This type is low-grade cancer and tends to grow slowly, having the best prognosis;

- Luminal B: are ER/PR positive, either HER2 positive or negative and have high levels of Ki67. Luminal B cancers generally grow slightly faster than the luminal A sub-type and have a slightly worse prognosis;

- Triple-negative: are ER, PR and HER2 negative. These tumours tend to occur more in younger women, are often aggressive and have a poorer prognosis than the luminal sub-types. However, they can be successfully treated;

[1] https://www.teresewinslow.com/breast/89t264tvm8t2fx014uof1ajok0twll

- HER2-enriched: are ER/PR negative, and HER2 positive. These cancers tend to grow faster than luminal A and B, and usually have a worse prognosis. However, they are often successfully treated with targeted therapies;
- Normal-like: similar to luminal A, with ER/PR positive, HER2 negative but with high levels of Ki67. Despite having a good prognosis, it is slightly worse than luminal A sub-type [131].

Table 8.1: Breast cancer molecular sub-types IHC profile and prognosis (adapted from [132]).

	IHC markers					**Proliferation**	**Outcome**
	HER2	**ER**	**PR**	**Ki67**	**basal**		
Luminal A	-	+	+	low	-	low	good
Luminal B	- or +	+	+	high	-	high	intermediate/ poor
Triple negative	-	-	-	high	+	high	poor
HER2-enriched	+	-	-	low/ intermediate	-/+	high	poor
Normal-like	-	+	+	high	-/+	low/ intermediate	intermediate

The analysis of tissue sections of cancer specimens (Figure 8.2) obtained by preoperative biopsy, commonly starts with haematoxylin and eosin (H&E) staining, which is usually followed by immunohistochemistry (IHC), a more advanced staining technique, used to highlight the presence of specific protein receptors [3]. In fact, according to the current clinical guidelines [133] for breast cancer management, Human epidermal growth factor receptor 2 (HER2) quantification in immunohistochemistry (IHC) must be routinely tested in all patients with invasive BCa, recurrence cases, and metastatic tumours. The overexpression of this receptor is observed in 10%–20%[133] of BCa cases and has been associated with aggressive clinical behaviour and poor prognosis [134]. However, patients diagnosed with HER2-positive BCa have a better response to targeted therapies and consequent improvements in healing and overall survival, which emphasises the importance of an accurate evaluation of HER2 status [135, 134].

The current guidelines [137], revised by the American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP), in 2018, indicate the following criteria for HER2 scoring:

- IHC 0+: no staining or incomplete membrane staining in 10% of tumour cells or less;
- IHC 1+: incomplete, barely perceptible membrane staining in $>$ 10% of tumour cells;
- IHC 2+: weak to moderate complete membrane staining in $>$ 10% of tumour cells;
- IHC 3+: circumferential, complete, intense membrane staining in $>$ 10% of tumour cells.

Moreover, cases scoring 0+ or 1+ are classified as HER2 negative, while cases with a score of 3+ are classified as HER2 positive. Cases with score 2+ are classified as equivocal and are further assessed by *in situ* hybridisation (ISH), to test for gene amplification (see Figure 8.2). In these cases, the HER2 status is given by the ISH result [137].

8.3 Computational pathology in BCa

Similarly to colorectal and cervical cancers, also the breast digital pathology field has opened many research opportunities in computer vision, as reviewed by Robertson *et al.* [2], with the analysis of digitised slides being used, for example, for mitosis detection [138–140], tissue classification [141–143], cancer grading [17, 144, 145] or histological sub-type classification [146, 19, 147]. However, only recently has research focused on predicting molecular markers directly on H&E-stained images and thus, the particular task of predicting HER2 status on H&E stained slides has not yet been extensively addressed in the literature. To the extent of our knowledge, the work proposed in the next chapter was one of the first studies on this purpose, and the first one developed on WSI. Despite some prior works on tissue micro-arrays (TMA), all the remaining were published after the HEROHE challenge [148], at ECDP 2020, that inspired the work presented in this thesis.

8.3.1 HER2 overexpression classification on H&E-stained slides

One of the first works on the feasibility of using digitised H&E–stained BCa slides for the prediction of molecular expression of biomarkers, was developed by Shamei *et al.* [149], in 2019. Particularly for the HER2 prediction, the authors used data from 4944 patients, which includes 3 H&E-stained TMA images and 1 IHC-stained TMA image per patient, in total 12789 TMA images. Each H&E TMA image is divided into non-overlapping tiles of 512×512 pixels, from which local features are extracted with a ResNet network. Then, all the features are averaged across all patches and across all the TMA images belonging to a patient, to obtain a final patient feature vector that is fed to a logistic regression classifier. The proposed approach achieved an AUC of 0.74, with an accuracy of 68%. Also using TMA, Rawat *et al.* [150] hypothesised that learning H&E morphological differences between patients could be seen and "fingerprints" to predict ER, PR, and HER2 status. In this sense, they proposed the usage of the ResNet-34 for the "fingerprints" extraction (combined with a cGAN for stain normalisation) from tiles with 224×224 pixels, that are used in a custom

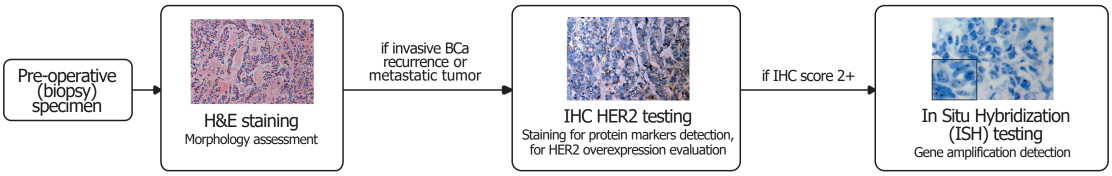

Pre-operative (biopsy) specimen
H&E staining
Morphology assessment
if invasive BCa recurrence or metastatic tumor
IHC HER2 testing
Staining for protein markers detection, for HER2 overexpression evaluation
if IHC score 2+
In Situ Hybridization (ISH) testing
Gene amplification detection

Figure 8.2: Schema of the process of BCa HER2 evaluation, involving H&E staining, IHC testing and, in specific cases, ISH testing. The proposed method aims to evaluate HER2 using only H&E stained slide images. Image examples were adapted from [136].

tile classifier, and which predictions are aggregated by average. They tested this framework in 124 TCGA samples and in 487 samples from the Australian Breast Cancer Tissue Bank (ABCTB), achieving an AUC of 0.71 and 0.79, respectively, for the HER2+ status prediction.

In 2020, following the participation in the HEROHE challenge mentioned above, Barbera *et al.* [151] proposed an approach based on a cascade of DL classifiers and MIL. Firstly, using the BACH Challenge dataset, they trained a classifier, based on the DenseNet-201 architecture, to distinguish tiles of 512×512 pixels into normal, benign, *in situ* carcinoma and invasive carcinoma tissue. Then, using the tumour tiles identified from the slides of the HEROHE challenge ($n = 360$), the authors trained a ResNet classifier to predict the HER2 status in all tiles, that are finally aggregated, by majority-voting, for the slide-level prediction. In the challenge test set, with 150 slides, the proposed model attained an accuracy of 68.7%, precision of 57.0%, recall of 88.3% and a F1-score of 0.69%.

Last year, Gamble *et al.* [152] developed individual learning systems to predict the ER, PR and HER2 status for both H%E tiles and WSI. Their work used annotated data from 3 different institutions, with reported AUC of 0.808 (95% confidence interval (CI): 0.802–0.813) and 0.60 (95%CI: 0.56–0.64) for HER2 predictions on the tile and slide levels, respectively. To develop the patch-level classifier, the authors acquired paired and aligned H&E and IHC slides, annotated by 16 expert pathologists, from which they randomly sample tiles of 512×512 pixels, to be categorised as biomarker positive, biomarker and "non-tumour", using an Inception-V3 network. Then, the slide-level prediction is based on features from the tiles predictions distribution. The proposed model was also tested on a subset of 870 slides from the TCGA dataset, achieving an AUC of 0.58 (95% CI: 0.53-0.63). Also in 2021, Bychkov *et al.* [153] proposed a model trained on TMA images and tested on WSI. The authors trained a tile classifier based on the "squeeze-and-excitation" network on TMA images and for inference on slides, they pulled the tile level scores by taking a median value within each WSI. The authors assessed the proposed model in two test sets: first on 354 TMA images and finally on an independent set of 712 WSI, with a reported AUC of 0.70 (95% CI: 0.63–0.77) and 0.67 (95% CI, 0.62–0.71), respectively.

In 2022, Farahmand *et al.* [154] trained an Inception-v3 classifier on non-overlapping tiles of 512×512 pixels of 188 H&E breast slides, manually annotated for tumour ROIs, to differentiate HER2 positive and negative areas. Then, the slide-level probabilities were computed by averaging the output of the tile classifier and final diagnosis was decided upon a 0.5 cutoff threshold. The proposed classifier achieved an AUC of 0.90 in cross-validation with a private dataset, and 0.81 on a set of 187 slides from the TCGA collection. More recently, Lu *et al.* [155] proposed a novel graph neural network (GNN) model, called SlideGraph+, to predict HER2 status directly from H&E-stained breast slides. The network was trained and tested on 709 TCGA slides, with cross-validtaion, and further tested on two other, the HER2Contest challenge dataset (HER2SC) and a private dataset, with a reported AUC of 0.75, 0.78 and 0.80, respectively, when differentiating unequivocal negatives (0+) and positive (3+) cases. Instead of extracting small tiles from the slide, the pipeline is based on a graph at the entire WSI-level, giving not only the overall prediction but also showing the most important image regions.

8.4 Summary

Breast cancer is a heterogeneous disease, affecting many women all over the world each year but, nowadays, despite the increased incidence, it has a decreased mortality. In fact, its earlier detection is of utmost importance for an efficient treatment. Moreover, an accurate disease sub-typing, staging and grading, assessed on histopathology data, is also vital for tailored clinical management, towards cancer elimination and quality of life improvement. Particularly for disease prognosis and treatment planning, assessing IHC markers is of utmost importance, with the HER2 status being associated with disease aggressiveness. Thus, a preliminary assessment of this marker expression on standard H&E samples could be a valuable indicator before IHC testing, enabling earlier identification of more severe cases.

Chapter 9

HER2 profile from H&E breast slides

Author Contributions The research work described in this chapter was conducted in collaboration with João R. Pinto, Tiago Gonçalves and Rita C. Marques, under the clinical supervision of Maria J. Cardoso, and the technical supervision of Hélder Oliveira and Jaime S. Cardoso. The author of this thesis contributed to this work on the problem conceptualisation, data curation, the preparation and conduction of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- S.P. Oliveira, J.R. Pinto, T. Gonçalves, R. Canas-Marques, M.J. Cardoso, H.P. Oliveira and J.S. Cardoso. Weakly-Supervised Classification of HER2 Expression in Breast Cancer Haematoxylin and Eosin Stained Slides. *Applied Sciences*, 10(14):4728, 2020
- S.P. Oliveira, J.R. Pinto, T. Gonçalves, H.P. Oliveira and J.S. Cardoso. "IHC Classification in Breast Cancer H&E Slides with a Weakly-Supervised Approach. *In 26th Portuguese Conference in Pattern Recognition (RECPAD)*, 2020

At the moment, besides very well-differentiated tumours, with low nuclear/cytoplasm area ratio, which typically are hormonal driven and therefore generally not positive for HER2, there are no morphological features on H&E slides that allow a reliable prediction of the HER2 status. Therefore, the standard procedure is to perform an additional immunohistochemical study, with an additional molecular study in case of equivocal results. However, despite the efficiency of IHC and ISH, the additional cost and time spent on these tests might be avoided if all the information needed to infer the HER2 status could be extracted only from H&E whole slide images (WSI), as a preliminary indication of the IHC result. Thus, in this chapter we propose a method using a convolutional neural network (CNN), inspired by multiple-instance learning (MIL), to automatically identify the HER2 status on BCa H&E stained slides. To deal with the sheer dimensions of the slides, tiles are extracted from the original images and separately processed by the model, which learns to aggregate the individual tile predictions into a single, image-wide label. Moreover, to introduce some prior knowledge about the morphology of tissue structures into the model, the CNN has been pretrained with HER2 IHC-stained slides.

9.1 Methodology

The proposed method (Figure 9.1) comprises a convolutional neural network (CNN), which is pretrained for the task of HER2 scoring of tiles extracted from IHC stained slides. The pretrained parameters are then transferred to the task of HER2 status prediction on H&E staining slide tiles, to provide the network with some knowledge of the tissue structures' appearance. Individual tile scores are then combined to obtain a single label for the respective slide. The data preprocessing methodology and the implemented networks are described below.

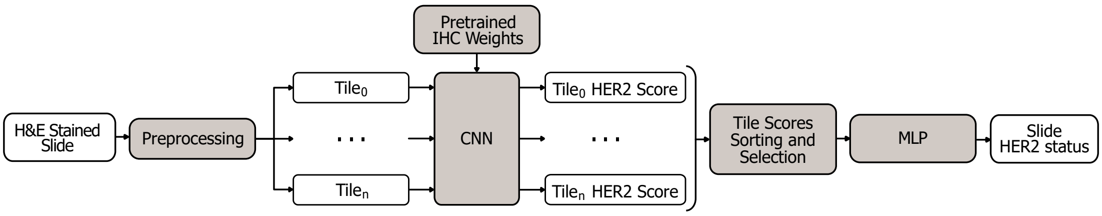


Figure 9.1: The proposed approach for weakly-supervised HER2 status classification on BCa H&E stained slides.

9.1.1 IHC-stained slides pre-processing

For the IHC stained slides of classes 2+ and 3+, the preprocessing begins with automatic tissue segmentation with Otsu's thresholding on the saturation (S) channel of the HSV colour space, obtaining the regions with more intense staining, that correspond to the HER overexpression areas. For slides of classes 0+ and 1+, the segmentation consists of the simple removal of pixels with the greatest HSV value (V) intensity, corresponding to background pixels, which do not contain essential information to the problem. These processes, which are performed at 32× downsampled slides, return the masks used in tile extraction.

Tiles with size 256 × 256 are extracted from the slide with original dimensions (without downsampling), provided they are completely within the mask region. These tiles are converted from RGB to HSL colour space, of which only the lightness (L) channel is used. Each tile inherits the class from the respective slide (examples in Figure 9.2a–d), turning the learning task into a weakly-supervised problem.

9.1.2 H&E-stained slides pre-processing

According to the ASCO/CAP guidelines for IHC evaluation, the diagnosis is based only on the tumour region of the slides. Hence, the preprocessing of H&E stained slides begins with an automatic invasive tissue segmentation with the HASHI method [158, 159], which consists of an adaptive gradient-based sampling approach that iteratively refines an initial coarse invasive BCa probability map, from CNN inference. The algorithm begins with a WSI as input, sampled in 100 tiles, each classified using a CNN-trained model, to obtain the probability of invasive BCa presence. By interpolating each tile probability, a heatmap is generated for the entire WSI. Then, the

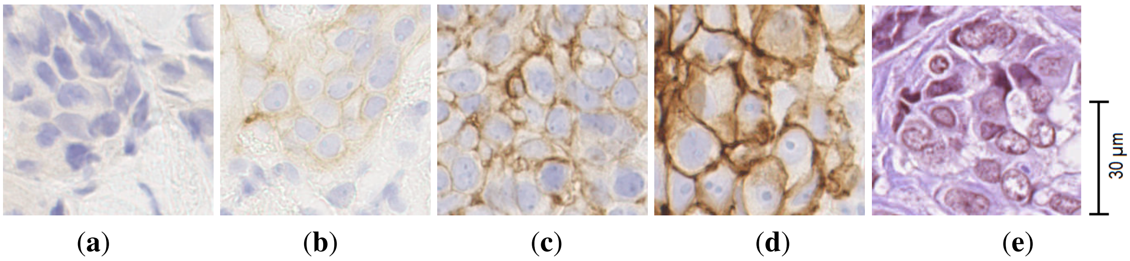


Figure 9.2: Tile examples extracted from IHC 0+ (**a**), IHC 1+ (**b**), IHC 2+ (**c**), IHC 3+ (**d**), H&E (**e**) slides. Tiles from IHC 2+ and 3+ and H&E slides were obtained by Otsu's thresholding and the remaining were obtained by simply removing the pixels with background value. The IHC tiles were obtained from slides of the HER2SC dataset [156] and the H&E tile was obtained from a slide of the BRCA dataset [91, 157].

gradient of the map is calculated and used to prioritise the sampling selection on the next iteration. The process is repeated during 20 iterations [158]. The method was implemented in the images referred to by Cruz-Roa et al. [158] as the test set, using the original magnification and extracting squared 512 × 512 tiles. Moreover, to exclude eventual small background zones included in HASHI segmentation, this mask region was intersected with the segmentation using Otsu's thresholding on the saturation (S) channel of the HSV colour space. The final segmentation mask was then used to generate H&E tiles (example in Figure 9.2e), extracted and processed accordingly to the methodology described for IHC slides. The number of tiles per slide varies according to the extent of the tissue region.

9.1.3 CNN for IHC tile scoring

The CNN architecture (Figure 9.3) consists of four convolutional layers (16, 32, 64 and 128 filters, respectively, with ReLU activation). The first layer has 5 × 5 square kernels, while the remaining have 3 × 3 square kernels. Each convolutional layer is followed by one pooling layer (a max-pooling function without overlap, with kernel 2 × 2). The network is topped with three fully-connected layers, with 1024, 256, and 4 units, respectively. The first two have ReLU activation, while the third is followed by softmax activation for the output of probabilities for each class.

9.1.4 CNN for H&E-stained slide classification

The network parameters pretrained with IHC stained slides were used as initial network weights for HER2 status classification on H&E stained slides. It is worth mentioning that IHC data is only used for the network pretrain, and not during the inference/test phase. To achieve a single prediction per tile, instead of four (as it was initially trained for on the IHC setting), a soft-argmax activation [160, 161] replaces the softmax activation, following the equation

$$\text{soft-argmax}(s) = \sum_i \text{softmax}(\beta s_i) i, \tag{9.1}$$

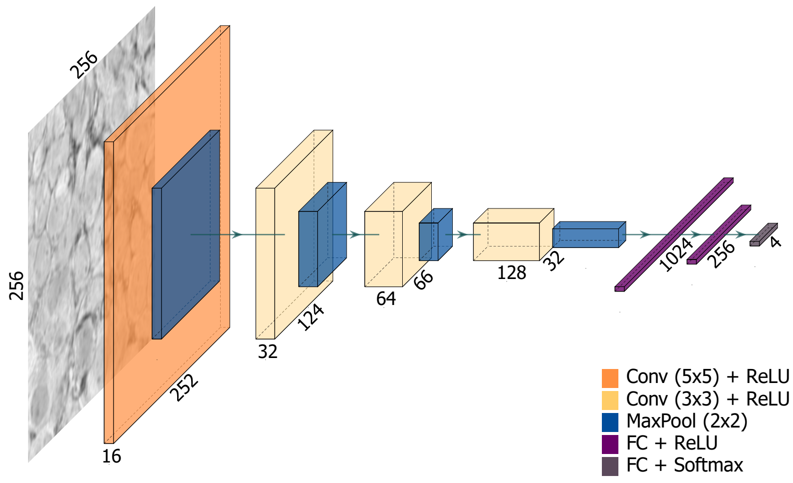


Figure 9.3: Architecture of the implemented convolutional neural network.

where β is an adjustment factor which controls the range of the probability map given by the softmax, s is the tile score array, and i is the index that corresponds to each class.

Having a single value per tile enables the easy sorting of tiles, which is performed before the aggregation into a single HER2 label. With the HER2 scores of each tile, output by the soft-argmax activation, tiles are sorted from 3+ to 0+. Then, the 15% highest scores are selected to serve as input to the aggregation process. This percentage was chosen to limit the information given to the aggregation network, while still including and barely exceeding the reference 10% of tumour area considered in the HER2 scoring guidelines.

The score aggregation is performed by a multilayer perceptron (MLP), composed of four layers, with 256, 128, 64, and 2 neurons, respectively. All layers are followed by ReLU activation, except the last layer, which is followed by softmax activation. Since the input dimension M of the MLP is fixed (we set $M = 300$ in our experimental analysis, to limit memory cost), for images where 15% of the number of tiles exceeds M, we downsample to 300 using evenly distributed tile selection. In cases where 15% of the number of tiles is lower than M, tiles are extracted with overlap, to guarantee that M tiles can be selected. The MLP will process these 300 HER2 scores and output a single HER2 status label for the respective slide.

9.1.5 Training Details

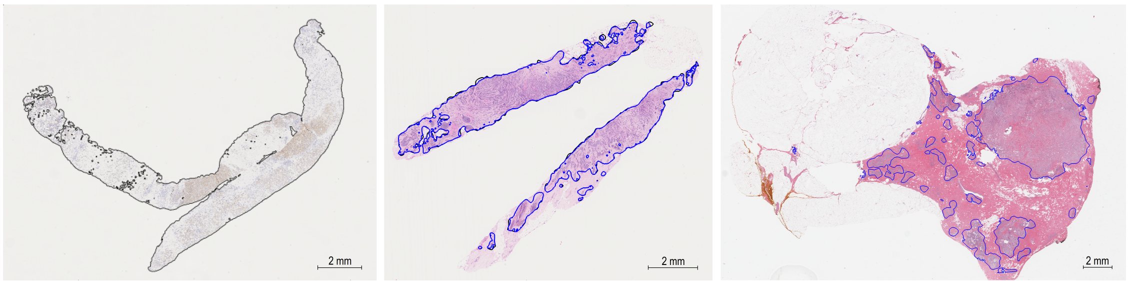
The hyperparameters used during training were empirically set to maximise performance. The CNN model for IHC tile scoring was randomly initialised and trained using the Adaptive Moment Estimation (Adam) [94] optimiser (learning rate of 1×10^{-5}), to minimise a cross-entropy loss function, during 200 epochs, with mini-batches of 128 tiles. The soft-argmax used a parameter $\beta = 1000$. The aggregation MLP was trained using the Adam optimiser, with a learning rate of 10^{-5} for 150 epochs and mini-batches of 1 WSI (consisting of soft-argmax scores of the respective 300 tiles), saving the best considering validation accuracy.

9.2 Datasets

The dataset is composed of subsets of WSI from two public datasets: the HER2 Scoring Contest (HER2SC) training set [156] and the TCGA-TCIA-BRCA (BRCA) collection [91, 157]. The HER2SC training set (the subset with available labelling) comprises WSI of sections of 52 cases of invasive BCa stained with both IHC and H&E (example in Figure 9.4a,b). From this set, all IHC and H&E stained slides were used, except 4 H&E excluded because of manual ink markings. The subset from the BRCA dataset includes 54 H&E stained WSI (example in Figure 9.4c). All slides have the same original resolution and are weakly annotated with HER2 status (negative/positive) and score (0+, 1+, 2+, 3+), obtained from the corresponding histopathological reports.

The IHC stained slides were manually segmented into regions of interest (ROI), using the Sedeen Viewer software [79]. However, it is noteworthy that these slides were only used for training and, thus this step is not needed for testing.

The training and validation sets, used for model parameter tuning and optimisation, have 40 and 12 IHC slides, respectively. A total of 7591 tiles per class have been extracted for training (30,364 tiles total) and 624 tiles per class extracted for validation (2496 tiles total), to keep a class balance.



(**a**) (**b**) (**c**)

Figure 9.4: Image examples from used datasets: HER2SC [156] IHC stained slides (**a**), HER2SC [156] H&E stained slides (**b**), BRCA [91, 157] H&E stained slides (**c**). The tile extraction was solely done on tissue, here denoted by the delineated regions.

9.3 Experimental results and discussion

9.3.1 Individual IHC tile scoring results

After training, the model offered 76.8% accuracy (see Table 9.1). This indicates that the model was able to discriminate against the IHC tiles between the four classes adequately. This model was subsequently transferred for HER2 scoring in tiles from H&E slides.

9.3.2 Invasive tumour tissue segmentation

Tiles from H&E WSI are extracted from the intersection area between the HASHI-based invasive tumour segmentation and the Otsu-based tissue segmentation. The HASHI segmentation method was trained on the BRCA data reported as the test set by Cruz-Roa et al. [158], with 179 WSI on

Table 9.1: Confusion matrix of the CNN for HER2 scoring in IHC tiles.

		Actual class			
		0	*1*	*2*	*3*
Predicted	*0*	**490**	132	2	0
	1	176	**384**	64	0
	2	45	159	**419**	1
	3	0	0	1	**623**

their original magnification. The results were comparable to the original paper (see examples in Figure 9.4) and were further evaluated by a pathology specialist, who confirmed the adequacy of the invasive tumour segmentation results.

9.3.3 Slide Scoring

On the HER2SC test set, the method achieved an F1-score of 86.7% and a weighted accuracy of 83.3% (Table 9.2). Despite the small size of the test set, the proposed method was able to correctly classify all positive WSI and only misclassify one negative sample. In this context, one might consider this a desirable behaviour, as false positives are less impactful than false negatives.

Table 9.2: Evaluation of the proposed method on the HER2SC and BRCA test sets.

	Accuracy	**F1-Score**	**Precision**	**Recall**
HER2SC	83.3%	86.7%	89.6%	87.5%
BRCA	53.8%	21.5%	81.2%	31.5%

When tested on the BRCA test set, this method achieved an F1-score of 21.5% and a weighted accuracy of 53.8% (see Table 9.2). The method retains the behaviour presented in HER2SC, preferring to err on the side of false positives than the alternative. On the other hand, the performance metrics on BRCA differ considerably from those obtained on HER2SC. While the method was trained on HER2SC data, which is expected similar to the test data, the WSI of the BRCA dataset presents some notable differences. These slides have a greater extent of tissue, which generates more tiles, impacting the distribution of the scores, which may influence the method's behaviour.

The evaluation results in single-database (HER2SC) and cross-database (BRCA) settings show the potential of the proposed method in standard and more challenging situations. However, the method appears to be dataset-dependent: it performed much better in conditions similar to the training. This should be addressed with additional efforts regarding domain adaptation.

The other shortcomings of the method appear to be related to the invasive tumour tissue segmentation and the individual tile scoring network, which could be improved with additional data and more accurate ground truth information. With these additional efforts, the proposed method could offer robust weakly-supervised WSI HER2 classification without IHC information.

9.3.4 Ablation study

Considering the lack of literature methods, to perform a benchmark, an ablation study was performed to confirm the capabilities of the proposed method. Experiments were conducted without IHC individual tile scoring CNN initialisation, and using alternative statistical methods for individual tile score aggregation instead of MLP (median and mean), as can be seen in Tables 9.3 and 9.4. The results show that these alternatives are, in most settings, less adequate for the task at hand.

Table 9.3: Results on the HER2SC test set.

Method	**Accuracy**	**F1-Score**	**Precision**	**Recall**
MLP Aggregation:				
proposed method	**83.3%**	**86.7%**	**89.6%**	**87.5%**
w/out pretrained CNN weights	62.5%	48.1%	39.1%	62.5%
Median Aggregation:				
w/pretrained CNN weights	50.0%	43.3%	78.6%	50.0%
w/out pretrained CNN weights	62.5%	48.1%	39.1%	62.5%
Mean Aggregation:				
w/pretrained CNN weights	50.0%	43.3%	78.6%	50.0%
w/out pretrained CNN weights	62.5%	48.1%	39.1%	62.5%

Table 9.4: Results on the BRCA test set.

Method	**Accuracy**	**F1-Score**	**Precision**	**Recall**
MLP Aggregation:				
proposed method	**53.3%**	**21.5%**	**81.2%**	**31.5%**
w/out pretrained CNN weights	50.0%	60.3%	51.8%	72%
Median Aggregation:				
w/pretrained CNN weights	50.0%	12.3%	7.80%	28.0%
w/out pretrained CNN weights	52.2%	63.5%	66.5%	72.0%
Mean Aggregation:				
w/pretrained CNN weights	50.0%	12.3%	7.80%	28.0%
w/out pretrained CNN weights	52.2%	63.5%	66.5%	72.0%

It is noteworthy that the median and mean-based aggregation are followed by a conversion to binary classes (0+ and 1+ are considered negative, while 2+ and 3+ are considered positive) since tiles have four possible labels. According to the guidelines, 2+ cases can be either negative or positive, but in an uncertain diagnosis scenario, it is preferable to classify them as positive.

9.4 Summary

In this chapter, a framework is proposed for the weakly supervised classification of HER2 overexpression status on H&E stained BCa WSI. The proposed approach integrates a CNN trained for HER2 scoring of individual H&E-stained slide tiles, initialised with the network parameters pretrained with data from IHC-stained images. The objective of this initialisation is to transfer some domain knowledge to the final training. The individual scores are aggregated on a single prediction per slide, returning the HER2 status label.

Tested with the BRCA data subset, the proposed method attained suitable performance. These preliminary results indicate that it is possible to accurately infer BCa HER2 status solely from H&E-stained slides. The results of an ablation study suggest that the proposed method with MLP tile score aggregation is more promising than simpler aggregation methods (mean or median).

Despite these results, further efforts should be devoted to performance improvements in the proposed task of diagnosing an IHC marker directly on H&E. Particularly for BCa HER2, firstly, the classifier and the aggregator could be integrated into a single optimisation process. On the other hand, the aggregation of individual scores could incorporate information on tile location, to take spatial consistency into account. Finally, the knowledge embedded in the networks through the pre-trained parameters could be better seized if input H&E tiles could be previously converted into IHC, using generative adversarial models (GANs), for example.

Part V

Epilogue

Chapter 10

Conclusion

More and more cancer cases and precursor lesions from screening programs are arriving at surgical pathology laboratories, significantly increasing the pathologists' workload. An early diagnosis and treatment promote an optimistic prognosis, so, a fast and accurate diagnosis is mandatory. Additionally, pathologists are experiencing added pressure as personalised therapeutics require a more detailed histological evaluation and precise biomarker assessment, while, at the same time, there are gradually fewer trained pathology specialists in the world [8]. Moreover, as previously described, grading and diagnosing depend on the pathologists' knowledge and experience, which means there is inevitable subjectivity in this process.

Using AI technology could help to automatically classify and diagnose pathological samples, improving diagnostic accuracy, while reducing time and resources [43]. In fact, AI can take up the effort of laborious, tedious tasks, leaving time for pathologists to appraise the most relevant lesions. For instance, it could help pathologists to prioritise high-risk samples, or could also be used as a second opinion, to help to confirm a diagnosis. Additionally, there are now many studies evaluating the value of AI solutions as prediction tools, in an attempt to be able to extrapolate molecular features and predict survival and therapy response [162].

In sum, due to the workload increase and added difficulties for pathologists in recent years, procedures will need to be revised and adapted in order to have the best balance between diagnostic/prognostic capacity and daily routine feasibility. Digital pathology, along with AI/CAD solutions, will certainly help in this purpose.

10.1 Summary of contributions

This thesis addressed the problem of developing medical image diagnostic tools for digital pathology, particularly for colorectal (Part II), cervical (Part III) and breast cancers (Part IV), without the need for models' full supervision. With the proposed setups, CAD systems can be developed without the need for extensively annotated datasets, reducing the burden of annotation for pathologists, and maintaining good performance, which is essential for the application in clinical practice. Its outcomes have been published in international journals and conferences, taking advantage of collaborations with medical institutions, such as IMP Diagnostics and Champalimaud Foundation, and also joint efforts with other members of the VCMI research group, at INESC TEC. These collaborations enhanced the focus of the work, keeping the medical reasoning always present and helping to bridge the machine learning challenges with the practical needs of the clinical side. The contributions throughout this document are summarised as follows:

- a feasibility study on using partially annotated datasets to drive the development of CAD systems for digital pathology, directly from WSI. This study was performed particularly for CRC diagnosis, but can be generalised to other pathologies. This work served as the basis for building one of the largest pathology datasets publicly available, with about 10,000 colorectal H&E-stained slides;
- a semi-supervised learning approach to construct a novel system for CRC automatic diagnosis on slides, with the capability to guide pathologists' attention towards the most relevant tissue areas within each WSI;
- an AI-based clinical software prototype for colorectal samples grading and tissue mapping, developed as a server-side web application, together with a clinical validation of the last proposed model;
- a weakly-supervised framework to grade dysplasia directly from cervical WSI, that firstly identifies ROIs and then classifies them. In this approach, we propose a three-level annotation: only slide labelling for most of the dataset, epithelium annotation in a subset of slides, and identification of small areas within the annotated epithelium areas, that serve as both epithelium labelling and individual tiles annotation;
- the first work on the classification of HER2 overexpression status on H&E-stained BCa slides, without pixel-level annotations.

10.2 Final remarks & future work

In general, despite the growing popularity and availability of computational pathology works, there are relatively few published works on diagnosis from WSI, and most of these are based on relatively small and private datasets, which renders them fragile and makes direct comparisons not so fair. Moreover, another issue has been largely ignored in digital pathology research: a developed model should not be specific to scanning machine output or to particular laboratory configurations. The broad knowledge acquired by the model during the training phase should not be wasted or useless. It is then necessary that the models can either be directly generalised to other scanning machines, or that a few samples of non-annotated WSIs are sufficient to fine-tune a model with similar performance to the original.

From a computational point of view, the high dimensionality of pathology data probably remains the biggest problem to solve. Images are too large to be fed to GPUs in their original dimensions, which is not easy to solve without either decomposing them into smaller parts (tiles), at the cost of losing spatial context, or resizing them, at the cost of losing resolution, essential especially in diagnostic tasks. For improved performance, attention mechanisms, multiclass frameworks and/or multi-level approaches can be beneficial for models to learn to diagnose better, even if they don't see all the information at once.

From a clinical point of view, although achieving remarkable performance, medical applications of DL-based methods have been severely criticised due to their natural black-box structure. In fact, for these approaches to be used in practice, researchers must develop techniques to inform pathologists about the spatial location that was most responsible for the diagnosis and to explain the reasons for the prediction. The ultimate goal should be to create, not only robust but also transparent systems that clinicians can trust and rely on.

In conclusion, despite the highlighted contributions, with state-of-the-art results in different tasks, the construction of CAD systems for decision support in digital pathology is far from being completely solved. In fact, computational pathology is still a young area, with many challenges to solve, so machine learning models can effectively get closer to clinical applicability. However, this doctoral project is, hopefully, a step forward in the path that is still left.

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[2] Stephanie Robertson, Hossein Azizpour, Kevin Smith, and Johan Hartman. Digital image analysis in breast pathology - from image processing techniques to artificial intelligence. *Translational Research*, 194:19–35, April 2018.

[3] M. Veta, J. P. W. Pluim, P. J. van Diest, and M. A. Viergever. Breast Cancer Histopathology Image Analysis: A Review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411, 2014.

[4] Catarina Eloy, João Vale, Mónica Curado, António Polónia, Sofia Campelos, Ana Caramelo, Rui Sousa, and Manuel Sobrinho-Simões. Digital pathology workflow implementation at ipatimup. *Diagnostics*, 11(11), 2021.

[5] Filippo Fraggetta, Alessandro Caputo, Rosa Guglielmino, Maria Giovanna Pellegrino, Giampaolo Runza, and Vincenzo L'Imperio. A survival guide for the rapid transition to a fully digital workflow: The "caltagirone example". *Diagnostics*, 11(10), 2021.

[6] Diana Montezuma, Ana Monteiro, João Fraga, Liliana Ribeiro, Sofia Gonçalves, André Tavares, João Monteiro, and Isabel Macedo-Pinto. Digital pathology implementation in private practice: Specific challenges and opportunities. *Diagnostics*, 12(2):529, 2022.

[7] Emad A Rakha, Michael Toss, Sho Shiino, Paul Gamble, Ronnachai Jaroensri, Craig H Mermel, and Po-Hsuan Cameron Chen. Current and future applications of artificial intelligence in pathology: a clinical perspective. *Journal of Clinical Pathology*, 2020.

[8] B. Acs, M. Rantalainen, and J. Hartman. Artificial intelligence as the next step towards precision pathology. *J Intern Med*, 288(1):62–81, 2020.

[9] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016.

[10] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D. Caie. Deep learning for whole slide image analysis: an overview. *Front. Med.*, 6:264, 2019.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[12] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J Pathol Inform*, 7, 2016.

[13] Ludovic. Roux, Daniel. Racoceanu, Nicolas. Loménie, Maria. Kulikova, Humayun. Irshad, Jacques. Klossa, Frédérique. Capron, Catherine. Genestie, Gilles. Naour, and Metin. Gurcan. Mitosis detection in breast cancer histological images: an ICPR 2012 contest. *J Pathol Inform*, 4(1), 2013.

[14] Mitko Veta, Paul J. van Diest, Stefan M. Willems, Haibo Wang, Anant Madabhushi, Angel Cruz-Roa, Fabio Gonzalez, Anders B.L. Larsen, Jacob S. Vestergaard, Anders B. Dahl, Dan C. Cireşan, Jürgen Schmidhuber, Alessandro Giusti, Luca M. Gambardella, F. Boray Tek, Thomas Walter, Ching-Wei Wang, Satoshi Kondo, Bogdan J. Matuszewski, Frederic Precioso, Violet Snell, Josef Kittler, Teofilo E. de Campos, Adnan M. Khan, Nasir M. Rajpoot, Evdokia Arkoumani, Miangela M. Lacle, Max A. Viergever, and Josien P.W. Pluim. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis*, 20(1):237–248, 2015.

[15] Nathan Ing, Zhaoxuan Ma, Jiayun Li, Hootan Salemi, Corey Arnold, Beatrice S. Knudsen, and Arkadiusz Gertych. Semantic segmentation for prostate cancer grading by convolutional neural networks. In John E. Tomaszewski and Metin N. Gurcan, editors, *Medical Imaging 2018: Digital Pathology*, volume 10581, pages 343–355. SPIE, 2018.

[16] Shidan Wang, Donghan M. Yang, Ruichen Rong, Xiaowei Zhan, and Guanghua Xiao. Pathology image analysis using segmentation deep learning algorithms. *American Journal of Pathology*, 189(9):1686–1698, September 2019.

[17] Tao Wan, Jiajia Cao, Jianhui Chen, and Zengchang Qin. Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features. *Neurocomputing*, 229:34–44, 2017. Advances in computing techniques for big medical image data.

[18] An Hoai Truong, Viktoriia Sharmanska, Clara Limbäck-Stanic, and Matthew Grech-Sollars. Optimization of deep learning methods for visualization of tumor heterogeneity and brain tumor grading through digital pathology. *Neuro-Oncology Advances*, 2(1), 2020.

[19] Duc My Vo, Ngoc-Quang Nguyen, and Sang-Woong Lee. Classification of breast cancer histology images using incremental boosting convolution networks. *Information Sciences*, 482:123–138, 2019.

[20] Nelson Zange Tsaku, Sai Chandra Kosaraju, Tasmia Aqila, Mohammad Masum, Dae Hyun Song, Ananda M. Mondal, Hyun Min Koh, and Mingon Kang. Texture-based deep learning for effective histopathological cancer image classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 973–977, 2019.

[21] Luís A Vale-Silva and Karl Rohr. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1):1–12, 2021.

[22] Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2022.

[23] Hamid. Tizhoosh and Liron. Pantanowitz. Artificial intelligence and digital pathology: challenges and opportunities. *J. Pathol. Inform*, 9(1), 2018.

[24] Esther Abels, Liron Pantanowitz, Famke Aeffner, Mark D Zarella, Jeroen van der Laak, Marilyn M Bui, Venkata NP Vemuri, Anil V Parwani, Jeff Gibbs, Emmanuel Agosto-Arroyo, Andrew H Beck, and Cleopatra Kozlowski. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *J. Pathol*, 249(3):286–294, 2019.

[25] Andrey Bychkov and Junya Fukuoka. Evaluation of the global supply of pathologists (1257). In *USCAP 2022 Abstracts: Quality and Patient Safety*, volume 35(Suppl. 2), page 1473–1522. Modern Pathology, 2022.

[26] International Agency for Research on Cancer (IARC). Global cancer observatory. `https://gco.iarc.fr/`, 2021.

[27] WHO Classification of Tumours Editorial Board. *Digestive System Tumours*. WHO Tumour series. International Agency for Research on Cancer, 5th edition, 2019.

[28] Herb Brody. Colorectal cancer. *Nature*, 521:S1, 2015.

[29] David Holmes. A disease of growth. *Nature*, 521:S2–S3, 2015.

[30] Rebecca L. Siegel, Kimberly D. Miller, Ann Goding Sauer, Stacey A. Fedewa, Lynn F. Butterly, Joseph C. Anderson, Andrea Cercek, Robert A. Smith, and Ahmedin Jemal. Colorectal cancer statistics 2020. *CA A Cancer J Clin*, 70(3):145–164, 2020.

[31] Digestive Cancers Europe (DiCE). Colorectal screening in europe. `www.digestivecancers.eu/wp-content/uploads/2020/02/466-Document-DiCEWhitePaper2019.pdf`, 2019.

[32] Cesare Hassan, Giulio Antonelli, Jean-Marc Dumonceau, Jaroslaw Regula, Michael Bretthauer, Stanislas Chaussade, Evelien Dekker, Monika Ferlitsch, Antonio Gimeno-Garcia, Rodrigo Jover, Mette Kalager, Maria Pellisé, Christian Pox, Luigi Ricciardiello, Matthew Rutter, Lise Mørkved Helsingen, Arne Bleijenberg, Carlo Senore, Jeanin E van Hooft, Mario Dinis-Ribeiro, and Enrique Quintero. Post-polypectomy colonoscopy surveillance: European society of gastrointestinal endoscopy guideline - update 2020. *Endoscopy*, 52(8):687–700, 2020.

[33] Lucie de Jonge, Joachim Worthington, Francine van Wifferen, Nicolas Iragorri, Elisabeth F. P. Peterse, Jie-Bin Lew, Marjolein J. E. Greuter, Heather A. Smith, Eleonora Feletto, Jean H. E. Yong, Karen Canfell, Veerle M. H. Coupé, and Iris Lansdorp-Vogelaar. Impact of the covid-19 pandemic on faecal immunochemical test-based colorectal cancer screening programmes in Australia, Canada, and the Netherlands: a comparative modelling study. *The Lancet Gastroenterology & Hepatology*, 6(4):304–314, 2021.

[34] Luigi Ricciardiello, Clarissa Ferrari, Michela Cameletti, Federica Gaianill, Francesco Buttitta, Franco Bazzoli, Gian Luigi de'Angelis, Alberto Malesci, and Luigi Laghi. Impact of SARS-CoV-2 pandemic on colorectal cancer screening delay: effect on stage shift and increased mortality. *Clinical Gastroenterology and Hepatology*, 2020.

[35] Dipti Mahajan, Erinn Downs-Kelly, Xiuli Liu, Rish K. Pai, Deepa T. Patil, Lisa Rybicki, Ana E. Bennett, Thomas Plesec, Oscar Cummings, Douglas Rex, and John R. Goldblum. Reproducibility of the villous component and high-grade dysplasia in colorectal adenomas <1 cm: Implications for endoscopic surveillance. *American Journal of Surgical Pathology*, 37(3):427–433, March 2013.

[36] Jeff K Turner, Geraint T Williams, Meleri Morgan, Melissa Wright, and Sunil Dolwani. Interobserver agreement in the reporting of colorectal polyp pathology among bowel cancer screening pathologists in wales. *Histopathology*, 62(6):916–924, 2013.

[37] Allison Osmond, Hector Li-Chang, Richard Kirsch, Dimitrios Divaris, Vincent Falck, Dong Feng Liu, Celia Marginean, Ken Newell, Jeremy Parfitt, Brian Rudrick, Heidi Sapp, Sharyn Smith, Joanna Walsh, Fasahat Wasty, and David K Driman. Interobserver variability in assessing dysplasia and architecture in colorectal adenomas: a multicentre canadian study. *Journal of Clinical Pathology*, 67(9):781–786, 2014.

[38] Samir Gupta, David Lieberman, Joseph C. Anderson, Carol A. Burke, Jason A. Dominitz, Tonya Kaltenbach, Douglas J. Robertson, Aasma Shaukat, Sapna Syngal, and Douglas K. Rex. Recommendations for follow-up after colonoscopy and polypectomy: A consensus update by the us multi-society task force on colorectal cancer. *Gastrointestinal Endoscopy*, 2020.

[39] Public Health England. Reporting lesions in the NHS BCSP: guidelines from the bowel cancer screening programme pathology group. `https://www.gov.uk/government/publications/bowel-cancer-screening-reporting-lesions#history`, 2018.

[40] Phil Quirke, Mauro Risio, René Lambert, Lawrence von Karsa, and Michael Vieth. Quality assurance in pathology in colorectal cancer screening and diagnosis - european recommendations. *Virchows Archiv*, 458(1):1–19, 2011.

[41] Pathology Working Group of the Canadian Partnership Against Cancer. Pathological reporting of colorectal polyps: pan-canadian consensus guidelines. `http://canadianjournalofpathology.ca/wp-content/uploads/2016/11/cjp-volume-4-isuue-3.pdf`, 2012.

[42] Nishant Thakur, Hongjun Yoon, and Yosep Chong. Current trends of artificial intelligence for colorectal cancer pathology image analysis: a systematic review. *Cancers*, 12(7), 2020.

[43] Yutong Wang, Xiaoyun He, Hui Nie, Jianhua Zhou, Pengfei Cao, and Chunlin Ou. Application of artificial intelligence to the diagnosis and therapy of colorectal cancer. *Am. J. Cancer Res.*, 10(11):3575–3598, 2020.

[44] Habil Kalkan, Marius Nap, Robert P. W. Duin, and Marco Loog. Automated colorectal cancer diagnosis for whole-slice histopathology. In Nicholas Ayache, Hervé Delingette, Polina Golland, and Kensaku Mori, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 550–557, 2012.

[45] Hiroshi Yoshida, Yoshiko Yamashita, Taichi Shimazu, Eric Cosatto, Tomoharu Kiyuna, Hirokazu Taniguchi, Shigeki Sekine, and Atsushi Ochiai. Automated histological classification of whole slide images of colorectal biopsy specimens. *Oncotarget*, 8(53):90719–90729, 2017.

[46] Eric Cosatto, Pierre-Francois Laquerre, Christopher Malon, Hans-Peter Graf, Akira Saito, Tomoharu Kiyuna, Atsushi Marugame, and Ken'ichi Kamijo. Automated gastric cancer diagnosis on H&E-stained sections; training a classifier on a large scale with multiple instance machine learning. In Metin N. Gurcan and Anant Madabhushi, editors, *Medical Imaging 2013: Digital Pathology*, volume 8676, pages 51–59. SPIE, 2013.

[47] Bruno Korbar, Andrea Olofson, Allen Miraflor, Catherine Nicka, Matthew Suriawinata, Lorenzo Torresani, Arief Suriawinata, and Saeed Hassanpour. Deep learning for classification of colorectal polyps on whole-slide images. *J. Pathol. Inf.*, 8(1), 2017.

[48] Osamu Iizuka, Fahdi Kanavati, Kei Kato, Michael Rambeau, Koji Arihiro, and Masayuki Tsuneki. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific Rep.*, 10, 2020.

[49] Zhigang Song, Chunkai Yu, Shuangmei Zou, Wenmiao Wang, Yong Huang, Xiaohui Ding, Jinhong Liu, Liwei Shao, Jing Yuan, Xiangnan Gou, Wei Jin, Zhanbo Wang, Xin Chen, Huang Chen, Cancheng Liu, Gang Xu, Zhuo Sun, Calvin Ku, Yongqiang Zhang, Xianghui Dong, Shuhao Wang, Wei Xu, Ning Lv, and Huaiyin Shi. Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists. *BMJ Open*, 10(9), 2020.

[50] Jason W. Wei, Arief A. Suriawinata, Louis J. Vaickus, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Naofumi Tomita, Behnaz Abdollahi, Adam S. Kim, Dale C. Snover, John A. Baron, Elizabeth L. Barry, and Saeed Hassanpour. Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. *JAMA Network Open*, 3(4), 2020.

[51] Lin. Xu, Blair. Walker, Peir-In. Liang, Yi. Tong, Cheng. Xu, Yu. Su, and Aly. Karsan. Colorectal cancer detection based on deep learning. *J. Pathol. Inf.*, 11(1), 2020.

[52] Kuan-Song Wang, Gang Yu, Chao Xu, Xiang-He Meng, Jianhua Zhou, Changli Zheng, Zhenghao Deng, Li Shang, Ruijie Liu, Shitong Su, et al. Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence. *BMC medicine*, 19(1):1–12, 2021.

[53] Gang Yu, Kai Sun, Chao Xu, Xing-Hua Shi, Chong Wu, Ting Xie, Run-Qi Meng, Xiang-He Meng, Kuan-Song Wang, Hong-Mei Xiao, et al. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. *Nature communications*, 12(1):1–13, 2021.

[54] Niccolò Marini, Sebastian Otálora, Francesco Ciompi, Gianmaria Silvello, Stefano Marchesin, Simona Vatrano, Genziana Buttafuoco, Manfredo Atzori, and Henning Müller. Multiscale task multiple instance learning for the classification of digital pathology images with global annotations. In Manfredo Atzori, Nikolay Burlutskiy, Francesco Ciompi, Zhang Li, Fayyaz Minhas, Henning Müller, Tingying Peng, Nasir Rajpoot, Ben Torben-Nielsen, Jeroen van der Laak, Mitko Veta, Yinyin Yuan, and Inti Zlobec, editors, *Proceedings of the MICCAI Workshop on Computational Pathology*, volume 156 of *Proceedings of Machine Learning Research*, pages 170–181. PMLR, 27 Sep 2021.

[55] Cowan Ho, Zitong Zhao, Xiu Fen Chen, Jan Sauer, Sahil Ajit Saraf, Rajasa Jialdasani, Kaveh Taghipour, Aneesh Sathe, Li-Yan Khor, Kiat-Hon Lim, et al. A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer. *Scientific Reports*, 12(1):1–9, 2022.

[56] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour. Looking under the hood: deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 821–827, 2017.

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[58] J. Weinstein et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[59] H. Yoon, J. Lee, J.E. Oh, H.R. Kim, S. Lee, H.J. Chang, and D.K. Sohn. Tumor identification in colorectal histology images using a convolutional neural network. *Journal of Digital Imaging*, 32(1):131–140, 2019.

[60] Patrik Sabol, Peter Sinčák, Pitoyo Hartono, Pavel Kočan, Zuzana Benetinová, Alžbeta Blichárová, Ludmila Verbóová, Erika Štammová, Antónia Sabolová-Fabianová, and Anna Jašková. Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. *Journal of Biomedical Informatics*, 109, 2020.

[61] E. W. Teh and G. W. Taylor. Learning with less data via weakly labeled patch classification in digital pathology. In *IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 471–475, 2020.

[62] Elene Firmeza Ohata, João Victor Souza das Chagas, Gabriel Maia Bezerra, Mohammad Mehedi Hassan, Victor Hugo Costa de Albuquerque, and Pedro Pedrosa Rebouças Filho. A novel transfer learning approach for the classification of histological images of colorectal cancer. *The Journal of Supercomputing*, 2021.

[63] Sang-Hyun Kim, Hyun Min Koh, and Byoung-Dai Lee. Classification of colorectal cancer in histological images using deep neural networks: an investigation. *Multimedia Tools and Applications*, 2021.

[64] Y. Xu, Zhipeng Jia, L. Wang, Yuqing Ai, F. Zhang, Maode Lai, and E. Chang. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics*, 18, 2017.

[65] Kai Yang, Bi Zhou, Fei Yi, Yan Chen, and Yingsheng Chen. Colorectal cancer diagnostic algorithm based on sub-patch weight color histogram in combination of improved least squares support vector machine for pathological image. *Journal of Medical Systems*, 43(9), 2019.

[66] Matheus Gonçalves Ribeiro, Leandro Alves Neves, Marcelo Zanchetta do Nascimento, Guilherme Freire Roberto, Alessandro Santana Martins, and Thaína Aparecida Azevedo Tosta. Classification of colorectal cancer based on the association of multidimensional and multiresolution features. *Expert Systems with Applications*, 120:262–278, 2019.

[67] Hawraa Haj-Hassan, Ahmad Chaddad, Youssef Harkouss, Christian Desrosiers, Matthew Toews, and Camel Tanougast. Classifications of multispectral colorectal cancer tissues using convolution neural network. *J. Pathol. Inf.*, 8(1):1, 2017.

[68] Francesco Ponzio, Enrico Macii, Elisa Ficarra, and Santa Di Cataldo. Colorectal cancer classification using deep convolutional networks - an experimental study. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOIMAGING)*, pages 58–66, 2018.

[69] University of Leeds. Virtual pathology at the University of Leeds. `http://www.virtualpathology.leeds.ac.uk/`, 2018.

[70] Paola Sena, Rita Fioresi, Francesco Faglioni, Lorena Losi, Giovanni Faglioni, and Luca Roncucci. Deep learning techniques for detecting preneoplastic and neoplastic lesions in human colorectal histological images. *Oncology Lett.*, 18(6):6101–6107, 2019.

[71] K. Sirinukunwattana, D. R. J. Snead, and N. M. Rajpoot. A stochastic polygons model for glandular structures in colon histology images. *IEEE Transactions on Medical Imaging*, 34(11):2366–2378, 2015.

[72] Philipp Kainz, M. Pfeiffer, and M. Urschler. Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization. *PeerJ*, 5, 2017.

[73] Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 388–398, 2019.

[74] Muhammad Shaban, Ruqayya Awan, Muhammad Moazam Fraz, Ayesha Azam, Yee-Wah Tsang, David Snead, and Nasir M Rajpoot. Context-aware convolutional neural network for grading of colorectal cancer histology images. *IEEE Transactions on Medical Imaging*, 39(7):2395–2405, 2020.

[75] Ming Y. Lu, Richard J. Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding, 2019.

[76] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.*, 25(8):1301–1309, 2019.

[77] Sara P. Oliveira, João Ribeiro Pinto, Tiago Gonçalves, Rita Canas-Marques, Maria-João Cardoso, Hélder P. Oliveira, and Jaime S. Cardoso. Weakly-supervised classification of HER2 expression in breast cancer haematoxylin and eosin stained slides. *Applied Sciences*, 10(14):4728, 2020.

[78] Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

[79] Pathcore. Sedeen viewer. `https://pathcore.com/sedeen`, 2020.

[80] S. J. Winawer, M. J. O'Brien, J. D. Waye, O. Kronborg, J. Bond, P. Frühmorgen, L. H. Sobin, R. Burt, A. Zauber, and B. Morson. Risk and surveillance of individuals with colorectal polyps. WHO collaborating centre for the prevention of colorectal cancer. *Bull World Health Organ.*, 68(6):789–795, 1990.

[81] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022.

[82] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, 40(2):244–252, 2016.

[83] Wilson Silva, Kelwin Fernandes, Maria J. Cardoso, and Jaime S. Cardoso. Towards complementary explanations using deep neural networks. In Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne Martel, Lena Maier-Hein, Andre F. Marquand, Edouard Duchesnay, Tommy Löfstedt, Bennett Landman, M. Jorge Cardoso, Carlos A. Silva, Sergio Pereira, and Raphael Meier, editors, *Understanding and Interpreting Machine Learning in Medical Image Computing Applications (IMIMIC)*, pages 133–140. Springer, 2018.

[84] Milda Pocevičiūtė, Gabriel Eilertsen, and Claes Lundström. Survey of XAI in digital pathology. In Andreas Holzinger, Randy Goebel, Michael Mengel, and Heimo Müller, editors, *Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges*, pages 56–88. Springer, 2020.

[85] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[86] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[87] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[88] D Montezuma, J Fraga, S Oliveira, P Neto, A Monteiro, and I Macedo Pinto. Annotation in digital pathology: how to get started? our experience in classification tasks in pathology. In *VIRCHOWS ARCHIV*, volume 479, pages S320–S320, 2021.

[89] S. Kirk, Y. Lee, C. A. Sadow, S. Levine, C. Roche, E. Bonaccio, and J. Filiippini. Radiology data from the cancer genome atlas colon adenocarcinoma [TCGA-COAD] collection., 2016.

[90] S. Kirk, Y. Lee, C. A. Sadow, and S. Levine. Radiology data from the cancer genome atlas rectum adenocarcinoma [TCGA-READ] collection., 2016.

[91] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and F. Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26:1045–1057, 2013.

[92] Pathology AI Platform. Paip, 2020. `http://www.wisepaip.org`, last accessed on 20/01/22.

[93] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[94] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[95] J Ferlay, M Ervik, F Lam, M Colombet, L Mery, M Piñeros, A Znaor, I Soerjomataram, and F Bray. Global cancer observatory: Cancer today. https://gco.iarc.fr, 2020. Accessed: 2022-09-19.

[96] WHO Classification of Tumours Editorial Board. *Female Genital Tumours*. Medicine Series. International Agency for Research on Cancer, 2020.

[97] Maxime Bonjour, Hadrien Charvat, Eduardo L Franco, Marion Piñeros, Gary M Clifford, Freddie Bray, and Iacopo Baussano. Global estimates of expected and preventable cervical cancers among girls born between 2005 and 2014: a birth cohort analysis. *The Lancet Public Health*, 6(7):e510–e521, 2021.

[98] Mark H Stoler, Mark Schiffman, et al. Interobserver reproducibility of cervical cytologic and histologic interpretations: realistic estimates from the ascus-lsil triage study. *Jama*, 285(11):1500–1505, 2001.

[99] Mary T Galgano, Philip E Castle, Kristen A Atkins, William K Brix, Sarah R Nassau, and Mark H Stoler. Using biomarkers as objective standards in the diagnosis of cervical biopsies. *The American journal of surgical pathology*, 34(8):1077, 2010.

[100] Xin Hou, Guangyang Shen, Liqiang Zhou, Yinuo Li, Tian Wang, and Xiangyi Ma. Artificial intelligence in cervical cancer screening and diagnosis. *Frontiers in Oncology*, 12, 2022.

[101] Ching-Wei Wang, Yi-An Liou, Yi-Jia Lin, Cheng-Chang Chang, Pei-Hsuan Chu, Yu-Ching Lee, Chih-Hung Wang, and Tai-Kuang Chao. Artificial intelligence-assisted fast screening cervical high grade squamous intraepithelial lesion and squamous cell carcinoma diagnosis and treatment planning. *Scientific Reports*, 11(1):1–14, 2021.

[102] Liron Pantanowitz and Marilyn M Bui. Computer-assisted pap test screening. *Modern Techniques in Cytopathology*, 25:67–74, 2020.

[103] Lawrence von Karsa, Marc Arbyn, Hugo De Vuyst, Joakim Dillner, Lena Dillner, Silvia Franceschi, Julietta Patnick, Guglielmo Ronco, Nereo Segnan, Eero Suonio, et al. European guidelines for quality assurance in cervical cancer screening. summary of the supplements on hpv screening and vaccination. *Papillomavirus Research*, 1:22–31, 2015.

[104] Susan J Curry, Alex H Krist, Douglas K Owens, Michael J Barry, Aaron B Caughey, Karina W Davidson, Chyke A Doubeni, John W Epling, Alex R Kemper, Martha Kubik, et al. Screening for cervical cancer: Us preventive services task force recommendation statement. *Jama*, 320(7):674–686, 2018.

[105] Elizabeth TH Fontham, Andrew MD Wolf, Timothy R Church, Ruth Etzioni, Christopher R Flowers, Abbe Herzig, Carmen E Guerra, Kevin C Oeffinger, Ya-Chen Tina Shih, Louise C Walter, et al. Cervical cancer screening for individuals at average risk: 2020 guideline update from the american cancer society. *CA: a cancer journal for clinicians*, 70(5):321–346, 2020.

[106] Chen Li, Hao Chen, Xiaoyan Li, Ning Xu, Zhijie Hu, Dan Xue, Shouliang Qi, He Ma, Le Zhang, and Hongzan Sun. A review for cervical histopathology image analysis using machine vision approaches. *Artificial Intelligence Review*, 53(7):4821–4862, 2020.

[107] C. Li, D. Xue, X. Zhou, J. Zhang, H. Zhang, Y. Yao, F. Kong, L. Zhang, and H. Sun. Transfer learning based classification of cervical cancer immunohistochemistry images. In *Proceedings of the Third International Symposium on Image Computing and Digital*

Medicine, ISICDM 2019, page 102–106, New York, NY, USA, 2019. Association for Computing Machinery.

[108] Chen Li, Hao Chen, Dan Xue, Zhijie Hu, Le Zhang, Liangzi He, Ning Xu, Shouliang Qi, He Ma, and Hongzan Sun. Weakly supervised cervical histopathological image classification using multilayer hidden conditional random fields. In *International Conference on Information Technologies in Biomedicine*, pages 209–221. Springer, 2019.

[109] Yuan Xue, Qianying Zhou, Jiarong Ye, L Rodney Long, Sameer Antani, Carl Cornwell, Zhiyun Xue, and Xiaolei Huang. Synthetic augmentation and feature-based filtering for improved cervical histopathology image classification. In *International conference on medical image computing and computer-assisted intervention*, pages 387–396. Springer, 2019.

[110] Sudhir Sornapudi, R Joe Stanley, William V Stoecker, Rodney Long, Zhiyun Xue, Rosemary Zuna, Shelliane R Frazier, and Sameer Antani. Feature based sequential classifier with attention mechanism. *arXiv preprint arXiv:2007.11392*, 2020.

[111] Pan Huang, Shuailei Zhang, Min Li, Jing Wang, Cailing Ma, Bowei Wang, and Xiaoyi Lv. Classification of cervical biopsy images based on lasso and el-svm. *IEEE Access*, 8:24219–24228, 2020.

[112] Yuan Xue, Jiarong Ye, Qianying Zhou, L Rodney Long, Sameer Antani, Zhiyun Xue, Carl Cornwell, Richard Zaino, Keith C Cheng, and Xiaolei Huang. Selective synthetic augmentation with histogan for improved histopathology image classification. *Medical image analysis*, 67:101816, 2021.

[113] Abdulkadir Albayrak, Asli Unlu Akhan, Nurullah Calik, Abdulkerim Capar, Gokhan Bilgin, Behcet Ugur Toreyin, Bahar Muezzinoglu, Ilknur Turkmen, and Lutfiye Durak-Ata. A whole-slide image grading benchmark and tissue classification for cervical cancer precursor lesions with inter-observer variability. *Medical & Biological Engineering & Computing*, 59(7):1545–1561, 2021.

[114] Bum-Joo Cho, Jeong-Won Kim, Jungkap Park, Gui-Young Kwon, Mineui Hong, Si-Hyong Jang, Heejin Bang, Gilhyang Kim, and Sung-Taek Park. Automated diagnosis of cervical intraepithelial neoplasia in histology images via deep learning. *Diagnostics*, 12(2):548, 2022.

[115] Lidiya Wubshet Habtemariam, Elbetel Taye Zewde, and Gizeaddis Lamesgin Simegn. Cervix type and cervical cancer classification system using deep learning techniques. *Medical Devices (Auckland, NZ)*, 15:163, 2022.

[116] Sudhir Sornapudi, Ravitej Addanki, R Joe Stanley, William V Stoecker, Rodney Long, Rosemary Zuna, Shellaine R Frazier, and Sameer Antani. Automated cervical digitized histology whole-slide image analysis toolbox. *Journal of Pathology Informatics*, 12(1):26, 2021.

[117] Murat Gultekin, Pedro T Ramirez, Nathalie Broutet, and Raymond Hutubessy. World health organization call for action to eliminate cervical cancer globally. *International Journal of Gynecological Cancer*, 30(4):426–427, 2020.

[118] World Health Organization. Global strategy to accelerate the elimination of cervical cancer as a public health problem and its associated goals and targets for the period 2020–2030., 2020.

[119] Natalia M Rodriguez. Participatory innovation for human papillomavirus screening to accelerate the elimination of cervical cancer. *The Lancet Global Health*, 9(5):e582–e583, 2021.

[120] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015.

[121] Sara P Oliveira, Pedro C Neto, João Fraga, Diana Montezuma, Ana Monteiro, João Monteiro, Liliana Ribeiro, Sofia Gonçalves, Isabel M Pinto, and Jaime S Cardoso. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Scientific Reports*, 11(1):14358, 2021.

[122] Pedro C Neto, Sara P Oliveira, Diana Montezuma, João Fraga, Ana Monteiro, Liliana Ribeiro, Sofia Gonçalves, Isabel M Pinto, and Jaime S Cardoso. iMIL4PATH: A Semi-Supervised Interpretable Approach for Colorectal Whole-Slide Images. *Cancers*, 14(10):2489, 2022.

[123] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[124] World Health Organization. Breast cancer, 18-02-2019.

[125] World Cancer Research Fund. Breast cancer, 18-02-2019.

[126] American Cancer Society. Breast cancer facts & figures 2017-2018, 18-02-2019.

[127] G. Viale. The current state of breast cancer classification. *Annals of Oncology*, 23(suppl_10):x207–x210, 09 2012.

[128] Xiaofeng Dai, Ting Li, Zhonghu Bai, Yankun Yang, Xiuxia Liu, Jinling Zhan, and Bozhi Shi. Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research*, 5:2929–43, 12 2015.

[129] The American Cancer Society. Breast cancer, 19-02-2019.

[130] Kimberly H. Allison. Molecular pathology of breast cancer: What a pathologist needs to know. *American Journal of Clinical Pathology*, 138(6):770–780, 2012.

[131] Breastcancer.org. What is breast cancer?, 19-02-2019.

[132] Jorge S Reis-Filho and Lajos Pusztai. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet*, 378(9805):1812–1823, 2011.

[133] American Society of Clinical Oncology (ASCO). Breast Cancer Guide. `https://www.cancer.net/cancer-types/breast-cancer/introduction`, 2005-2020. [Online: accessed on 28.01.2020].

[134] Emad A Rakha, Sarah E Pinder, John M S Bartlett, Merdol Ibrahim, Jane Starczynski, Pauline J Carder, Elena Provenzano, Andrew Hanby, Sally Hales, Andrew H S Lee, and Ian O Ellis. Updated UK Recommendations for HER2 assessment in breast cancer. *Journal of Clinical Pathology*, 68(2):93–99, 2015.

[135] Katrina Goddard, Sheila Weinmann, Kathryn Richert-Boe, Chuhe Chen, Joanna Bulkley, and C Wax. HER2 Evaluation and Its Impact on Breast Cancer Treatment Decisions. *Public Health Genomics*, 15:1–10, 2011.

[136] Wedad M Hanna, Penny J. Barnes, Martin Chang, C Blake Gilks, Anthony M. Magliocco, Henrike Rees, Louise Quenneville, Susan J. Robertson, Sandip K Sengupta, and Sharon Nofech-Mozes. Human epidermal growth factor receptor 2 testing in primary breast cancer in the era of standardized testing: a canadian prospective study. *Journal of Clinical Oncology*, 32(35):3967–73, 2014.

[137] Antonio C. Wolff, M. Elizabeth Hale Hammond, Kimberly H. Allison, Brittany E. Harvey, Pamela B. Mangu, John M.S. Bartlett, Michael Bilous, Ian O. Ellis, Patrick Fitzgibbons, Wedad Hanna, Robert B. Jenkins, Michael F. Press, Patricia A. Spears, Gail H. Vance, Giuseppe Viale, Lisa M. McShane, and Mitchell Dowsett. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *Journal of Clinical Oncology*, 36(20):2105–2122, 2018.

[138] Haibo Wang, Angel Cruz-Roa, Ajay Basavanhally, Hannah Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio González, and Anant Madabhushi. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1:1–8, 12 2014.

[139] Chao Li, Xinggang Wang, Wenyu Liu, and Longin Jan Latecki. Deepmitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical image analysis*, 45:121–133, 2018.

[140] Anabia Sohail, Asifullah Khan, Humaira Nisar, Sobia Tabassum, and Aneela Zameer. Mitotic nuclei analysis in breast cancer histopathology images using deep ensemble classifier. *Medical image analysis*, 72:102121, 2021.

[141] M. M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, and M. N. Gurcan. Computerized classification of intraductal breast lesions using histopathological images. *IEEE Transactions on Biomedical Engineering*, 58(7):1977–1984, 2011.

[142] Xingyu Li, Marko Radulovic, Ksenija Kanjer, and Konstantinos N Plataniotis. Discriminative pattern mining for breast cancer histopathology image classification via fully convolutional autoencoder. *arXiv preprint arXiv:1902.08670*, 2019.

[143] Francisco Perdigón Romero, An Tang, and Samuel Kadoury. Multi-level batch normalization in deep networks for invasive ductal carcinoma cell discrimination in histopathology images. *arXiv preprint arXiv:1901.03684*, 2019.

[144] Suzanne C Wetstein, Vincent MT de Jong, Nikolas Stathonikos, Mark Opdam, Gwen MHE Dackus, Josien PW Pluim, Paul J van Diest, and Mitko Veta. Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. *Scientific reports*, 12(1):1–12, 2022.

[145] Y. Wang, B. Acs, S. Robertson, B. Liu, L. Solorzano, C. Wählby, J. Hartman, and M. Rantalainen. Improved breast cancer histological grading using deep learning. *Annals of Oncology*, 33(1):89–98, 2022.

[146] Hongliu Cao, Simon Bernard, Laurent Heutte, and Robert Sabourin. Improve the performance of transfer learning without fine-tuning using dissimilarity-based multi-view learning for breast cancer histology images. In Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny, editors, *Image Analysis and Recognition*, pages 779–787. Springer International Publishing, 2018.

[147] Andrew Lagree, Audrey Shiner, Marie Angeli Alera, Lauren Fleshner, Ethan Law, Brianna Law, Fang-I Lu, David Dodington, Sonal Gandhi, Elzbieta A Slodkowska, et al. Assessment of digital pathology imaging biomarkers associated with breast cancer histologic grade. *Current Oncology*, 28(6):4298–4316, 2021.

[148] Eduardo Conde-Sousa, João Vale, Ming Feng, Kele Xu, Yin Wang, Vincenzo Della Mea, David La Barbera, Ehsan Montahaei, Mahdieh Baghshah, Andreas Turzynski, et al. Herohe challenge: Predicting her2 status in breast cancer from hematoxylin–eosin whole-slide imaging. *Journal of Imaging*, 8(8):213, 2022.

[149] Gil Shamai, Yoav Binenbaum, Ron Slossberg, Irit Duek, Ziv Gil, and Ron Kimmel. Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer. *JAMA network open*, 2(7):e197700–e197700, 2019.

[150] Rishi R Rawat, Itzel Ortega, Preeyam Roy, Fei Sha, Darryl Shibata, Daniel Ruderman, and David B Agus. Deep learned tissue "fingerprints" classify breast cancers by er/pr/her2 status from h&e images. *Scientific reports*, 10(1):1–13, 2020.

[151] David La Barbera, António Polónia, Kevin Roitero, Eduardo Conde-Sousa, and Vincenzo Della Mea. Detection of her2 from haematoxylin-eosin slides through a cascade of deep learning classifiers via multi-instance learning. *Journal of Imaging*, 6(9):82, 2020.

[152] Paul Gamble, Ronnachai Jaroensri, Hongwu Wang, Fraser Tan, Melissa Moran, Trissia Brown, Isabelle Flament-Auvigne, Emad A Rakha, Michael Toss, David J Dabbs, et al. Determining breast cancer biomarker status and associated morphological features using deep learning. *Communications medicine*, 1(1):1–12, 2021.

[153] Dmitrii Bychkov, Nina Linder, Aleksei Tiulpin, Hakan Kücükel, Mikael Lundin, Stig Nordling, Harri Sihto, Jorma Isola, Tiina Lehtimäki, Pirkko-Liisa Kellokumpu-Lehtinen, et al. Deep learning identifies morphological features in breast cancer predictive of cancer erbb2 status and trastuzumab treatment efficacy. *Scientific reports*, 11(1):1–10, 2021.

[154] Saman Farahmand, Aileen I Fernandez, Fahad Shabbir Ahmed, David L Rimm, Jeffrey H Chuang, Emily Reisenbichler, and Kourosh Zarringhalam. Deep learning trained on hematoxylin and eosin tumour region of interest predicts her2 status and trastuzumab treatment response in her2+ breast cancer. *Modern Pathology*, 35(1):44–51, 2022.

[155] Wenqi Lu, Michael Toss, Muhammad Dawood, Emad Rakha, Nasir Rajpoot, and Fayyaz Minhas. Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer. *Medical Image Analysis*, page 102486, 2022.

[156] Talha Qaiser, Abhik Mukherjee, Chaitanya Reddy PB, Sai D Munugoti, Vamsi Tallam, Tomi Pitkäaho, Taina Lehtimäki, Thomas Naughton, Matt Berseth, Aníbal Pedraza, Ramakrishnan Mukundan, Matthew Smith, Abhir Bhalerao, Erik Rodner, Marcel Simon, Joachim Denzler, Chao-Hui Huang, Gloria Bueno, David Snead, Ian O Ellis, Mohammad Ilyas, and Nasir Rajpoot. HER2 challenge contest: a detailed assessment of automated HER2 scoring

algorithms in whole slide images of breast cancer tissues. *Histopathology*, 72(2):227–238, 2018.

[157] Wilma Lingle, Bradley J. Erickson, Margarita L. Zuley, Rose Jarosz, Ermelinda Bonaccio, Joe Filippini, Jose M. Net, Len Levi, Elizabeth A. Morris, Gloria G. Figler, Pierre Elnajjar, Shanah Kirk, Yueh Lee, Maryellen Giger, and Nicholas Gruszauskas. Radiology data from the cancer genome atlas breast invasive carcinoma [tcga-brca] collection, 2016.

[158] Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, Anant Madabhushi, and Fabio González. High-throughput adaptive sampling for whole-slide histopathology image analysis (hashi) via convolutional neural networks: Application to invasive breast cancer detection. *PLOS ONE*, 13(5):1–23, 2018.

[159] Angel Cruz-Roa, Hannah L. Gilmore, Ajay Basavanhally, Michael D. Feldman, Shridar Ganesan, N. C. Shih, John P. Tomaszewski, Fabio A. Gonzalez, and Anant Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific reports*, 7(46450), 2017.

[160] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2018.

[161] Olivier Chapelle and Mingrui Wu. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 13(3):216–235, 2010.

[162] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer*, 124(4):686–696, 2021.

This work was financed by the Portuguese science and technology foundation, Fundação para a Ciência e a Tecnologia (FCT) and co-financed by the European Social Fund through the North Regional Operational Programme (NORTE 2020), under the grant SFRH/BD/139108/2018.