UK Research and Innovation

UNIVERSITY OF LEEDS

Artificial Intelligence for Ovarian Cancer Diagnosis from Digital Pathology Slides

Jack Breen

A thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy in the

School of Computing

The University of Leeds

February 2025

Declaration of Authorship

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and the other authors' contributions to this work have been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

- Chapter 3 is based on the 2023 published paper *Artificial Intelligence in Ovarian Cancer Histopathology: A Systematic Review* in NPJ Precision Oncology by J. Breen, K. Allen, K. Zucker, P. Adusumilli, A. Scarsbrook, G. Hall, N. M. Orsi, and N. Ravikumar.

- Chapter 5 is based on the 2023 published paper *Efficient Subtyping of Ovarian Cancer Histopathology Whole Slide Images using Active Sampling in Multiple Instance Learning.* in Proceedings of SPIE 12471 by J. Breen, K. Allen, K. Zucker, G. Hall, N. M. Orsi, and N. Ravikumar.

- Chapter 6 is based on the 2024 published paper *Reducing Histopathology Slide Magnification Improves the Accuracy and Speed of Ovarian Cancer Subtyping* in 2024 IEEE International Symposium on Biomedical Imaging (ISBI) by J. Breen, N. Ravikumar, K. Allen, K. Zucker and N. M. Orsi.

- Chapter 7 is based on the 2025 published paper *A Comprehensive Evaluation of Histopathology Foundation Models for Ovarian Cancer Subtype Classification* in NPJ Precision Oncology by J. Breen, K. Allen, K. Zucker, L. Godson, N. M. Orsi, and N. Ravikumar.

- Chapter 8 is based on the recently accepted paper *Multi-Resolution Histopathology Patch Graphs for Ovarian Cancer Subtyping* in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) by J. Breen, K. Allen, K. Zucker, N. M. Orsi, and N. Ravikumar.

- Chapters 1 and 4 are partially based on the aforementioned publications, as well as the published book chapter *Generative Adversarial Networks for Stain Normalisation in Histopathology* in Applications of Generative AI by J. Breen, K. Zucker, K. Allen, N. Ravikumar, and N. M. Orsi.

For each listed publication, I was the primary author, conducted the primary research, wrote the manuscript, and generated all tables/figures. Other authors created the internal datasets, aided in project design, and provided feedback on the manuscripts. For the systematic review paper used in Chapter 3, other authors also screened papers and performed some of the risk of bias assessments and data extraction.

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

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

© 2025 The University of Leeds and Jack Breen

List of Publications

This thesis was primarily based on two published journal papers and three conference papers [1, 2, 5, 6, 7], with additional publications listed for completeness.

Journal Papers

- [1] **J. Breen**, K. Allen, K. Zucker, P. Adusumilli, A. Scarsbrook, G. Hall, N. M. Orsi, and N. Ravikumar (2023). Artificial Intelligence in Ovarian Cancer Histopathology: A Systematic Review. NPJ Precision Oncology, 7(1), 83.
- [2] **J. Breen**, K. Allen, K. Zucker, L. Godson, N. M Orsi, N. Ravikumar (2025). A Comprehensive Evaluation of Histopathology Foundation Models for Ovarian Cancer Subtype Classification. NPJ Precision Oncology, 9, 33.
- [3] M. Aubreville, N. Stathonikos, C. A. Bertram, R. Klopfleisch, N. Ter Hoeve, F. Ciompi, ... **J. Breen**, N. Ravikumar, ... and K. Breininger (2023). Mitosis Domain Generalization in Histopathology Images—the MIDOG Challenge. Medical Image Analysis, 84, 102699.

Conference Papers

- [4] **J. Breen**, K. Zucker, N. M. Orsi, and N. Ravikumar (2021). Assessing Domain Adaptation Techniques for Mitosis Detection in Multi-Scanner Breast Cancer Histopathology Images. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 14-22). Cham: Springer International Publishing.
- [5] **J. Breen**, K. Allen, K. Zucker, G. Hall, N. M. Orsi, and N. Ravikumar (2023). Efficient Subtyping of Ovarian Cancer Histopathology Whole Slide Images using Active Sampling in Multiple Instance Learning. In Proceedings of SPIE 12471 (Vol. 12471). SPIE.
- [6] **J. Breen**, N. Ravikumar, K. Allen, K. Zucker and N. M. Orsi (2024). Reducing Histopathology Slide Magnification Improves the Accuracy and

Speed of Ovarian Cancer Subtyping. In 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece.

- [7] **J. Breen**, K. Allen, K. Zucker, G. Hall, N. M. Orsi, and N. Ravikumar (2024). Multi-Resolution Histopathology Patch Graphs for Ovarian Cancer Subtyping. Accepted for publication through MICCAI GRAIL Workshop 2024.

Conference Abstracts

- [8] K. Allen, **J. Breen**, G. Hall, K. Zucker, N. Ravikumar, and N, M. Orsi (2023). Comparative Evaluation of Ovarian Carcinoma Subtyping in Primary versus Interval Debulking Surgery Specimen Whole Slide Images using Artificial Intelligence. International Journal of Gynecologic Cancer. 2023;33(Suppl 3):A429-30.

Book Chapters

- [9] **J. Breen**, K. Zucker, K. Allen, N. Ravikumar, and N. M. Orsi (2024). Generative Adversarial Networks for Stain Normalisation in Histopathology. In Applications of Generative AI (pp. 227-247). Cham: Springer International Publishing.

- [10] K. Allen, P. Adusumilli, **J. Breen**, G. Hall, and N. M. Orsi (2024). Artificial Intelligence in Ovarian Digital Pathology. In Pathology of the Ovary, Fallopian Tube and Peritoneum (pp. 731-749). Cham: Springer International Publishing.

Preprints

- [11] **J. Breen**, K. Allen, K. Zucker, G. Hall, N. Ravikumar, and N. M. Orsi (2023). Predicting Ovarian Cancer Treatment Response in Histopathology using Hierarchical Vision Transformers and Multiple Instance Learning. arXiv preprint arXiv:2310.12866.

- [12] M. Eisenmann, A. Reinke, V. Weru, M. D. Tizabi, F. Isensee, T.J. Adler, ... **J. Breen**, ... and Filipiak, P. (2022). Biomedical Image Analysis Competitions: The State of Current Participation Practice. arXiv preprint arXiv:2212.08568.

List of Awards

Turing Institute Enrichment **Community Award** - £1,500 award for academic costs and access to *The Alan Turing Institute* events and resources during 2022/23.

Turing-Roche **Community Scholar Award** - £3,000 award and access to *The Alan Turing Institute* and *Roche* events and resources during 2023/24 for contributing to the Turing Way research handbook.

Best Poster Award at the 2023 *Health Data Hub* AI4Health Summer School.

Best Presentation Award at the 2023 *University of Leeds* School of Computing postgraduate symposium.

Best Presentation Award at the 2024 UKRI AI CDTs in Healthcare Conference.

SPIE Medical Imaging 2023 **Conference Fee Waiver**.

Leeds Institute of Data Analytics (LIDA) **Conference Grant** for AIUK 2024.

Acknowledgements

This thesis is the culmination of years of support from many educators, academics, funders, friends and family, without whom it would have never even started.

To my supervisors, Nishant Ravikumar, Nicolas M. Orsi, Kieran Zucker, and Geoff Hall, and my PhD colleague Katie Allen, I express sincere gratitude for their invaluable support throughout the process. Their expertise, effort, and kindness have not gone unnoticed, and their mentorship has made the PhD a genuinely enjoyable and enriching experience.

This PhD has been a large investment by the British taxpayer and would not have been possible without the Tony Bramall Charitable Trust supporting concurrent ovarian cancer research. Further, hundreds of patients allowed their data to be used in this research despite knowing that this would not provide any direct benefit to themselves. I sincerely hope that I can realise the value of these investments.

Being a member of the Leeds Medical AI CDT has been a pleasure, and I would like to thank the CDT directors, David Hogg, Vania Dimitrova, and Owen Johnson, for their tireless work in establishing and maintaining such a flourishing research community. I would also like to thank the wider CDT management team, particularly Hien Nguyen and Sunjeev Ghir, for their ongoing support. Further, I would like to thank all of the CDT students for their technical input and help in generating research ideas and, most of all, for making a city I had never previously visited feel like home.

I would also like to acknowledge the support of my friends. To those I met in Nottingham, I am grateful that we have remained close despite now being dispersed around the country. To those in Leeds, it is such a shame to now be leaving, but I hope we can continue in the same way.

Finally, I would like to thank my family. Chris, your achievements have been a constant source of inspiration, and I am grateful to have shared your hobbies and interests. Charlie, I am deeply thankful for you, you have always been my favourite. I am confident you will achieve whatever you set your mind to. Mum, thank you for everything, you have always supported me, even at your own sacrifice.

Rachel, I have been so fortunate to have your love and support throughout this process, despite the distance between us we have grown closer than ever. I love you so much and I cannot wait to start the next chapter of our lives together.

Abstract

Digital pathology is a rapidly growing field, allowing for the development of assistive diagnostic tools. Many tools use artificial intelligence (AI) to automatically provide insights from whole slide images (WSIs), aiming to improve the accuracy, objectivity, and efficiency of the diagnostic process. Research has typically focused on the most common cancers, but less common cancers have received comparatively little attention. We focus on the histological subtyping of ovarian cancer, an essential diagnostic task for determining optimal treatments and prognoses. Through a systematic literature review, we find that previous research has been limited to model prototyping with small homogeneous datasets, with little focus on clinical utility. We perform the most thorough analyses of automated ovarian cancer histological subtyping to date, using the largest training dataset and evaluating models through cross-validation, hold-out testing, external validations, bootstrapping, and hypothesis testing. Analyses are based on attention-based multiple instance learning (ABMIL) with an ImageNet-pretrained ResNet50 backbone, a commonly used WSI classifier. The computational complexity of current AI models is a key limitation, with pathology labs typically not having sufficient hardware for model deployment. We propose an active tissue sampling technique and show that this approach can drastically reduce the computational burden of inference with minimal impact on diagnostic performance. ABMIL analyses tissue at only a single magnification, with high magnifications offering more cellular detail and low magnifications providing broader tissue context. We find that 10x magnification balances the cellular and histoarchitectural details to give the most accurate ovarian cancer subtyping performance, while drastically reducing the computational burden compared to the clinical standard 40x magnification. Recently, histopathology foundation models have promised to revolutionise diagnostic AI. We analyse 14 foundation models and confirm that they give significantly greater performance than previous feature extractors. In ABMIL, tissue patches are treated as independent of each other. We propose a multi-resolution patch graph network to better model spatial context and find this marginally improves performance. The optimal model, a combination of a foundation model and a graph, achieved five-class balanced accuracies of 88%, 99%, and 77% in three validation sets, where our baseline model achieved only 66%, 69%, and 52%, and individual pathologists achieved 74-91% concordance with similarly determined labels. This gives us confidence that AI models could have clinical utility, so future work should focus on the practicalities of implementation and real-world validation.

Contents

List of Figures

List of Tables

List of Abbreviations

ABMIL	Attention-based Multiple Instance Learning
AI	Artificial Intelligence
ATEC23	Automated Prediction Of Treatment Effectiveness In Ovarian Cancer Using Histopathological Images
AUROC	Area Under the Receiver Operating Characteristic Curve
CCC	Clear Cell Carcinoma
CLAM	Cluster-constrained Attention MIL
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DRAS-MIL	Discriminative Region Active Sampling For Multiple Instance Learning
EC	Endometrioid Carcinoma
FFPE	Formalin-fixed, Paraffin-embedded
GNN	Graph Neural Network
GPU	Graphics Processing Unit
H&E	Haematoxylin and Eosin
HGSC	High-grade Serous Carcinoma
HPC	High-performance Computer
IDS	Interval Debulking Surgery
IHC	Immunohistochemistry
LGSC	Low-grade Serous Carcinoma
LTHT	Leeds Teaching Hospitals NHS Trust
MC	Mucinous Carcinoma
MIL	Multiple Instance Learning
ML	Machine Learning
NLP	Natural Language Processing
OCEAN	Ovarian Cancer SubtypE ClAssification and Outlier DetectioN
PC	Personal Computer
PRISMA	Preferred Reporting Items For Systematic Reviews and Meta-Analyses
PROBAST	Prediction Model Risk Of Bias ASsessment Tool
RGB	Red, Green and Blue

SSL	Self-supervised Learning
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TMA	Tissue Microarray
TRIPOD	*Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis*
ViT	Vision Transformer
WSI	Whole Slide Image

Chapter 1

Introduction

Pathology is an increasingly digitised discipline, with tissue slides scanned at high magnification to allow diagnostic pathologists to analyse them using a computer rather than a microscope. Digital pathology offers an opportunity for the development of computer-aided diagnosis tools to improve the efficiency and accuracy of diagnostic workflows, which may help to mitigate the severe shortage of available pathologists as diagnostic workloads continue to grow rapidly. There is a wealth of research in this field for the most common cancers, such as breast and lung cancers, but much less for lower-incidence malignancies such as ovarian cancer. Addressing such disparities in research will be central to ensuring that any benefits from clinical artificial intelligence (AI) are not confined to a handful of common diseases.

Digital pathology images are vastly larger than the images analysed in typical AI models, making them particularly challenging to assess. Diagnostically relevant tissue may form a very small proportion of the entire tissue area, and missing such diagnoses may have severe consequences for the health of a patient. However, thoroughly assessing every pixel in such huge images presents computational challenges, both in the training of a model and in the inference of clinical samples. Therefore, it is pertinent to design models specifically for use on digital pathology data, with mechanisms to determine and focus on the most diagnostically relevant tissue, and with particular consideration for computational efficiency.

1.1 Aims and Objectives

The primary aim of this research is the development and thorough validation of AI techniques to classify ovarian cancer subtypes from digitised pathology whole slide images. It is hoped that such tools may offer significant clinical benefit in improving the efficiency and accuracy of an essential aspect of cancer diagnosis which is often time-consuming and resource-intensive. The specific objectives are to:

- Conduct a systematic literature review to characterise and quantify the risks of bias associated with all previous research investigating diagnostic & prognostic AI using ovarian cancer digital pathology slides.
- Apply state-of-the-art AI approaches from other pathologies to ovarian cancer subtyping using the largest ovarian cancer histopathology dataset to date.
- Build upon these previous approaches using novel classification techniques to boost the efficiency and discriminative ability of the models.
- Rigorously analyse optimal model configurations using hyperparameter tuning procedures and thoroughly validate classification performance using multiple validation datasets. Measure discriminative power through multiple metrics, assess model efficiency, and qualitatively investigate performance.

1.2 Thesis Structure

Following this introductory chapter are eight subsequent chapters. Chapter 2 provides a clinical overview of ovarian cancer and digital pathology, and a technical overview of relevant AI methods. Chapter 3 is an extensive review of published literature in the domain of AI for ovarian cancer diagnosis and prognosis. Chapter 4 describes the methodology applied throughout the thesis, including explanations of the standard models, validation methods, and datasets which are used in subsequent chapters. Chapters 5-8 are primary research chapters, with each focusing on a different aspect of the slide classification process. Chapter 5 explores approaches for active sampling for a more efficient slide classification. Chapter 6 is a thorough analysis of how the tissue magnification affects the efficiency and discriminative ability of a whole slide classifier. Chapter 7 is a thorough analysis of feature extraction techniques, including recently developed histopathology foundation models. Chapter 8 introduces a novel multi-resolution graph network for slide classification, building upon the lessons learned in previous chapters. Finally, Chapter 9 provides a summary of the thesis, including the limitations of the presented work and a view towards the future of research in this field.

These chapters are based on my primary author peer-reviewed articles published in journals and conference proceedings. Chapter 3 is based on a published systematic literature review [1] in *NPJ Precision Oncology*. Chapters 5, 6, and 8 are underpinned by conference papers in *SPIE Medical Imaging 2023* [5], *ISBI 2024* [6], and *MICCAI 2024* [7], respectively. The work underpinning Chapter 7 has recently been published in *NPJ Precision Oncology* [2]. Chapters 2 and 4 include work from each of these publications, as well as a published book chapter [9].

Chapter 2

Clinical and Technical Background

In this chapter, we provide a clinical background to ovarian cancer, pathological diagnostics, and in particular, ovarian cancer histological subtyping. We then describe the wider context of AI in digital pathology the modelling techniques used in the field.

2.1 Ovarian Cancer

In the female reproductive system, egg cells are produced in the ovaries and travel through the fallopian tubes to the uterus. The ovaries are situated in the lower (infracolic) compartment of the abdominal cavity, close to the lower digestive and urinary systems, a dense array of lymph nodes, and a large section of fatty tissue called the omentum. Ovarian cancer encompasses primary malignant tumours of the ovaries, fallopian tubes, and peritoneum (the inner lining of the abdominal cavity), with some research suggesting that these all originate in the fallopian tubes due to the findings of precursor lesions (serous tubal intraepithelial carcinoma) in patients who received prophylactic surgery (salpingo-oophorectomy) due to genetic risk factors [13].

Ovarian cancer is the eighth most common malignancy in women worldwide [14]. It is a notoriously difficult disease to detect due to the disease having vague symptoms similar to those caused by menopause [15], which is particularly problematic since ovarian cancer typically affects menopausal and post-menopausal women [16]. Furthermore, a randomised controlled trial with 200,000 participants found that screening based on ultrasound imaging and the blood biomarker CA125 did not improve early-stage diagnosis rates sufficiently to save lives [17].

Without effective screening, ovarian cancer is typically only detected once it has spread beyond the pelvis, giving a relatively poor prognosis. While overall survival trends have somewhat improved [18], ovarian cancer remains a particularly deadly disease. Worldwide, there are 324,000 new cases of ovarian cancer diagnosed each year, leading to 207,000 deaths [14]. In the UK, overall 1-year and 5-year survival rates are around 76% and 38%, respectively, with the length of survival depending upon the histological subtype and stage at diagnosis [19].

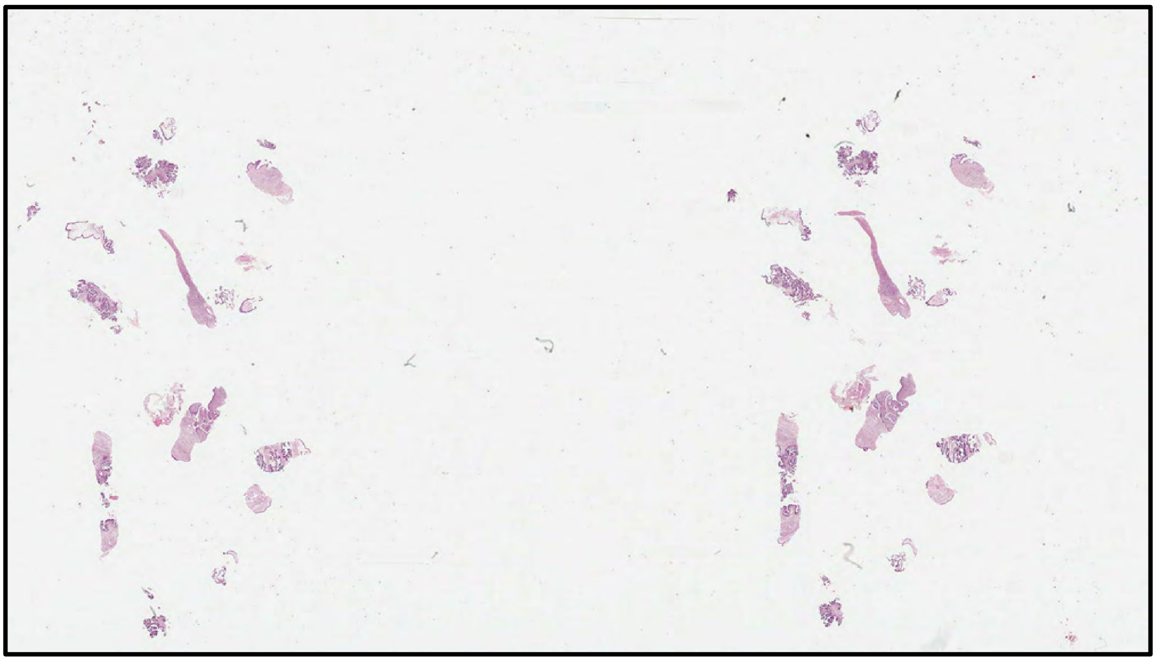


Figure 2.1 A digital pathology image containing ovarian biopsies.

In suspected cases, numerous tests may be used to confirm the presence of ovarian cancer, including diagnostic imaging (radiology), blood tests, and biopsies. Biopsies (Figure 2.1) contain very little tissue and are often only relied upon for confirming the presence of cancer, with deeper pathological analysis performed after resection surgery (Figure 2.2). In the initial surgery, resected tissue typically includes the ovaries, fallopian tubes, uterus, omentum, and local lymph nodes.

Pathological diagnosis includes classification of the stage, grade, and subtype of the cancer. International Federation of Gynecology and Obstetrics (FIGO) staging [20] is used to quantify the spread of primary ovarian cancer based on the primary tumour, local lymph nodes, and distant metastases. Stage I ovarian cancer is confined to the ovaries, stage II has spread within the pelvis, stage III has spread within the peritoneum or retroperitoneal lymph nodes, and stage IV has metastasized further away. Grading instead measures the abnormality of the cancer cells, which in turn represents how aggressively a cancer is likely to behave. This was historically a three-tier system, with grade 1 tumours containing well-differentiated cells (most similar to normal cells), grade 3 tumours containing poorly differentiated cells, and grade 2 falling somewhere

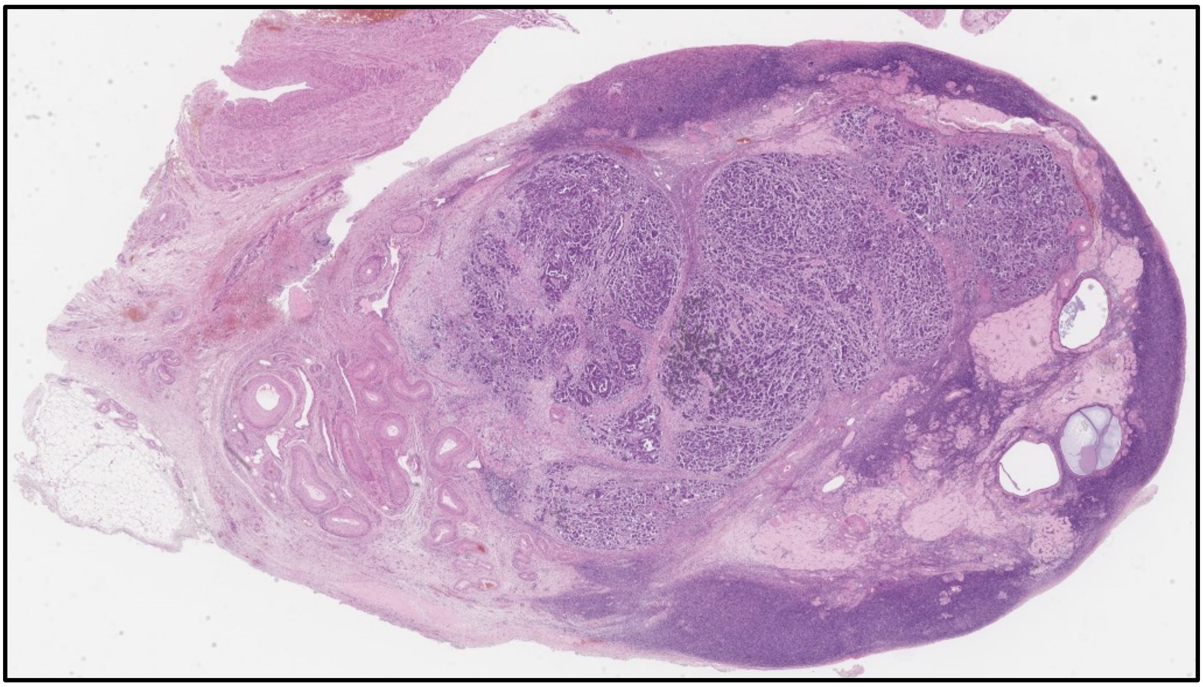


Figure 2.2 A digital pathology image containing a slice of an entire resected ovary from staging surgery. The upper left side shows connective tissue. The pale regions towards the right side are *corpus albicans*, the scars left after egg cells are released.

in the middle. This system is still used for some ovarian cancer histological subtypes, though *serous* ovarian cancers are instead categorised as being either high-grade or low-grade, and clear cell ovarian cancers are all categorised as high-grade.

Treatment decision-making is influenced by a range of factors, including radiological and pathological analysis, comorbidities, age, and the patient's personal decisions [21]. Most patients will undergo both surgery and chemotherapy, though in some cases, one of these will be sufficient alone, or treatment may not be administered due to other factors. These treatments are highly variable, with variations in the extent of surgery, the chemotherapy drugs used, and the number and timing of chemotherapy cycles [22]. Additional pharmacological treatments may include VEGF inhibitors and/or PARP inhibitors, which prevent blood vessel formation and DNA repair respectively. The effectiveness of the latter depends upon genetic factors [23–25]. Optimal treatment decisions require integrating data from a range of sources, with histological analysis being an essential component, without which patients may be subjected to ineffective treatments and have worse overall outcomes.

The primary focus of this thesis is the diagnosis of histological subtypes, which are the specific characteristics of the cancer determined by the cellular and histoarchitectural features present in pathology samples. Most ovarian cancers are carcinomas (cancers of epithelial origin), for which the World Health Organisation defines five main subtypes [26] - high-grade serous carcinoma (HGSC), endometrioid carcinoma (EC), clear cell carcinoma (CCC), low-grade serous carcinoma (LGSC), and mucinous carcinoma (MC) (Figure 2.3). HGSC is the most common form of ovarian cancer, accounting for approximately 70% of all cases [27]. Non-epithelial ovarian cancers account for less than 10% of all ovarian malignancies and include germ cell, sex cord-stromal, and mesenchymal tumours [28]. Histological subtypes are distinct in their genetics, prognoses, and treatment options [29, 30], making their classification an essential component of ovarian cancer diagnosis.

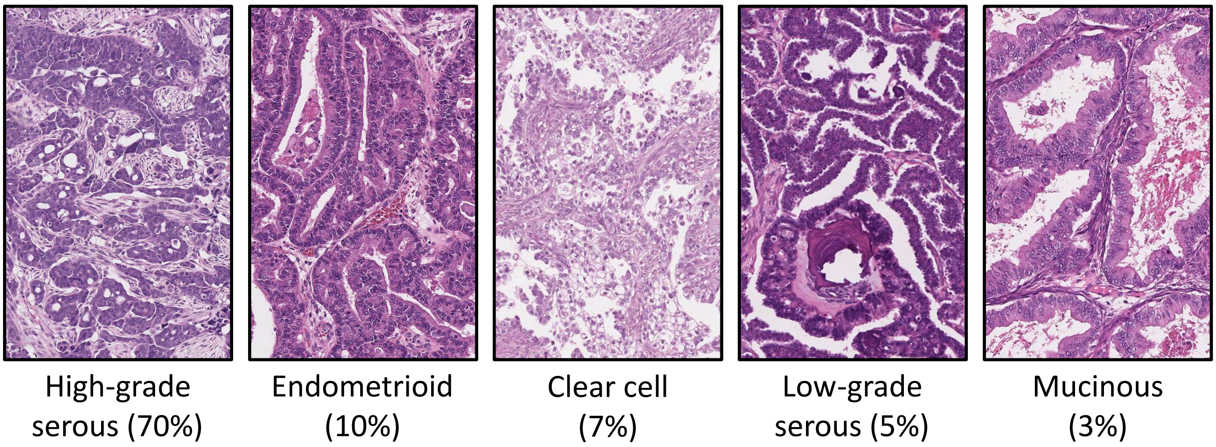


Figure 2.3 Examples of the five major morphological subtypes of epithelial ovarian cancer with corresponding frequencies [30]. These frequencies do not sum to 100% due to the existence of rarer subtypes, which are not shown here.

2.2 Histopathology

Histopathology is the microscopic evaluation of tissue for medical diagnosis. It is an essential part of the diagnostic pathway for many diseases, including autoimmune disorders, infections, and cancers. Tissue samples are taken either as small biopsies or larger tissue resections, and typically they are fixed in formalin, embedded in paraffin, sectioned, and stained. Formalin-fixed, paraffin-embedded (FFPE) samples are the diagnostic gold standard and are suitable for long-term storage at ambient temperatures. Some samples are instead flash-frozen, a faster process which allows pathologists to provide rapid information during surgery, but at the expense of increased cell damage and inferior staining quality. Pathologists typically interpret tissue stained with haematoxylin and eosin (H&E), where haematoxylin stains cell nuclei blue and eosin stains other cellular structures, such as cytoplasm and cell membranes, varying shades of pink and red.

Histological subtypes are diagnosed by pathologists assessing standard H&E-stained tissue samples for their varied morphologies. Important features include histological patterns and architecture of tumour cells (solid, papillary, glandular etc.), the frequency of typical and atypical mitotic figures (dividing cells), the degree of cellular and nuclear pleomorphism (variation in size and shape), the nuclear to cytoplasmic ratio, the colour and consistency of cytoplasm, and the presence or absence of necrosis.

Ideally, histological subtypes are diagnosed using primary surgery resection specimens, where the surgical removal of the tumour was the initial treatment. However, in many cases a patient will receive neoadjuvant chemotherapy to reduce the size of the tumour before surgery, with any resection surgery performed after chemotherapy (or after the primary surgery) referred to as interval debulking surgery (IDS). IDS samples are not typically considered appropriate for subtyping because of chemotherapy-induced morphological changes, such as varying amounts of cell death and associated changes in surrounding stroma. If it is not possible to analyse a primary surgery resection specimen the next-best option is a pre-treatment biopsy, with IDS samples only relied upon in cases where such a biopsy is not available.

The interpretation of H&E slides can be a subjective, time-consuming process, with some tasks having a high level of inter-observer variation [31–33]. From an individual ovarian carcinoma slide, pathologists only achieved a median 86% concordance rate with the central review subtype diagnosis, and individual pathologists varied between 74-91% [32]. In the assessment of difficult cases, generalist pathologists may seek assistance from subspecialty experts (such as gynaecological pathologists), and/or use ancillary tests, such as immunohistochemistry (IHC) staining. IHC stains indicate the presence of specific antigens to aid pathologists in identifying known phenotypic profiles, helping to distinguish primary tumour types or histological subtypes [29]. For example, p53 protein expression can be profiled, with abnormalities suggesting *TP53* gene mutations, which are particularly common in HGSC but not in LGSC. IHC can also provide some indication of prognosis, with the quantity of CD8+ tumour-infiltrating lymphocytes being associated with longer overall survival of HGSC patients [34], and the level of oestrogen receptor and progesterone receptor expression being associated with disease-specific length of survival in HGSC and EC [35]. However, ancillary testing increases the complexity of diagnosis, so we instead focus on improving the accuracy and objectivity of the information extracted from the standard H&E slides.

Pathological workloads are currently increasing [36, 37] alongside increasing cancer rates [14], causing pathology departments to often be unable to meet demand. Most NHS pathology departments resort to outsourcing work or hiring temporary locums [36] despite the United Kingdom being one of the best-resourced countries worldwide [38]. The number of histopathologists in the NHS is projected to slightly decrease in the coming years, exacerbating current issues [39].

There are significant variations between pathology departments, with smaller departments typically having only a few generalist pathologists whereas larger departments often have many subspeciality experts. When seeking a second opinion from these experts, pathologists often need to send samples over a long distance with an associated financial cost and delay in diagnosis. While pathologists prioritise cases to reduce delays in urgent cases, the constantly rising workload threatens to overwhelm the system. Any delays resulting from demand outstripping diagnostic resources risk catastrophic impacts on patient outcomes, with a four-week delay in cancer treatment being associated with an approximately 10% increased mortality rate among patients [40].

In recent years digital pathology scanners have started to be adopted in some pathology departments, allowing pathologists to assess histology specimens using a computer rather than a microscope. Digitisation can drastically improve the efficiency of the diagnostic process [41, 42] with minimal impact on diagnostic decisions [43, 44]. Digital pathology images can be stored and transported much more easily than physical slides, significantly reducing the logistical burden of outsourcing or seeking a second opinion. Digital images are also much more accessible for the training of new pathologists. However, high start-up costs and technical training requirements have slowed the rate of adoption of digital pathology, with very few departments routinely digitising all pathology slides [45]. Implementing a digital pathology workflow typically costs hundreds of thousands of dollars, causing adoption to be much more common in the largest university hospitals and cancer centres than in smaller community hospitals [45].

2.3 Pathology AI

While the digital pathology workflow has primarily been developed for logistical and long-term financial reasons, it has also allowed for the development of diagnostic AI tools by facilitating the creation of huge digital pathology data repositories. Models have been developed for a wide array of diagnostic and prognostic tasks, including diagnostic classification, tissue type segmentation, cell detection, treatment response prediction, and overall survival prediction [1, 46]. These computer-aided diagnosis tools aim to further increase the efficiency, accuracy, and objectivity of diagnosis. Such tools may be able to automate the most routine aspects of pathological analysis and offer assistance to pathologists in more complex aspects.

One key factor limiting the development of digital pathology tools is the huge size of its images relative to other 2D medical imaging modalities. While a typical 2D slice of a resection sample is only around one square inch in size, the whole slide image (WSI) generated by scanning it at a standard 40x magnification is around 100,000 x 100,000 pixels. At a standard printing resolution of 240 pixels per inch, this single image would cover half of a tennis court. These gigapixel WSIs are typically stored in a pyramidal file format (Figure 2.4), including lower-resolution versions of the same image to facilitate the multi-scale analysis which is required for many diagnostic tasks. Each WSI file comprises gigabytes of information, and a single case of ovarian cancer can generate dozens of samples. Individual data centres generate terabytes of digital pathology data each year [45], and the largest studies are starting to utilise petabytes of data [47]. Automatically analysing such a huge quantity of image data can be incredibly computationally demanding, often requiring servers with expensive graphics processing units (GPUs).

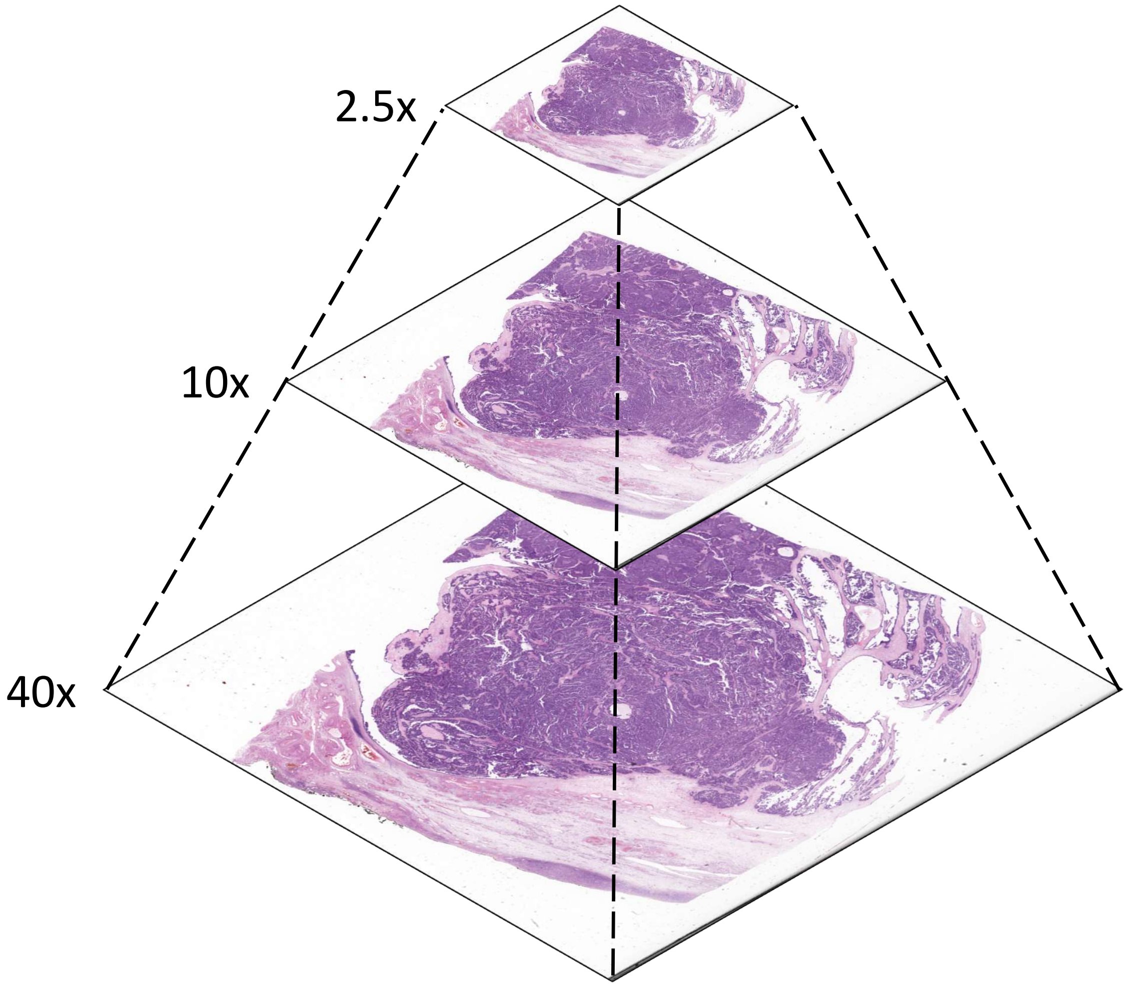


Figure 2.4 Illustration of a pyramidal file including three native tissue magnifications.

The objective of diagnostic pathology is to classify the relevant anatomical differences between samples without being distracted by irrelevant sources of variability (Figure 2.5). Digital pathology slides vary visually due to differences in the sample processing [48] (e.g. cut-up, fixation, and staining protocol) and digitisation [49] (e.g. scanner, magnification, file formatting), as well as anatomical differences [50] (e.g. tissue type, disease, genetics). Such variations are likely to be minor over short periods within a pathology lab, meaning that single-centre data will typically be more homogeneous than multi-centre data. Models trained with single-centre data are likely to generalise poorly to data from different data centres, and even to data from the same centre over time due to changes in tissue processing and digitisation procedures. Similarly, a model trained for a particular disease is unlikely to generalise well to other diseases, with these models typically seen as *narrow AI* due to their focus on very specific tasks.

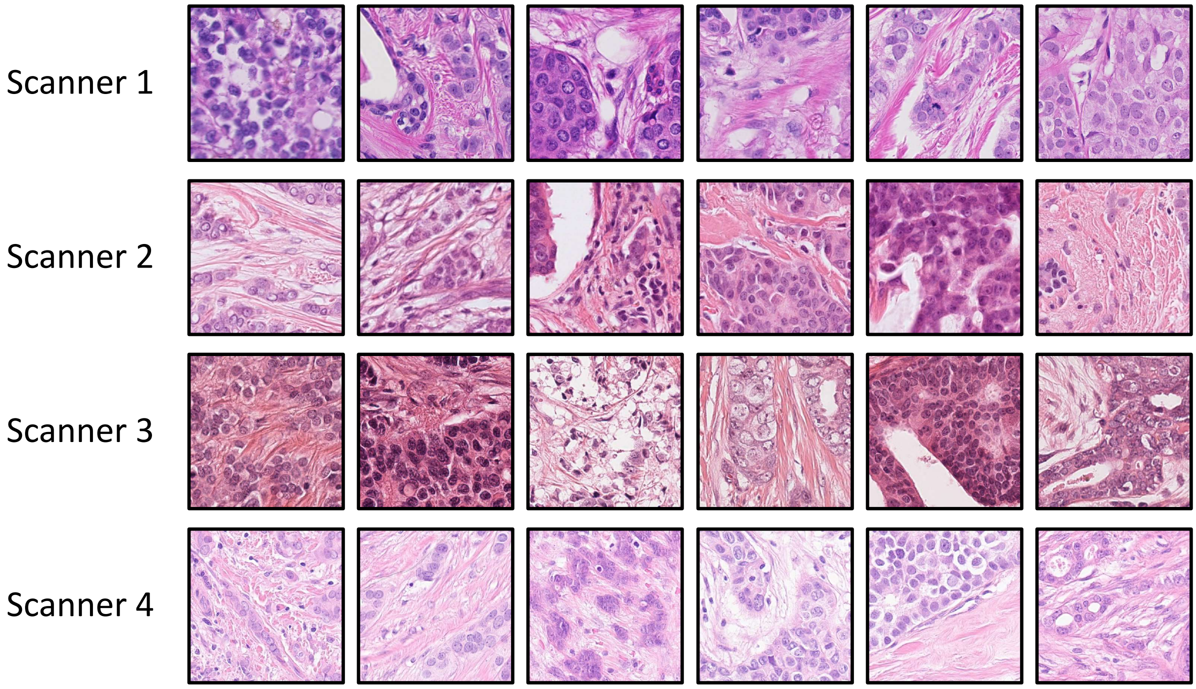


Figure 2.5 Examples of the visual variation in digital pathology images caused by different scanners from the MIDOG 2021 Challenge training set [3]. Each tissue sample was processed in the same laboratory following the same protocol and then digitised with one of four available scanners.

The clinical implementation of digital pathology AI is at a very early stage, with the United States Food and Drug Administration (FDA) having only approved the first *AI-enabled medical device* in digital pathology imaging in 2021. This tool classifies whether prostate biopsies contain malignant cells and indicates the most likely affected area within the WSI [51]. While this is a success story for digital pathology AI, the task of prostate biopsy malignancy classification has many enabling traits - it is a very common disease [14], biopsy slides contain orders of magnitude less tissue than resection slides, there are only two possible classes, and it has a relatively low level of inter-rater variation, with pairwise consensus rates over 90% for the three pathologists in the original study [51]. The high incidence rate of prostate cancer makes it possible to collect vast quantities of varied data, and the relatively small size of biopsy samples makes it possible to train a model with a huge number of samples, allowing for the development of a robust model.

As of September 2023, pathology AI products had only been approved in Europe for the analysis of primary breast, prostate, and gastrointestinal cancers (as well as the detection of lymph node metastases) [52], which all have much greater incidence rates than ovarian cancer [14]. Most approved products focus on the detection/quantification of cancer, with relatively few tools approved for more complex diagnostic or prognostic tasks. It is notable that despite these products being approved, the evidence of their efficacy is limited, with less than half of the products associated with peer-reviewed external validations [52]. Further, evidence is extremely limited on the cost-effectiveness and real-world usability of these models, with clinical validations only starting to be published in 2023 [53]. To be used in the UK, it is expected that AI tools will be evaluated by the Medicines and Healthcare products Regulatory Agency (MHRA) as *software as a medical device* (SaMD). Only a handful of products have successfully undergone this process and received the *UK Conformity Assessed* (UKCA) marking [52]. While some AI tools are starting to achieve very limited clinical adoption [45], none have yet undergone full evaluation by the National Institute for Health and Care Excellence (NICE), so they are not routinely used in NHS pathology departments [52]. Despite the limited adoption, most of the pathologists who responded to an international survey reported optimistic views that AI would complement their work in the future and many expressed an interest in tools to aid with tumour classification (though with greater interest expressed in tools for the objective scoring of IHC and the detection of lymph node metastases) [45].

2.4 Technical Background

2.4.1 Neural Networks

AI comprises a very wide array of automated tools which mimic *intelligence*, typically by making seemingly reasonable interpretations or taking seemingly reasonable actions when given a specific type of input data. This includes machine learning (ML) approaches, which learn the relevant patterns within the input data without explicit instructions. The most common type of AI model applied to digital pathology is the neural network.

Neural networks are mathematical models inspired by the neural pathways in the brain, with nodes (neurons) connected by edges (synapses). Mathematical operations are performed at each node using inputs from any previous nodes, and the result is passed along the edges to any subsequent nodes. The mathematical operations in a node are often described as a linear combination of the input values with a specific weight θ_i applied to each input value x_i, before an additional bias β value is added and an *activation function* $f(x)$ is applied to the result, giving the formula for the output y of the given node:

$$y = f\left(\sum_{i=1}^{N} \theta_i x_i + \beta\right), \tag{2.1}$$

for N inputs. The most simple approaches are feed-forward neural networks, where the initial data is passed to an *input layer*, passed through a series of *hidden layers*, before reaching a final *output layer*. Feed-forward neural networks are typically *fully connected*, meaning every node is connected to all nodes in the immediate prior layer and the immediate subsequent layer. The output layer contains a series of nodes with numerical outputs to be interpreted for the given task (for example, in classification each node may represent a possible class and the one with the highest output value is taken to be the class predicted by the model).

A neural network with more than one hidden layer is considered to be a deep neural network, with *deep learning* increasing the abstraction between the input and output. The abstraction is so great that these models are often referred to as *black-box* models because of how difficult it is to interpret the steps between the input and output of the model.

Feed-forward neural networks are trained using *backpropagation* [54], which starts by calculating a *loss function* at the output node, then steps backwards through the layers of the neural network calculating the gradient of the loss at each node. The size and direction of the gradient is used to determine the adjustments to be made to the model weights in a *gradient descent* approach. The loss function must be differentiable to allow for the application of the chain rule to calculate gradients, and should ideally be continuous to allow for stable convergence to a local minimum.

2.4.2 Computer Vision

Image data is computationally represented as a 2D spatial matrix with red, green and blue (RGB) colour channels, giving a total matrix size of height×width×3 per image. Inputting this matrix directly to a neural network would disregard the inherent spatial patterns, so instead a plethora of methods have been developed specifically for image analysis. Traditional computer vision approaches used pre-defined image features to capture the colours, textures, and edges within an image through approaches such as filtering, thresholding, and histogram analyses [55]. This has been largely supplanted in modern research by approaches which automatically learn relevant features, such as convolutional neural networks (CNNs) and vision transformers (ViTs).

CNNs use *convolutional layers* in which many small filters (often only 3x3 pixels) are passed over the entire image, with the trainable weights used in these filters multiplied by the corresponding pixel values to quantify specific patterns. Stacking convolutional layers iteratively increases the abstraction from the input image, allowing more complex patterns to be modelled. These layers are interleaved with *pooling layers*, which combine neighbouring pixels (often taking the average or maximum value) to reduce the dimensionality of the feature representations. For image classification, the outputs from the final pooling layer are typically flattened to generate a vector of image features which are input to a fully connected neural network.

These models were first popularised in the 1990s, with CNNs increasingly capable of classifying handwritten digits (0-9) in 28 x 28 greyscale images from the MNIST dataset [56]. Their usage expanded in the 2010s as classification performance drastically improved [57] on the 224 x 224 RGB images from the ImageNet dataset [58], which included 1.4 million natural images from 1000 classes. Newer CNNs have been increasingly computationally intensive, from the five layers used in MNIST classification and eight layers used in early winners of the ImageNet challenge [56, 57], to newer models with tens or hundreds of layers [59], which can be impractical to train without GPUs.

2.4.3 Transformers

Many computer vision approaches decompose the input image into many smaller subsections, typically referred to as patches (Section 2.4.4). A sequence of patches in an image is analogous to a sequence of words in a sentence, and so many modelling approaches have been adapted from the field of natural language processing (NLP). One such approach is the *attention mechanism*, which assigns weights to the tokens in the sequence representing their relative importance, with these weights considered when aggregating information from the tokens to make inferences about the sequence as a whole [60].

The *transformer* extended the attention mechanism to *self-attention*, capturing the pairwise relationships between tokens in a sequential input [61]. These were extended from their origin in NLP to create the vision transformer (ViT), which captures relationships between patches in images [62]. The patches are typically small to maximise the learned relational information between pairs of patches (14 x 14, 16 x 16, or 32 x 32 pixels in the original paper [62]). The patch embeddings input to the transformer blocks are simple linear encodings, where the flattened raw pixel values are linearly projected to a desired dimension, which can be understood as passing the input pixels through a single neural network layer.

The attention score matrix A in a transformer is calculated as a function of a query matrix Q and a key matrix K, which is then multiplied by a value matrix V to give the transformer block output [61]. The query, key, and value matrices are calculated by multiplying the patch embeddings matrix X by a trainable weight matrix $U_{\{Q,K,V\}}$. The queries $Q = XU_Q$ represent what each token is seeking, and the keys $K = XU_K$ are compared to the queries to determine the relevance of each token. The values $V = XU_V$ are the encodings of the input tokens which will be aggregated using the attention mechanism.

The scaled dot product attention function is calculated as:

$$A = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right), \tag{2.2}$$

where the scaling factor $\frac{1}{\sqrt{d_k}}$ is a function of the dimension of the keys and values d_k. The output of the transformer block is then simply the attention-weighted values, AV. Several transformer blocks are typically used in parallel with different weights to create a multihead self-attention. We use *self-attention*, where the queries, keys, and values all originate from the same input. The alternative, *cross-attention*, allows the use of multi-modal data by taking keys and values from one data source and queries from another.

Transformers need *position embeddings* to be input alongside the input tokens in order to learn the importance of the arrangement of tokens [61]. In NLP there is a 1D input which typically does not have a fixed sequence length, so a flexible, unbounded function needs to be used to encode the relative positions of tokens (typically based on a sine wave). When using ViTs it is typical to use a fixed input image size, giving a fixed number of patches n, and thus a 1D position embedding can be applied simply by flattening the patch sequence and numbering it from 1 to n. Some more complex approaches have been attempted to accurately map the 2D spatial positions or to calculate relative distances between patches, though these have not demonstrated a benefit over the simple 1D embedding [62], which appears to provide sufficient information to allow the model to learn the spatial structure [63].

ViTs are rapidly growing in popularity as an alternative to CNNs in medical imaging [64]. These models are incredibly scalable [65] and (as with the transformers used in NLP) they benefit from pre-training with huge datasets to create *foundation models* [66, 67], which can be adapted to specific use cases. ViTs and foundation models are extensively evaluated for the task of ovarian cancer histological subtyping in Chapter 7.

2.4.4 Multiple Instance Learning

Slide-level classification is a difficult task because WSIs are too large to be directly input into standard computer vision methods such as CNNs and ViTs. Some researchers have applied these methods to small tissue subsections or heavily compressed WSIs, but such approaches discard a wealth of potentially relevant diagnostic information. We focus on the much more popular approach of multiple instance learning (MIL) [68], where multiple *instances* with shared properties are aggregated in a *bag* for modelling. In digital pathology, this means splitting the WSI into smaller *patches* for modelling, with patch-level information aggregated to make inferences about the whole slide.

In previous literature, the term *multiple instance learning (MIL)* does not have a single agreed definition with clear boundaries. In this thesis, MIL refers to any approach in which patch-level information is aggregated to make slide-level inferences, including voting-based approaches, attention-based aggregations, and graph models. It may be argued that patch-level encoders that decompose inputs into tiny sub-patches (such as ViTs) are a form of MIL, though we focus on approaches which can be applied to WSIs.

While MIL approaches have been researched since the 1990s [69], applications to digital pathology classification did not occur until the mid-2010s [70]. Earlier work would have been impractical due to the rarity and expense of digital pathology scanners and the limitations of computational hardware. Generating pixel-wise annotations for such large images is very time-consuming, so digital pathology MIL is typically used in a weakly-supervised setting, where only slide-level class labels are provided for model training [68]. Most MIL methods in histopathology can be described in five stages - preprocessing, patching, embedding, aggregation, and classification.

Preprocessing refers to the initial adjustments applied to the digital pathology image. Given that a large proportion of most WSIs is the non-tissue background region, it is common to perform tissue segmentation as a preprocessing method [1], which allows for the removal of the plain background to improve the efficiency of the model and to increase the focus on tissue. Some researchers take this a step further by performing tumour segmentation during preprocessing, further focusing the model on relevant tissue. More complex methods may be employed to perform quality control, with artefacts being detected and adjusted. Other typical preprocessing approaches include downsampling to reduce the effective tissue magnification and hence the overall size of the image, and chromatic adjustments to either reduce visual variability (normalisation) or increase it (augmentation). Most approaches may be applied either before or after patching.

Patching is the process of splitting the WSI into computationally manageable subsections. Patch sizes can be variable, though 256 x 256 or 512 x 512 pixel patches will typically be used for CNN-based models and 224 x 224 pixel patches for ViT-based models. Larger patches give a greater context window, whereas smaller patches are typically more computationally efficient. A standard 40x magnification WSI is around 100,000 x 100,000 pixels in size, so using a patch size of 256 x 256 pixels gives around 150,000 unique, non-overlapping patches.

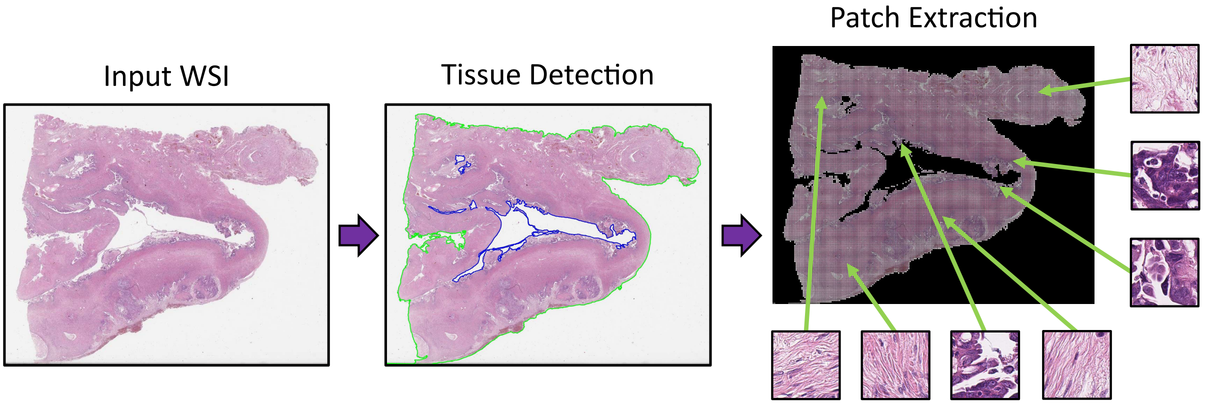


Figure 2.6 Example preprocessing and tissue patch extraction procedure, with tissue segmentation preprocessing and 256 x 256 pixel patches extracted from a 40x magnification WSI.

Embedding is the process of extracting features from the patches. While traditional AI methods in histopathology used hand-crafted features, it is now more common to use CNNs or ViTs to automatically learn relevant features [1]. Embedding patches can drastically reduce their dimensionality to make further modelling computationally tractable.

Transfer learning is typically employed during this stage, meaning that the feature extractor is pre-trained using a particular *source* dataset and then adapted for usage on a *target* dataset. Ideally a deep learning model would be trained end-to-end, with all model weights updated in a single backpropagation pass, but this may be computationally impractical when applying MIL to such large images. It is common for the feature extractor to be *frozen*, meaning it is kept in a fixed state during the training of the subsequent model weights, allowing patch features to be pre-computed and stored before model training. The feature extractor may, however, be fine-tuned to the target domain before being frozen within the MIL model. Transfer learning often improves the speed of model convergence and may also benefit the final classification performance, especially when there is not a large enough target dataset to thoroughly train a feature extractor from scratch, or when computational hardware is insufficient to train a model end-to-end.

Aggregation approaches collate information from the different patches in a slide. Aggregation methods can be grouped into *instance classification* and *instance embedding* approaches based on whether the instances are individually classified. The simplest instance classification approach is max-pooling, where the instance with the highest individual classification score represents the entire bag. The simplest instance embedding approach is mean-pooling, where a slide-level embedding is generated as the mean of all patch-level embeddings and then passed through a slide-level classifier.

Classification approaches depend upon aggregation approaches. If an aggregation approach generates slide-level embeddings then these can be classified through standard classification approaches (support vector machines (SVMs), decision trees, k-nearest neighbours, etc.), with neural networks being the most popular choice in modern research [1]. For instance classification techniques, slide-level classification depends upon instance scores/classes, for example taking the most common patch-level class as the slide-level class. The phrases *mean-pooling* and *average-pooling* in previous literature may refer to either the aforementioned instance embedding approach [71] or to an instance classification approach in which the average patch classification score is taken as the slide classification score [70], though this is more commonly called *average-vote*. Bag classification techniques can also be adjusted to perform bag-level regression, clustering, and ranking, though classification remains the most common [72].

Instance Classification Approaches

Max-pooling is the simplest instance classification approach, but it is only suitable when the traditional MIL assumption holds - if any instance is positive then so is the entire bag, but if all instances are negative then the bag is negative. This applies to malignancy classification, with a WSI classified as malignant if any patch within it contains malignant tissue and classified as benign if no patch contains malignant tissue. However, the assumption does not hold for multi-class subtyping.

More complex instance classification approaches were introduced in the early 2000s, leveraging multiple instances per bag. These approaches used instance classification scores to identify a set of most relevant instances, then aggregated these using ML approaches such as SVMs [73] and neural networks [74]. Often the number of relevant instances per bag was set as a hyperparameter K, with the top-k instances used to represent the bag. Some of these approaches have utilised instance embeddings, but these are still *instance classification* methods since the aggregation depends upon individual instance classification scores. These more complex approaches increased the abstraction from the traditional MIL assumption and better accommodated multi-class classification.

A modern application of top-k patch selection used a recurrent neural network (RNN) to classify the slide based on the top-k patch feature embeddings [75]. This approach achieved very high accuracy for malignancy classification with three different types of cancer, and underpinned the first FDA-approved AI-enabled medical device in digital pathology [51]. However, this was found to be a particularly data-hungry approach, with an ablation study finding that at least 8,000 training WSIs were required to minimise validation error in a homogeneous set of prostate biopsy slides. It is further noteworthy that the top-k aggregation did not offer a significantly better performance compared to a simple max-pooling aggregation in either the original study or a similar study which applied the method in a fully-supervised manner [76].

Instance Embedding Approaches

Instance embedding is the more common MIL aggregation approach in digital pathology, and is more directly applicable to multi-class classification. In instance embedding approaches, patches are encoded to a latent embedding space and then combined to generate a latent representation of the entire slide for classification. The most basic approach is the aforementioned mean-pooling, with the average of the patch feature embeddings used as the slide embedding for classification.

Attention-based multiple instance learning (ABMIL) is a more advanced approach in which the mean aggregation of instance embeddings is weighted based on trainable attention scores [77]. The attention weight (a_k) of an instance is calculated with the following equation:

$$a_k = \frac{\exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_k^\top)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_j^\top)\}}, \tag{2.3}$$

for *1*$\times M$ instance embedding $\mathbf{h}_k \in \{\mathbf{h}_1, ..., \mathbf{h}_K\}$, $L\times$*1* parameter vector $\mathbf{w}$, and $L\times M$ parameter matrix $\mathbf{V}$. Dimensions L and M are pre-defined hyperparameters. An alternate version, *gated attention*, calculates weights similarly but includes a sigmoid non-linearity in an attempt to better learn complex relations:

$$a_k = \frac{\exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_k^\top) \odot \sigma(\mathbf{U}\mathbf{h}_k^\top))\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_j^\top) \odot \sigma(\mathbf{U}\mathbf{h}_j^\top))\}}, \tag{2.4}$$

where $\mathbf{U}$ is also an $L\times M$ parameter matrix, σ is the sigmoid function, and $\odot$ is an element-wise multiplication. The seminal approach also included trainable transformations before and after aggregation to increase model flexibility. In the original publication [77], ABMIL achieved state-of-the-art results on two histopathology datasets, though the images used were particularly small at less than 1000 x 1000 pixels each, where a typical WSI is around 100,000 x 100,000 pixels. It was unclear whether standard attention or gated attention was better, with each outperforming the other in some experiments.

Cluster-constrained attention MIL (CLAM) attempted to further refine the feature space by clustering within high-attention regions during training [78]. Within this study it was shown how attention-based methods could be extended to multi-class classification by using a parallel attention branch for each class, with each of the class-specific slide representations passed to the final classification layer of the network. Surprisingly, the classification performance of ABMIL was not compared in the original study despite CLAM being an adaptation of this approach. CLAM has since become a very commonly used benchmarking model despite it remaining unclear whether it is better than ABMIL and whether the multiple attention branches provide a benefit [79–82]. The popularity of CLAM may be influenced by its particularly well-developed open-source code repository.

2.4.5 Spatial MIL Networks

One key limitation of the MIL methods explored thus far is that patches are processed independently, with the WSI modelled as a bag of patches without any spatial structure. However, the spatial arrangement of a tissue sample is likely to be diagnostically relevant as it contains information such as the size of the tumour, the extent of invasion, and the immune response to the tumour. As such, many recent approaches have focused on modelling these spatial relationships, typically using either transformer or graph networks.

Transformers

ViTs have been successfully applied to the classification of natural images and written digits [62], but are limited in their applicability to WSIs by the quadratic computational complexity of capturing the pairwise relationship between patches. In the original vision transformer paper [62], a typical image contained only 196 patches (224 x 224 pixel inputs with 16 x 16 patches), but a typical WSI contains thousands or tens of thousands of patches. Taking larger patches may help to compensate but this results in less fine-grained spatial information in the transformer and it requires a more computationally complex patch embedding model, thus it is not a sufficient solution.

Approaches to implementing transformers for slide-level classification have included taking an approximation to self-attention which is of linear complexity $O(n)$ rather than quadratic $O(n^2)$ [71], and stacking multiple transformers at different tissue scales [83], with both approaches reported to improve performance over non-transformer aggregation techniques. Recent research has focused much more heavily on applying transformers during the patch embedding stage to create histopathology foundation models, which are trained on large domain-specific datasets to learn domain-specific features. We explore histopathology foundation models in Chapter 7.

Graph Networks

Graph networks offer a different approach to modelling spatial relationships. In a graph, *nodes* are connected by *edges*, and information is passed along the edges to allow nodes to gain contextual information. Computationally, a graph can be characterised by a pair of matrices $g = (\mathbf{X}, \mathbf{A})$. For a graph with n nodes, the feature matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ is composed of m-dimensional node feature vectors $\mathbf{x}_1, ..., \mathbf{x}_n$, and the adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a sparse matrix encoding which nodes are connected by edges.

In histopathology, graphs can model tissue as connected cells or tissue patches. Cell graphs are typically only applied to small regions of interest rather than entire resection specimens [84–94] as the incredibly vast number of cells per tissue sample is computationally limiting. When cell graphs have been applied to WSIs they have relied on subgraph sampling approaches, with analyses limited to relatively few WSIs [95, 96]. Only one previous approach has performed slide-level classification directly from a cell graph, and this was only applied to IHC specimens with drastically fewer visible cells than in a standard H&E sample [97].

Patch graphs are more directly applicable to slide-level classification as the number of patches is typically orders of magnitude smaller than the number of cells - a single WSI may contain millions of cells [95, 98] but only tens of thousands of patches. Patch graphs are a natural extension of MIL, with the bag and instances defined in the same way, but connections added between (spatially) related instances. Some graph approaches are neither cell graphs nor patch graphs, with adaptive graphs modelling regions determined by segmentation or clustering approaches [99–101]. This may reduce the computational complexity of the graph compared to a cell graph with more flexibility than a patch graph, though these approaches have been less thoroughly explored in previous literature [101].

When a graph structure has been defined, the graphs may be input into a graph neural network (GNN) [102]. GNNs utilise message-passing layers to share information between connected nodes, and graph-pooling layers to reduce the number of nodes, with several of each of these layers used to pass information to distant parts of the graph and to iteratively reduce the graph size. Graph-based MIL has been applied to image classification for many years, with the graph layers followed by a MIL aggregation approach to turn the remaining nodes into a whole-graph feature embedding for classification [103]. We explore graph-based MIL in Chapter 8.

2.5 Classification Metrics

The most straightforward metric to quantify WSI classification performance is accuracy, though this metric only provides information at a single decision threshold and becomes distorted by class imbalances. Balanced accuracy is an improvement as it takes the average of the accuracy score for each class, thus it is more robust to class imbalances. The F1 score is a similar metric which takes the harmonic mean of precision and recall at a single threshold. The most commonly reported metric in previous research is the area under the receiver operating characteristic curve (AUROC) [1]. The receiver operating characteristic curve compares true positive and false positive rates across classification thresholds, so the area under this curve (the AUROC) gives a more holistic measure of model performance which is independent of the classification threshold. However, this metric is too abstract to be a clear measure of clinical utility alone.

Each of the chosen metrics gives a score between 0 and 1 (with higher scores being better, and 1 being perfect), meaning they can be expressed as percentages. When classifying a dataset with N classes, where the most common class accounts for a proportion q of the dataset, to demonstrate predictive power a model should outperform a model which always selects the most common class, which gives an accuracy of q, balanced accuracy of $\frac{1}{\mathrm{N}}$, macro-averaged F1 score of $\frac{2q}{\mathrm{N}(q+1)}$, and AUROC of 0.5. For example, in our largest cross-validation experiments (in Chapters 7 and 8), we have $\mathrm{N} = 5$ classes and $q = 0.68$ as the proportion of the most common class, giving baseline scores of 68% accuracy, 20% balanced accuracy, 0.16 F1 score, and 0.5 AUROC.

		Prediction	
		Positive	Negative
Ground Truth	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Table 2.1 Confusion matrix.

Each metric can be expressed as a function of the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) class predictions (described in Table 2.1) as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

$$\begin{aligned}\text{Balanced Accuracy} &= \frac{\text{TPR} + \text{TNR}}{2} \\ &= \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right),\end{aligned}$$

$$\begin{aligned}\text{F1 Score} &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \\ &= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}},\end{aligned}$$

$$\begin{aligned}\text{AUROC} &= \int_0^1 \text{TPR}(\text{FPR}) \; d(\text{FPR}) \\ &= \int_0^1 \frac{\text{TP}}{\text{TP} + \text{FN}} \left(\frac{\text{FP}}{\text{FP} + \text{TN}}\right) \; d\left(\frac{\text{FP}}{\text{FP} + \text{TN}}\right),\end{aligned}$$

where TPR is the true positive rate, TNR is the true negative rate, and FPR is the false positive rate, and where TPR(FPR) is TPR as a function of FPR.

These metrics help to determine the discriminative performance of a model, but this is not the only aspect of a model that is relevant to the clinical utility. Due to the huge size of WSIs, classifiers are often computationally intensive, requiring multi-GPU servers for training and inference. Such hardware is unlikely to be directly available to clinicians, so it is also pertinent to measure the size and/or efficiency of models, which can be done in terms of speed, memory requirements, or number of model parameters. The efficiency of inference is more directly relevant to clinicians than the efficiency of model training, as models can be trained in a research setting and then deployed to the clinical setting for slide evaluation. Efficient model training may be beneficial in allowing models to be trained with more data, more extensively tuned, or trained with additional augmentation techniques, which can lead to benefits in the classification performance of a model.

Chapter 3

Systematic Literature Review

In this chapter, we explore published research for the diagnosis or prognosis of ovarian cancer from digital pathology images. We systematically review such research, characterising the methods used and the clinical tasks addressed. We assess the risks of bias in each study and provide recommendations for subsequent research.

3.1 Introduction

AI in digital pathology is a broad and rapidly growing field. To conduct relevant, high-impact research, it is essential to first understand the current state of the field. Previous reviews of AI in gynaecological cancers have given broad overviews of the field without a comprehensive synthesis of all published research using ovarian cancer histopathology data [104–107]. We instead systematically reviewed all literature in which AI techniques (comprising both traditional ML and deep learning methods) were applied to digital pathology images for the diagnosis or prognosis of ovarian cancer [1]. This included research which focused on a specific diagnostic factor (such as histological subtype), and studies that performed computer-aided diagnostic tasks (such as tumour segmentation).

In the review, we characterised relevant studies and assessed their quality. We then provided insights and recommendations based on published literature to improve the clinical utility of subsequent research, including reducing risks of bias, improving reproducibility, and increasing generalisability.

We developed and registered a study protocol (PROSPERO CRD42022334730) defining the scope and methodology of the review. There were two research questions to be addressed:

- What diagnostic/prognostic tasks have been addressed using AI methods for ovarian cancer using histopathology data?
- What underlying AI methodology did these studies use, how well did they perform, and how reliable was the research?

This was a multi-disciplinary effort in which I (JB) managed a group involving two pathologists (KA, NMO), an oncologist (KZ), and a computer science academic (NR). While I planned the review and wrote the manuscript, the other group members offered regular feedback and were directly involved in the literature selection, risk of bias assessment, and data synthesis stages to ensure a fair and balanced review process. The Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 guidelines for reporting systematic reviews were followed, with the checklist provided in Appendix B.

3.2 Literature Search and Selection

Literature Search

Searches were conducted in three research databases, PubMed, Scopus and Web of Science, and two trial registries, Cochrane Central Register of Controlled Trials (CENTRAL) and the World Health Organisation International Clinical Trial Registry Platform (WHO-ICTRP). The chosen research databases only included journal papers and conference proceedings which had undergone peer review, ensuring a basic level of integrity in the included research.

The search strategy was composed of three distinct concepts - artificial intelligence, ovarian cancer, and histopathology. For each concept, multiple relevant terms were combined using the *OR* operator (e.g. "artificial intelligence" OR "machine learning"), and then these were combined using the *AND* operator to ensure that retrieved research included all three concepts. Many AI approaches build on statistical models, such as logistic regression, which can blur the lines between disciplines.

When conducting searches, a previously reported methodology was adopted [108] in which typical AI approaches were identified by name (e.g. neural networks), and other methods were identified by the authors describing their work using terms such as *artificial intelligence*. Search terms are shown in Table 3.1 and the full search strategies for each research database are shown in Appendix A.

Artificial Intelligence	**Ovarian Cancer**	**Histopathology**
Machine Learning	Ovarian Carcinoma	Histology
AI	Ovarian Mass	Digital Pathology
ML	Ovarian Tumour	Whole Slide Image
Deep Learning	Ovarian Neoplasm	Tissue Slide
Active Learning	Ovarian Malignancy	Pathology Slide
Computer-aided Diagnosis	Fallopian Cancer†	Pathology Image
Computer-assisted Diagnosis	Fallopian Carcinoma†	Tissue Microarray
Computer Vision	Fallopian Mass†	Immunohistochemistry
Neural Network	Fallopian Tumour†	Haematoxylin and Eosin
Deep Network	Fallopian Neoplasm†	Computational Pathology†
Recurrent Network	Fallopian Malignancy†	
Convolutional Neural Network	Peritoneal Cancer†	
Graph Network	Peritoneal Carcinoma†	
Perceptron	Peritoneal Tumour†	
Multiple Instance Learning		
Support Vector Machine		
Random Forest		
Ensemble		
Nearest Neighbour		
Gradient Boosting		
Backpropagation		
Segmentation		

Table 3.1 Systematic review search terms grouped by concept. Wildcards were used to allow for different spellings and suffixes (e.g. "patholog*" to allow for "pathology" and "pathologist"), and "AND/OR" operators were used to allow different combinations of words (e.g. "fallopian AND cancer" to allow for "fallopian tubes cancer" or "cancer of the fallopian tubes"). †These terms were added after peer-review feedback.

The widest possible set of search fields was used for each search engine except for Scopus, where restrictions were imposed to avoid searching within the citation list of each article, which was not an available field in the other search engines. The terms *ML* and *AI* were restricted to specific fields due to the diversity of their possible meanings. To ensure the most rigorous literature search possible, no restrictions were placed on the publication date or article type during searching. Searches were initially conducted on 25/04/2022 and were most recently repeated for the systematic review on 19/05/2023 after the first round of peer review. To bring the thesis up-to-date, searches were repeated again on 25/06/2024, with the most recent literature described in Section 3.5.3.

Literature Selection

First, duplicate papers were manually removed with the assistance of the referencing software *EndNote X9*. Then, two researchers (JB, KA) independently screened all articles for inclusion in two stages, the first based on the title and abstract alone, and the second based on the full manuscript. In any case where these researchers disagreed on whether a paper should be included in their independent assessments, their inclusion was discussed and, if necessary, arbitrated by a third researcher (NR or NMO). Trials in WHO-ICTRP did not have associated abstracts, so only the titles were available for the initial screening.

The inclusion criteria required that research evaluated the use of at least one AI approach to make diagnostic or prognostic inferences on human histopathology images from suspected or confirmed cases of ovarian cancer. Studies were only included where AI methods were applied directly to the digital pathology images, or to features which were automatically extracted from the images. Fundamental tasks, such as segmentation and cell counting, were included as these could be used by pathologists for computer-aided diagnosis. Only conventional light microscopy images were considered, with other imaging modalities, such as fluorescence and hyperspectral imaging, excluded. Multi-modal approaches were included as long as the pathology modality met this criteria. Publications which did not include primary research were excluded (such as review papers and comments). Non-English language articles and research where a full manuscript was not accessible were also excluded.

The initial searches (25/04/2022) returned 1305 records, of which 28 were eligible for inclusion, with the final searches (19/05/2023) bringing this up to 1573 records and 45 inclusions. As shown in Figure 3.2, of the 1573 total records, 557 were duplicates, 930 were excluded during the screening of titles and abstracts, and 41 were excluded based on full paper screening, including 3 records for which full articles could not be obtained. The remaining 45 studies included 11 conference papers and 34 journal articles. All accepted studies had originally been identified through searches of research databases, with no records from trial registries meeting the inclusion criteria. While the searches returned literature from as early as 1949, all of the research which met the inclusion criteria had been published since 2010, with over 70% of the included literature published since 2020, as shown in Figure 3.1.

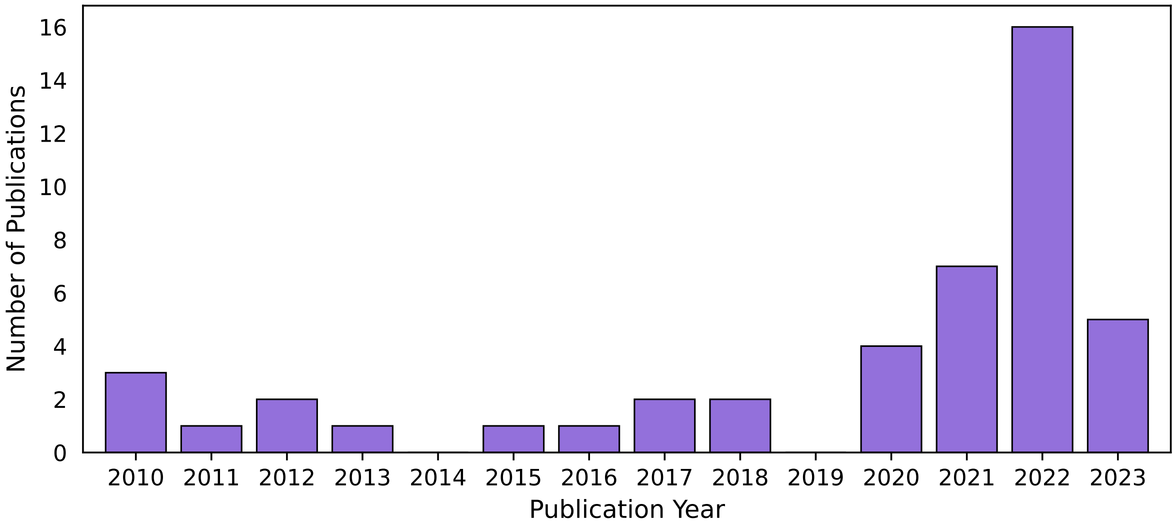


Figure 3.1 Number of publications included in the systematic literature review by publication year (final searches on 19/05/2023).

An AI model in an included study was considered to be a *model of interest* if it met the same inclusion criteria as was used for selecting papers. Where multiple models were compared for the same outcome, the model of interest was taken to be the newly proposed model, with the best performing model during validation taken if this was unclear. If multiple model outcomes were assessed in the same study, a model of interest was taken for each model outcome, regardless of any similarity in modelling approaches. Models investigating the same outcome at different levels of precision (e.g. patch-level, slide-level, patient-level) were not considered to be different model outcomes. Models didn't need to be entirely independent, for example, the output of one model of interest could have been used as the input of another model of interest on the condition that model performance was separately evaluated for each model. Applying these criteria, we found 80 models of interest in the 45 included studies, with up to six models of interest per paper.

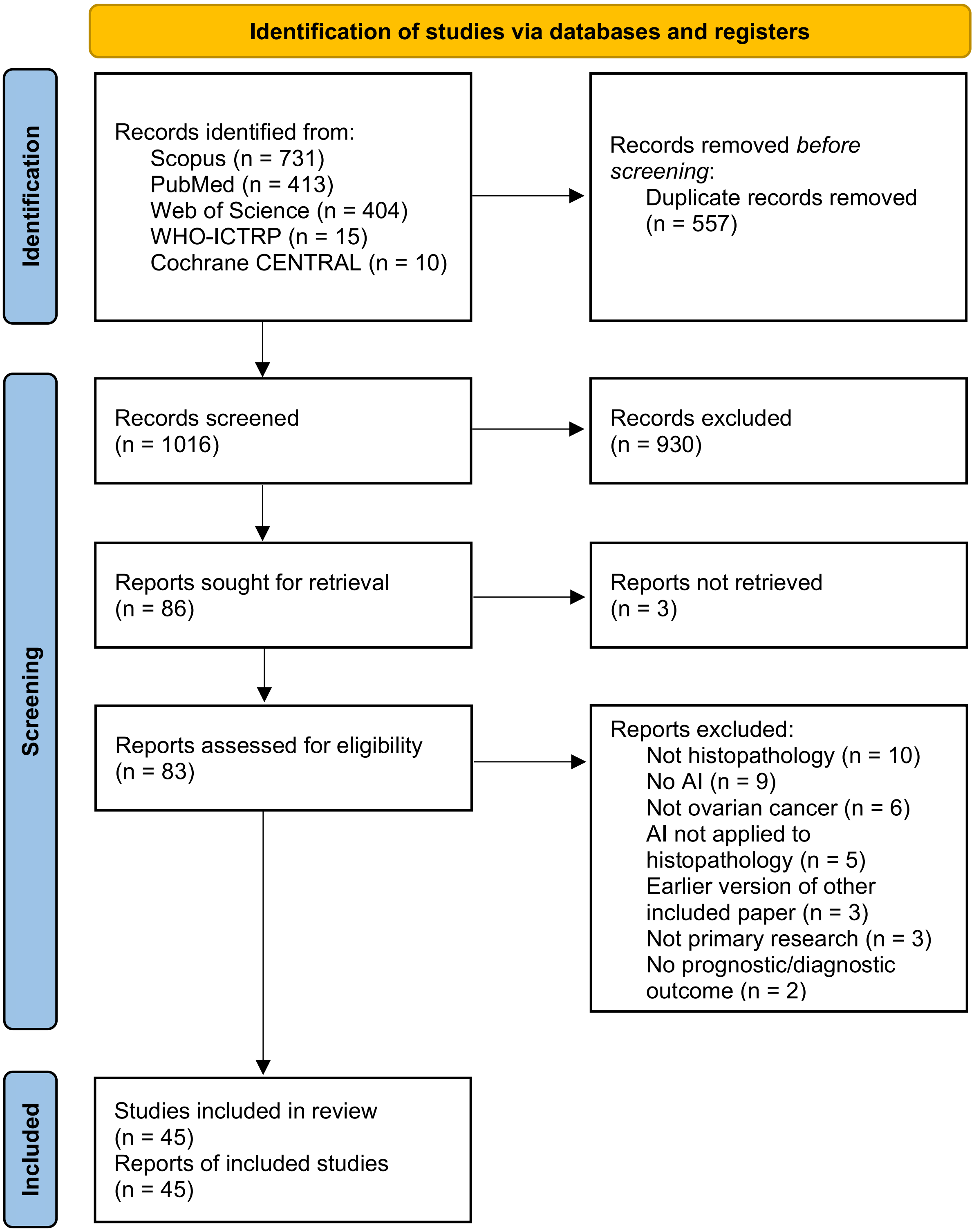

Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
Scopus (n = 731)
PubMed (n = 413)
Web of Science (n = 404)
WHO-ICTRP (n = 15)
Cochrane CENTRAL (n = 10)
Records removed *before screening*:
Duplicate records removed (n = 557)
Records screened (n = 1016)
Records excluded (n = 930)
Reports sought for retrieval (n = 86)
Reports not retrieved (n = 3)
Reports assessed for eligibility (n = 83)
Reports excluded:
Not histopathology (n = 10)
No AI (n = 9)
Not ovarian cancer (n = 6)
AI not applied to histopathology (n = 5)
Earlier version of other included paper (n = 3)
Not primary research (n = 3)
No prognostic/diagnostic outcome (n = 2)
Studies included in review (n = 45)
Reports of included studies (n = 45)

Figure 3.2 PRISMA 2020 flowchart of the finalised study identification and selection process for the systematic review. Records were screened on titles and abstracts alone, and reports were assessed based on the full-text content.

3.3 Data Synthesis

Data extraction was performed independently by two researchers (JB, KA) using a form containing 81 fields within the categories *Overview*, *Data*, *Methods*, *Results*, and *Miscellaneous*. Several of these fields were added or clarified during data extraction with the agreement of both researchers and retroactively applied to all accepted literature. The final data extraction form is available on GitHub (www.github.com/scjjb/OvCaReview), with a summary shown in Table 3.2.

Category	**Data Extraction Fields**
Overview	Internal ID. Lead author. Year. Conference/Journal name.
Data	Number of development images. Total number of images. Type of samples. FFPE/Frozen. Size of images. Tissue of origin. Number of development patients. Total number of patients. Number of data collection centres. Type of stain. Number of stainers. Scanners. Number of scanner types. Number of tissue processing centres. Data origin countries. Number of pathologists for data labelling. Online dataset. Prospective/retrospective. Clinical/research tissue. Data annotation. Maximum magnification available. Supplementary datatypes. Data exclusion reasons. Number of images excluded. Other cancer types included.
Methods	Outcome. Outcome measure/classes. Outcome standards/definition. Magnifications used. Patch sizes. Patches per image. Task type. Feature extraction type. Feature extractors. AI in main method. Other AI methods. Optimiser. Number of external validations. Differences to external validation set. Total external validation images. Number of cross-validation folds. Number of non-novel methods compared. Number of GPUs. Type of GPUs.
Results	Internal test accuracy. AUROC. Sensitivity/specificity. Other metric(s). External training type. External test accuracy. AUROC. Sensitivity/specificity. Other metric(s). Type of error bounds. Model training time. Visualisation type.
Miscellaneous	Code availability. Data availability. Notes.

Table 3.2 Summary of the fields used for data extraction in the systematic review.

Extracted data are presented in two tables, with Table 3.3 showing the 45 included studies and Table 3.4 showing the 80 models of interest. The term *model outcome* refers to the model output, whether this was a clinical outcome (diagnosis/prognosis), or a diagnostically relevant outcome that could be used for computer-aided diagnosis, such as tumour segmentation. Meta-analysis was not performed given the diversity of included methods and model outcomes.

Publication	Ovarian Cancer Data Source	Models of Interest	Outcome Type(s)	Model Outcome(s)	Published Code
Dong 2010(a) [109]	Unclear	1	Other	Stain segmentation	None
Dong 2010(b) [110]	Unclear	1	Other	Stain segmentation	None
Signolle 2010 [111]	Unclear	1	Other	Tumour segmentation	None
Janowczyk 2011 [112]	Unclear	1	Diagnosis	Malignancy	None
Janowczyk 2012 [113]	Unclear	1	Other	Stain segmentation	None
Kothari 2012 [114]	TCGA-OV (Multi-city, USA)	1	Diagnosis	Malignancy	None
Poruthoor 2013 [115]	TCGA-OV (Multi-city, USA)	2	Diagnosis, Prognosis	Grade; Overall survival	None
BenTaieb 2015 [116]	Transcanadian Study (Multi-city, Canada)	1	Diagnosis	Histological subtype	None
BenTaieb 2016 [117]	Transcanadian Study (Multi-city, Canada)	1	Diagnosis	Histological subtype	Inaccessible
BenTaieb 2017 [118]	Unclear	1	Diagnosis	Histological subtype	Inaccessible
Lorsakul 2017 [119]	Unclear	1	Other	Cell type	None
Du 2018 [120]	Unique (Oklahoma, USA)	1	Other	Tissue type	None
Heindl 2018 [121]	TCGA-OV (Multi-city, USA)	1	Other	Cell type	https://yuanlab.org/file/Ov3sweave2.pdf
Kalra 2020 [122]	TCGA-OV (Multi-city, USA)	4	Diagnosis	Primary cancer type	None
Levine 2020 [123]	OVCARE (Vancouver, Canada)	1	Diagnosis	Histological subtype	https://github.com/AIMLab-UBC/pathGAN
Yaar 2020 [124]	TCGA-OV (Multi-city, USA)	1	Prognosis	Treatment response	https://github.com/asfandasfo/LUPI
Yu 2020 [125]	TCGA-OV (Multi-city, USA)	4	Diagnosis, Prognosis	Malignancy, Grade, Transcriptomic subtype; Treatment response	https://github.com/khyu/ovarian_ca/
Gentles 2021 [126]	Unique (Newcastle, UK)	6	Other	Stain quantity/intensity	None
Ghoniem 2021 [127]	TCGA-OV (Multi-city, USA)	1	Diagnosis	Stage	None
Jiang 2021 [128]	Mayo Clinic (Rochester, USA)	1	Diagnosis	Malignancy	https://github.com/smujiang/CellularComposition
Laury 2021 [129]	Unique (Helsinki, Finland)	1	Prognosis	Progression-free survival	None
Paijens 2021 [130]	Unique (Groningen & Zwolle, The Netherlands)	1	Other	Tissue type	None
Shin 2021 [131]	TCGA-OV (Multi-city, USA) & Unique (Ajou, Korea)	1	Diagnosis	Malignancy	https://github.com/ABMI/HistopathologyStyleTransfer
Zeng 2021 [132]	TCGA-OV (Multi-city, USA) & Unique (Shanghai, China)	5	Diagnosis, Prognosis	Genetic mutation, Transcriptomic subtype, Microsatellite instability; Overall survival	None
Boehm 2022 [133]	TCGA-OV (Multi-city, USA) & MSKCC (New York, USA)	3	Diagnosis, Prognosis	Malignancy; Overall survival, Progression-free survival	https://github.com/kmboehm/onco-fusion
Boschman 2022 [134]	OVCARE (Vancouver, Canada)	1	Diagnosis	Histological subtype	None
Elie 2022 [135]	Unique (Caen, France)	3	Other	Stain quantity/intensity	None
Farahani 2022 [136]	OVCARE (Vancouver, Canada) & Unique (Calgary, Canada)	2	Diagnosis	Malignancy, Histological subtype	https://github.com/AIMLab-UBC/ModernPath2022
Hu 2022 [137]	TCGA-OV (Multi-city, USA)	1	Diagnosis	Epithelial-mesenchymal transition	https://github.com/superhy/LCSB-MIL
Jiang 2022 [138]	Mayo Clinic (Rochester, USA)	4	Diagnosis, Other	Tumour-stroma reaction; Tumour segmentation	https://github.com/smujiang/TumorStromaReaction
Kasture 2022 [139]	TCGA-OV* (Multi-city, USA)	1	Diagnosis	Histological subtype	https://github.com/kokilakasture/OvarianCancerPrediction
Kowalski 2022 [140]	Unclear	1	Other	Tumour segmentation	None
Lazard 2022 [141]	TCGA-OV (Multi-city, USA)	1	Diagnosis	Homologous recombination deficiency status	https://github.com/trislaz/wsi_mil
Liu 2022 [142]	TCGA-OV (Multi-city, USA)	1	Prognosis	Overall survival	https://github.com/RanSuLab/EOCprognosis
Mayer 2022 [143]	TCGA-OV (Multi-city, USA) & Unique (Frankfurt, Germany)	1	Diagnosis	Malignancy	None
Nero 2022 [144]	Unique (Rome, Italy)	2	Diagnosis, Prognosis	Genetic mutation; Relapse	None
Salguero 2022 [145]	TCGA-OV (Multi-city, USA)	1	Diagnosis	Malignancy	None
Wang 2022(a) [146]	Tri-Service (Taipei, Taiwan)	4	Prognosis	Treatment response	None
Wang 2022(b) [147]	Tri-Service (Taipei, Taiwan)	1	Prognosis	Treatment response	None
Yokomizo 2022 [148]	Unique (Tokyo, Japan)	3	Prognosis	Overall survival, Progression-free survival, Relapse	Inaccessible
Ho 2023 [149]	MSKCC (New York, USA)	2	Diagnosis, Other	Genetic mutation; Tumour segmentation	https://github.com/MSKCC-Computational-Pathology/DMMN-ovary
Meng 2023 [150]	Unique (Beijing, China)	1	Diagnosis	Malignancy	https://github.com/dreambamboo/STT-BOX-public
Ramasamy 2023 [151]	TCGA-OV* (Multi-city, USA)	2	Diagnosis, Other	Primary cancer type; Tumour segmentation	None
Wang 2023 [152]	Tri-Service (Taipei, Taiwan)	4	Prognosis	Treatment response	https://github.com/cwwang1979/OvaryTreatment_AnginPKM2VEGF
Wu 2023 [153]	TCGA-OV (Multi-city, USA)	1	Prognosis	Overall survival	None

Table 3.3 Characteristics of the 45 studies included in the systematic review. Details are shown for individual models in Table 3.4. Code is labelled as inaccessible where it could not be found despite a link being provided in the publication. *Indicates papers where significant discrepancies were found regarding the data source, as described in Section 3.5.

Diagnosis Outcome	Publication	Internal Participants	Internal Pathology Images	Other Data*	Stain Type	Original Image Size	Patch Size (Pixels)	Magnifications	Feature Extraction	Histopathological Features	Final Model	Prediction Precision	Classes	Validation Type	External Validation Data	Metric	Internal Results	Internal Variability (Measure)	External Results	External Variability (Measure)
Malignancy Status	Janowczyk 2011	6	11		IHC	1400x1400	Unclear	40x	Hand-crafted	Texture, cellular morphology	Probabilistic Boosting Tree	Patch	2 - Tumour, stroma	Monte Carlo cross-validation (5 reps)		AUC	0.8341			
	Kothari 2012	571	1301		H&E	WSI	512x512	Unclear	Hand-crafted	Colour, texture, cellular and nuclear morphology	SVM	Patch	2 - Tumour, non-tumour	Single train/test split		Accuracy	90%			
	Yu 2020	587	1375		H&E	WSI	Unclear	Unclear	Learned	CNN features (VGG16)	CNN (VGG16)	WSI	2 - Malignant, benign	Monte Carlo cross-validation (3 reps)		AUC	0.975	±0.001 (unclear)		
	Jiang 2021	30	≥30		H&E	WSI	512x512	40x	Hand-crafted	Colour, cellular and nuclear morphology	SVM	Patch	2 - HGSC, Serous borderline tumour	Unclear		Accuracy	90.64%			
																AUC	0.96			
	Shin 2021	142	≥142		H&E	WSI	256x256	Unclear	Learned	CNN features (Inception V3)	CNN (Inception V3)	Patch	2 - Cancer, non-cancer	External validation	32 WSIs from different centre	Accuracy	98.3%		80.8%	
																AUC	0.998	0.995-0.999 (95% CI)	0.916	0.899–0.930 (95% CI)
	Boehm 2022	283	≥283		H&E	WSI	128x128	Unclear	Learned	CNN features (ResNet18)	CNN (ResNet18)	Patch	4 - Tumour, stroma, fat, necrosis	4-fold cross-validation		Accuracy	88%			
	Farahani 2022	≤416	416		H&E	WSI	512x512	20x	Learned	CNN features (ResNet18)	CNN (ResNet18)	Patch	2 - Tumour, stroma	3-fold cross-validation		Balanced accuracy	96.99%			
																AUC	0.9441			
	Mayer 2022	≤101	101		H&E	WSI	512x512	Unclear	Learned	CNN features (ResNet18)	CNN Ensemble (ResNet18)	Patch	2 - Cancer, non-cancer	Monte Carlo cross-validation (10 reps) & external validation	41 WSIs from different centre	Accuracy per patient	56.3%-93.2%	Unclear plot (IQR & range)	Unclear plot	Unclear plot (IQR & Range)
	Salguero 2022	18	≥18		H&E	WSI	100x100	40x	Hand-crafted	Colour, texture, cellular morphology	SVM	Patch	2 - Cancer, non-cancer	Single train/test split		Accuracy	73%			
	Meng 2023	80	94		H&E	WSI	512x512	Unclear	Learned	CNN features (ResNet50)	CNN (novel STT-BOX)	WSI	2 - Malignant, benign	3-fold cross-validation (non-ovarian) & external validation (ovarian)	50 WSIs from 30 patients	AUC per subtype	0.9815-0.9953		0.8883	
Histological Subtype	BenTaieb 2015	80	80		H&E	WSI	Unclear	20x, 90x	Learned	CNN features (deconvolution network)	SVM	WSI	5 - HGSC, LGSC, CCC, MC, EC	Monte Carlo cross-validation (3 reps)		Accuracy	91.0%	±1.0% (unclear)		
																AUC	0.86			
	BenTaieb 2016	80	80		H&E	WSI	500x500	20x, 40x	Hand-crafted	Colour, texture, cellular morphology, cytology	SVM	WSI	5 - HGSC, LGSC, CCC, MC, EC	Leave-one-patient-out cross-validation (5 reps)		Accuracy	95.0%	±1.5% (one SD)		
	BenTaieb 2017	133	133		H&E	WSI	500x500	4x, 10x, 20x, 40x	Learned	CNN features (novel K-means)	SVM	WSI	5 - HGSC, LGSC, CCC, MC, EC	Single train/test split		Accuracy	90%			
	Levine 2020	≤406	406		H&E	WSI	256x256	40x	Learned	CNN features (VGG19)	CNN (VGG19)	Patch	5 - HGSC, LGSC, CCC, MC, EC	Monte Carlo cross-validation (10 reps)		Accuracy	70.87%	±6.35% (one SD)		
																Balanced accuracy	75.15%	±10.44% (one SD)		
																AUC	0.9177	±2.30% (one SD)		
	Boschman 2022	160	308		H&E	WSI	256x256	20x	Learned	CNN features (ResNet18)	CNN (ResNet18)	WSI	5 - HGSC, LGSC, CCC, MC, EC	External validation	60 WSIs from different centre	AUC	0.97	Unclear plot (unclear)	0.94	Unclear plot (unclear)
	Farahani 2022	485	948		H&E	WSI	512x512	20x	Learned	CNN features (VGG19)	CNN (VGG19)	WSI	5 - HGSC, LGSC, CCC, MC, EC	3-fold cross-validation & external validation	60 WSIs from different centre	Balanced accuracy	81.38%		80.97%	
																AUC	0.9475		0.9469	
	Kasture 2022	≤500	500		H&E	227x227	NA	20x	Learned	CNN features (novel KK-Net)	CNN (novel KK-Net)	Patch	5 - Serous, MC, CCC, EC, Non-cancer	10-fold cross-validation		Accuracy	91%			
																AUC	0.95			
Primary Cancer Type	Kalra 2020	933	1039		H&E	WSI	1000x1000	20x	Learned	NNs features (unclear architectures)	Yottixel Search	WSI	4 - Ovarian, uterine carcinosarcoma, uterine endometrial, cervical (FFPE slides)	Leave-one-patient-out cross-validation		Accuracy (Ovarian)	66.98%			
	Kalra 2020	1450	2216		H&E	WSI	1000x1000	20x	Learned	NNs features (unclear architectures)	Yottixel Search	WSI	4 - Ovarian, uterine carcinosarcoma, uterine endometrial, cervical (frozen slides)	Leave-one-patient-out cross-validation		Accuracy (Ovarian)	98.98%			
	Kalra 2020	9,484	11,561		H&E	WSI	1000x1000	20x	Learned	NNs features (unclear architectures)	Yottixel Search	WSI	13 - Gynaecological, brain, pulmonary, prostate/testis, breast, … (FFPE slides)	Leave-one-patient-out cross-validation		Accuracy (Gynaecological)	68.86%			
	Kalra 2020	10,571	14,887		H&E	WSI	1000x1000	20x	Learned	NNs features (unclear architectures)	Yottixel Search	WSI	13 - Gynaecological, brain, pulmonary, prostate/testis, breast, … (frozen slides)	Leave-one-patient-out cross-validation		Accuracy (Gynaecological)	66.89%			
	Ramasamy 2023	≤776	776		Unclear	WSI	Unclear	Unclear	Learned	CNN features (novel architecture)	CNN (novel)	WSI	2 - ovarian cancer, non-ovarian cancer	5-fold cross-validation		Accuracy	99.2%			
Genetic Mutation Status	Zeng 2021	229	≥229		H&E	WSI	1000x1000	Unclear	Hand-crafted	Texture, cellular and nuclear morphology	Random Forest	Patient	2 - BRCA1 Mutated, not mutated	Single train/test split		AUC	0.952			
	Zeng 2021	229	≥229		H&E	WSI	1000x1000	Unclear	Hand-crafted	Texture, cellular and nuclear morphology	Random Forest	Patient	2 - BRCA2 Mutated, not mutated	Single train/test split		AUC	0.912			
	Nero 2022	664	664		H&E	WSI	256x256	Unclear	Learned	CNN features (ResNet50)	CNN (CLAM)	WSI	2 - BRCA1/2 Mutated, wild-type	Single train/test split		AUC	0.59			
	Ho 2023	609	609		H&E	WSI	224x224	5x	Learned	CNN features (ResNet182)	CNN (ResNet182)	WSI	2 - BRCA1/2 Mutated, wild-type	Single train/test split		AUC	0.43			
Tumour-Stroma Reaction	Jiang 2022	≤306	≤306		H&E	WSI	256x256	Unclear	Learned	CNN features (Mask-RCNN)	CNN (VGG16)	Patch	3 - Low, intermediate, high (fibrosis score)	Single train/test split		Sensitivity per class	0.91-0.93			
	Jiang 2022	≤306	≤306		H&E	WSI	256x256	Unclear	Learned	CNN features (Mask-RCNN)	CNN (VGG16)	Patch	3 - Low, intermediate, high (cellularity score)	Single train/test split		Sensitivity per class	0.79-0.95			
	Jiang 2022	≤306	≤306		H&E	WSI	256x256	Unclear	Learned	CNN features (Mask-RCNN)	CNN (VGG16)	Patch	3 - Low, intermediate, high (orientation score)	Single train/test split		Sensitivity per class	0.74-0.95			
Grade	Poruthoor 2013	387	≥387		H&E	WSI	512x512	Unclear	Hand-crafted	Colour, texture, cellular and nuclear morphology	SVM	WSI	2 - Grade 1-2, Grade 3-4	Monte Carlo cross-validation (15 reps)		Accuracy	88%	Unclear plot (one SD)		
	Yu 2020	570	≤1358		H&E	WSI	Unclear	Unclear	Learned	CNN features (VGG16)	CNN (VGG16)	WSI	2 - Low-to-moderate, high	Monte Carlo cross-validation (3 reps)		AUC	0.812	±0.088 (unclear)		

Table 3.4 Characteristics of the 80 models of interest from the 45 papers included in this systematic review, grouped by model outcome. *Other data types are Genomics (G), Proteomics (P), Radiomics (R), and Transcriptomics (T). TMA refers to individual cores from tissue microarrays, WSI refers to whole slide images of biopsy or resection specimens. Continued on the next two pages.

Transcriptomic Subtype	Yu 2020	553	≤1341		H&E	WSI	Unclear	Unclear	Learned	CNN features (VGG16)	CNN (VGG16)	WSI	4 - Proliferative, differentiated, immunoreactive, mesenchymal	5-fold cross-validation		p-value	<0.0001			
	Zeng 2021	229	≥229		H&E	WSI	1000x1000	Unclear	Hand-crafted	Texture, cellular and nuclear morphology	Random Forest	Patient	4 - Proliferative, differentiated, immunoreactive, mesenchymal	Single train/test split		AUC per class	0.918-0.961			
Stage	Ghoniem 2021	587	587	G	H&E	WSI	224x224	Unclear	Learned	CNN features (altered VGG16)	CNN (altered VGG16)	WSI	5 - I, II, III, IV, Not available	5-fold cross-validation (20 reps)		Accuracy	98.87%			
Microsatellite Instability	Zeng 2021	229	≥229		H&E	WSI	1000x1000	Unclear	Hand-crafted	Texture, cellular and nuclear morphology	Random Forest	Patient	3 - High instability, stable, NA	Single train/test split		AUC (High instability)	0.919			
																AUC (Stable)	0.924			
Epithelial-Mesenchymal Transition Status	Hu 2022	≤70	70		H&E	WSI	256x256	40x	Learned	CNN features (ResNet18)	CNN (novel adInter-MIL)	WSI	2 - High, low	Monte Carlo cross-validation (10 reps)		Balanced accuracy	85.45%	±0.48% (variance)		
																AUC	0.7455	±0.0043 (variance)		
HRD Status	Lazard 2022	≤90	90		H&E	WSI	224x224	20x	Learned	CNN features (ResNet18)	NN	WSI	2 - Homologous Recombination Deficient, Proficient	Unclear		AUC	0.73			
Prognosis Outcome	**Publication**	**Internal Participants**	**Internal Pathology Images**	**Other Data***	**Stain Type**	**Original Image Size**	**Patch Size (Pixels)**	**Magnifications**	**Feature Extraction**	**Histopathological Features**	**Final Model**	**Prediction Precision**	**Classes**	**Validation Type**	**External Validation Data**	**Metric**	**Internal Results**	**Internal Variability (Measure)**	**External Results**	**External Variability (Measure)**
Treatment Response	Yaar 2020	220	≥220	G	H&E	WSI	512x512	20x	Learned	CNN features (novel)	CNN	WSI	2 - Chemo-resistant, chemo-sensitive	5-fold cross-validation		AUC	0.79	±0.07 (one SD)		
	Yu 2020	277	≤1065		H&E	WSI	Unclear	Unclear	Learned	CNN features (VGG16)	CNN (VGG16)	WSI	2 - Early relapse, late relapse	5-fold cross-validation		p-value	0.003			
	Wang 2022(a)	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (modified Inception V3)	CNN (modified Inception V3)	TMA	2 - Effective, invalid (AIM2 stain)	5-fold cross-validation		AUC	0.91	±0.05 (unclear)		
	Wang 2022(a)	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (modified Inception V3)	CNN (modified Inception V3)	TMA	2 - Effective, invalid (C3 stain)	5-fold cross-validation		AUC	0.78	±0.12 (unclear)		
	Wang 2022(a)	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (modified Inception V3)	CNN (modified Inception V3)	TMA	2 - Effective, invalid (C5 stain)	5-fold cross-validation		AUC	0.66	±0.07 (unclear)		
	Wang 2022(a)	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (modified Inception V3)	CNN (modified Inception V3)	TMA	2 - Effective, invalid (NLRP3 stain)	5-fold cross-validation		AUC	0.55	±0.08 (unclear)		
	Wang 2022(b)	78	288		H&E	WSI	512x512	Unclear (multiple)	Learned	CNN features (Inception V3)	CNN (Inception V3)	WSI	2 - Effective, invalid	5-fold cross-validation & external validation	175 TMAs from 71 patients	Accuracy	88.2%	±6% (unclear)	77.5%	
	Wang 2023	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (novel)	CNN (InceptionV3)	TMA	2 - Effective, invalid (PKM2 stain)	5-fold cross-validation		AUC	0.99	±0.01 (unclear)		
	Wang 2023	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (novel)	CNN (InceptionV3)	TMA	2 - Effective, invalid (Ang-2 stain)	5-fold cross-validation		AUC	1.00	±0.01 (unclear)		
	Wang 2023	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (novel)	CNN (InceptionV3)	TMA	2 - Effective, invalid (VEGF stain)	5-fold cross-validation		AUC	0.89	±0.08 (unclear)		
	Wang 2023	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (novel)	CNN Ensemble (InceptionV3)	TMA	2 - Effective, invalid (PKM2+Ang-2 stain)	5-fold cross-validation		AUC	1.00	±0.00 (unclear)		
Overall Survival	Poruthoor 2013	382	≥382	G,P	H&E	WSI	512x512	Unclear	Hand-crafted	Colour, texture, cellular and nuclear morphology	SVM	WSI	2 - <5 years, ≥5 years	Monte Carlo cross-validation (15 reps)		Accuracy	55%	Unclear plot (one SD)		
	Zeng 2021	229	≥229	G,P,T	H&E	WSI	1000x1000	Unclear	Hand-crafted	Texture, cellular and nuclear morphology	Random Forest	Patient	2 - High risk, low risk	External validation	TMAs from 92 patients	Hazard ratio	18.23	10.34-34.38 (95% CI)		
																p-value	<0.0001		0.0097	
	Boehm 2022	444	≥283	R	H&E	WSI	128x128	Unclear	Hand-crafted	Colour, texture, cellular and nuclear morphology	Cox model	WSI	Risk score	Single train/test split		C-index	0.61	0.594-0.625 (95% CI)		
	Liu 2022	583	1296		H&E	WSI	512x512	20x	Learned	CNN features (novel DeepConvAttentionSurv)	CNN (novel DCAS)	Patient	Risk score	Single train/test split		C-index	0.98	±0.0085 (unclear)		
	Yokomizo 2022	110	≥110		H&E	WSI	255x255	Unclear	Learned	CNN features (ResNet34)	CNN (ResNet34)	TMA	2 - Short, long	Monte Carlo cross-validation (10 reps)		AUC	"~0.95"			
	Wu 2023	90	90		H&E	WSI	256x256	Unclear	Learned	CNN features (ResNet50)	CNN (CLAM)	TMA	Risk score	5-fold cross-validation		C-index	0.5789	0.5096-0.6053 (CV range)		
Progression-free Survival	Laury 2021	52	227		H&E	WSI	Unclear	Unclear	Learned	CNN features (unclear architecture)	NN	WSI	2 - <6 months, >18 months	Single train/test split		Accuracy	82%			
	Boehm 2022	422	≥261	G,R	H&E	WSI	128x128	Unclear	Hand-crafted	Colour, texture, cellular and nuclear morphology	Cox model	WSI	2 - High, low	Single train/test split		p-value	0.040			
	Yokomizo 2022	110	≥110		H&E	WSI	255x255	Unclear	Learned	CNN features (ResNet34)	CNN (ResNet34)	TMA	2 - Short, long	Monte Carlo cross-validation (10 reps)		AUC	0.98			

Table 3.5 Model characteristics continued - diagnostic and prognostic outcomes.

Other Outcome	Publication	Internal Participants	Internal Pathology Images	Other Data*	Stain Type	Original Image Size	Patch Size (Pixels)	Magnifications	Feature Extraction	Histopathological Features	Final Model	Prediction Precision	Classes	Validation Type	External Validation Data	Metric	Internal Results	Internal Variability (Measure)	External Results	External Variability (Measure)
Stain Quantity/Intensity	Gentles 2021	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	ATM stain H-score (0-18)	Single test set		R^2	0.1833			
	Gentles 2021	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	ATR stain H-score (0-18)	Single test set		R^2	0.4330			
	Gentles 2021	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	DNAPKcs stain H-score (0-18)	Single test set		R^2	0.6296			
	Gentles 2021	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	Ku70 stain H-score (0-18)	Single test set		R^2	0.5891			
	Gentles 2021	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	PAR stain H-score (0-18)	Single test set		R^2	0.3978			
	Gentles 2021	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	RPA stain H-score (0-18)	Single test set		R^2	0.4453			
	Elie 2022	25	25		IHC	WSI	Unclear	20x	Hand-crafted	Colour, texture	Gaussian Mixture Model	Patch	3 - Mcl-1 high, medium, low	None		Accuracy per patient	96.94%-99.51%			
	Elie 2022	25	25		IHC	WSI	Unclear	20x	Hand-crafted	Colour, texture	Gaussian Mixture Model	Patch	3 - Bim high, medium, low	None		Accuracy per patient	92.77%-95.75%			
	Elie 2022	25	25		IHC	WSI	Unclear	20x	Hand-crafted	Colour, texture	Gaussian Mixture Model	Patch	3 - P-ERK high, medium, low	None		Accuracy per patient	89.08%-100%			
Tumour Segmentation	Signolle 2010	Unclear	Unclear		IHC	WSI	2048x2048	20x	Hand-crafted	Texture	Hidden Markov Tree	Pixel	5 - Cancer, inflammatory stroma, loose connective tissue, cellular stroma, background	Single train/test split		Accuracy	71.50%	±12.83 (one SD)		
	Jiang 2022	306	306		H&E	WSI	256x256	Unclear	Learned	CNN features (Mask-RCNN)	CNN (Mask-RCNN)	Pixel	2 - Tumour, stroma	Single train/test split		Dice coefficient	93.5%	Unclear plot (unclear)		
	Kowalski 2022	≤26	26		H&E	1698x1242	100x200	Unclear	Learned	CNN features (novel architecture)	CNN (novel)	Pixel	2 - Cancer, healthy	Single train/test split		Accuracy	82%			
	Ho 2023	39	39		H&E	WSI	256x256	5x, 10x, 20x	Learned	CNN features (novel architecture)	CNN (novel DMMN)	Pixel	2 - Cancer, non-cancer	Single train/test split		Intersection over union	0.74			
	Ramasamy 2023	≤776	776		Unclear	WSI	Unclear	Unclear	Learned	CNN features (novel architecture)	CNN (novel)	Pixel	2 - Tumour, non-tumour	5-fold cross-validation		Dice coefficient	92%			
Stain Segmentation	Dong 2010(a)	1	1		IHC	Unclear	NA	Unclear	Hand-crafted	Colour	ISODATA clustering	Pixel	2 - Positive, Negative	None		Qualitative	"Satisfactory"			
	Dong 2010(b)	1	1		IHC	Unclear	NA	Unclear	Hand-crafted	Colour	OTSU thresholding	Pixel	2 - Positive, Negative	None		Qualitative	"Satisfactory"			
	Janowczyk 2012	100	≥500		IHC	TMA	NA	20x	Hand-crafted	Colour	HNCuts (novel)	Pixel	2 - Positive, Negative	Single test set		Sensitivity	59.24%	±7.36% (variance)		
																Specificity	99.01%	±0.56% (variance)		
Cell Type	Lorsakul 2017	≤45	45		IHC	WSI	Unclear	20x	Hand-crafted	Nuclear morphology	Random Forest	Cell	5 - Cancer, carcinoma-associated fibroblast, non-tumour, background, artifact	5-fold cross-validation		Accuracy	91.7%			
	Heindl 2018	514	514		H&E	WSI	2000x2000	Unclear	Hand-crafted	Texture, cellular morphology	SVM	Cell	3 - Cancer, stroma, lymphocyte	Single train/test split		Balanced accuracy per class	80.64%-85.05%			
Tissue Type	Du 2018	≤154	154		H&E	Unclear	60x60	Unclear	Learned	CNN features (GoogLeNet)	SVM	Superpixel	2 - Epithelium, stroma	Single train/test split		Accuracy	91.8%			
																AUC	0.974			
	Paijens 2021	268	268		IHC	TMA	Unclear	Unclear	Learned	NN features (unclear architecture)	NN	Pixel	2 - Epithelium, stroma	None		None	NA			

Table 3.6 Model characteristics continued - other outcomes.

3.3.1 Data in Included Literature

The number of participants in internal datasets varied by orders of magnitude, with each study including 1 to 776 ovarian cancer patients, and one study including over 10,000 total patients across a range of 32 malignancies [122]. Most research only used data from the five most common subtypes of ovarian carcinoma, though one recent study included the use of sex cord-stromal tumours [150]. Only one study explicitly included any prospective data collection, and this was only for a small subset which was not used for external validation [133].

As shown in Figure 3.3, the number of samples used was often much greater than the number of patients included, with three studies using over 1,000 samples from ovarian cancer patients [114, 125, 142]. In most studies, models were developed using WSIs containing resected or biopsied tissue (34/45), with others using individual tissue microarray (TMA) core images (5/45) or pre-cropped digital pathology images (3/45). Most studies used H&E-stained tissue (33/45) and others used a variety of IHC stains (11/45), with no two papers reporting the use of the same IHC stains. Some studies included multi-modal approaches, using genomics [115, 124, 127, 132, 133], proteomics [115, 132], transcriptomics [132], and radiomics [133] data alongside histopathological data.

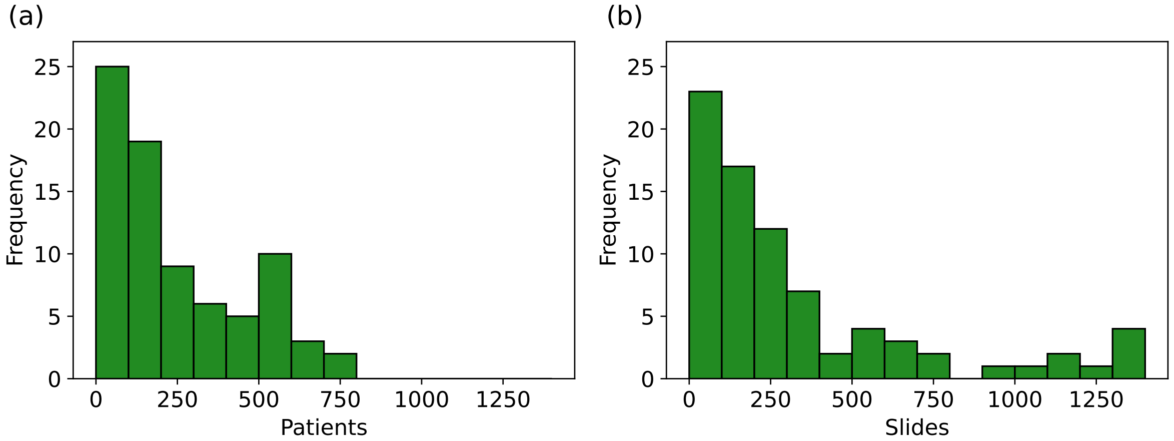


Figure 3.3 Histograms showing (a) the number of ovarian cancer patients and (b) the number of samples used in the development of each model. Many of these values are uncertain due to incomplete reporting, as reflected in Table 3.4.

The most commonly used data source was The Cancer Genome Atlas (TCGA) (18/45), a project from which over 30,000 digital pathology images from 33 malignancies are publicly available. The ovarian cancer subset, TCGA-OV [154], contains 1481 WSIs from 590 cases of ovarian serous carcinoma (mostly, but not exclusively, high-grade), with corresponding genomic, transcriptomic, and clinical data. This includes slides from eight data centres in the United States, with most slides containing frozen tissue sections (1374/1481) rather than FFPE sections. Other recurring data sources were the University of British Columbia Ovarian Cancer Research Program (OVCARE) repository [123, 134, 136], the Transcanadian study [116, 117], and clinical records at the Mayo Clinic [128, 138], Tri-Service General Hospital [146, 147, 152], and Memorial Sloan Kettering Cancer Center [133, 149]. All other researchers either used a unique data source (12/45) or did not report the provenance of their data (8/45). TCGA-OV, OVCARE, and the Transcanadian study are all multi-centre datasets. Aside from these, few studies reported the use of multi-centre data [130–133, 136, 143]. Only two studies reported the use of multiple slide scanners, with each slide scanned on one of two available scanners [134, 136]. The countries from which data were sourced included Canada, China, Finland, France, Germany, Italy, Japan, the Netherlands, South Korea, Taiwan, the United Kingdom, and the United States of America.

3.3.2 Methods in Included Literature

The 80 models of interest included 37 diagnostic models, 22 prognostic models, and 21 other models predicting diagnostically relevant information. Diagnostic model outcomes included the classification of malignancy status (10/37), histological subtype (7/37), primary cancer type (5/37), genetic mutation status (4/37), tumour-stroma reaction level (3/37), grade (2/37), transcriptomic subtype (2/37), stage (1/37), microsatellite instability status (1/37), epithelial-mesenchymal transition status (1/37), and homologous recombination deficiency status (1/37). Prognostic models included the prediction of treatment response (11/23), overall survival (6/23), progression-free survival (3/23), and recurrence (2/23). Other models performed tasks which could be used to assist pathologists in analysing pathology images, including measuring the quantity/intensity of staining (9/21), generating tumour (5/21) or stain (3/21) segmentation masks, and classifying tissue (2/21) or cell (2/21) types.

A variety of models were used, with the most common types being CNNs (41/80), SVMs (10/80), and random forests (6/80). CNN architectures included GoogLeNet [120], VGG16 [125, 138], VGG19 [123, 136], InceptionV3 [131, 146, 147, 152], ResNet18 [133, 134, 136, 137, 141, 143], ResNet34 [148], ResNet50 [144, 150, 153], ResNet182 [149], and MaskRCNN [138]. Novel CNNs typically used multiple standardised blocks involving convolutional, normalization, activation, and/or pooling layers [124, 139, 140], with two studies also including attention modules [142, 152]. One study generated their novel architecture by using a topology optimization approach on a standard VGG16 [127].

Most researchers split their original images into patches to be separately processed, with dimensions ranging from 60 to 2048 pixels, and the most common patch sizes being 512 x 512 pixels (19/56) and 256 x 256 pixels (12/56). A range of feature extraction techniques were employed, including both hand-crafted/pre-defined features (23/80) and features that were automatically learned by the model (51/80). Hand-crafted features included a plethora of textural, chromatic, and cellular and nuclear morphological features. Hand-crafted features were commonly used as inputs to classical ML methods, such as SVM and random forest models. Learned features were typically extracted using a CNN, which was often also used for classification.

Despite the common use of patches, most models made predictions at the WSI level (29/80), TMA core level (18/80), or patient level (6/80), requiring aggregation of patch-level information. The methods used for this aggregation may be referred to as MIL (Section 2.4.4), though few models of interest were reported using this terminology [124, 137, 141, 144]. Instance embedding approaches (Section 2.4.4) generated slide-level features using summation [150], averaging [115, 132, 149], attention-based weighted averaging [137, 141, 142, 144, 153], concatenation [117, 122], as well as more complex embedding approaches using Fisher vector encoding [116] and k-means clustering [118]. Instance classification approaches (Section 2.4.4) aggregated patch-level predictions by taking the maximum [124, 152], median [148], or average [127], using voting strategies [134, 147], or using a random forest classifier [136].

Most studies included segmentation at some stage, with many of these analysing tumour/stain segmentation as a model outcome [109–113, 130, 138, 140, 149, 151]. Some studies used segmentation to determine regions of interest for further modelling, either simply separating tissue from the background [114, 122, 144, 153], or using tumour segmentation to select the most relevant tissue regions [126, 129, 146, 147, 152]. One study also used segmentation to detect individual cells for classification [121]. Some studies used segmentation in determining features relating to the quantity and morphology of different tissues, cells, and nuclei [114, 115, 117, 128, 132, 133].

While attention-based approaches have been applied to other malignancies for several years [77, 78], they were only seen in the most recent ovarian cancer studies [136, 137, 141, 142, 144, 146, 147, 152, 153], and none of the methods included self-attention (Section 2.4.5), an increasingly popular method for other malignancies [155]. Most models were deterministic (in that they used fixed weights and would generate the same output if the same image were to be input multiple times), though hidden Markov trees [111], probabilistic boosting trees [112], and Gaussian mixture models [135] were also used.

Aside from the common use of low-resolution images to detect and remove non-tissue areas, images were typically analysed at a single resolution, with only six papers exploring multi-magnification techniques which may better leverage both cellular-level and broader tissue-level features. Four of these combined features from different resolutions for modelling [116–118, 149], and the other two used different magnifications for selecting informative tissue regions and for modelling [146, 147]. Out of the papers for which it could be determined, the most common modelling magnifications were 20x (35/41) and 40x (7/41). Few models integrated histopathology data with other modalities (6/80). Multi-modal approaches included the concatenation of separately extracted uni-modal features before modelling [115, 127, 132], the aggregation of unimodal predictions from separate models [133], and a teacher-student approach where multiple modalities were used in model training but only histopathology data was used for prediction [124].

3.3.3 Analyses in Included Literature

Analyses were limited, with less than half of all models being evaluated with cross-validation (39/80) and with very few externally validated using independent ovarian cancer data (7/80), despite small internal cohort sizes. Cross-validation methods included k-fold (22/39) with 3 to 10 folds, Monte Carlo (12/39) with 3 to 15 repeats, and leave-one-patient-out cross-validations (5/39). Some other papers included cross-validation on the training set to select hyperparameters but used only a small unseen test set from the same data source for evaluation. Externally validated models were all trained with WSIs, with validations either performed on TMA cores (2/7) or WSIs from independent data sources (5/7), with two of these explicitly using different scanners to digitize internal and external data [134, 136]. Some reported methods were externally validated with data from non-ovarian malignancies, but none of these included ovarian cancer data in any capacity so they were not included in the review. However, there was one method which trained with only gastrointestinal tumour data and externally validated with ovarian tumour data [150].

Most classification models were evaluated using accuracy, balanced accuracy, and/or AUROC, with one exception where only a p-value was reported measuring the association between histological features and transcriptomic subtypes based on a Kruskal-Wallis test [125]. Some models were also evaluated using the F1 score, which was not tabulated (in Table 3.4) as the other metrics were reported more consistently. Survival model performance was typically reported using AUROC, with other metrics including p-value, accuracy, hazard ratios, and C-index, which is similar to AUROC but can account for censoring. Segmentation models were almost all evaluated differently from each other, with different studies reporting AUROC, accuracy, Dice coefficient, intersection over union, sensitivity, specificity, and qualitative evaluations. Regression models were all evaluated using the coefficient of determination (R^2-statistic). For some models, performance was broken down per patient [135, 143], subtype [150], or class [121, 122, 132, 138], without an aggregated, holistic measure of model performance.

The variability of model performance was not frequently reported (33/94), and when it was reported it was often incomplete. This included cases where it was unclear what the intervals represented (95% confidence interval, one standard deviation, variance, etc.), or not clear what the exact bounds of the interval were due to results being plotted but not explicitly stated. Within the entire review, there were only three examples in which variability was reported during external validation [131, 134, 143], only one of which clearly reported both the bounds and the type of the interval [131]. No studies performed any Bayesian form of uncertainty quantification. Reported results are shown in Table 3.4, though direct comparisons between the performance of different models should be treated with caution due to the diversity of data and validation methods used to evaluate different models, the lack of variability measures, the consistently high risks of bias (Section 3.4.2), and the heterogeneity in reported metrics.

3.4 Risk of Bias Assessment

3.4.1 PROBAST Assessment Tool

The *risk of bias* in research is the chance of reported results being distorted by limitations within the study design, conduct, and analysis. The risks of bias of each publication in the review were assessed using the Prediction model Risk Of Bias ASsessment Tool (PROBAST) [156]. This includes 20 guiding questions which are categorised into four domains (participants, predictors, outcome, and analysis), which are summarised as either *high-risk* or *low-risk*, or *unclear* if there is insufficient information to make a comprehensive assessment and none of the available information indicates a high risk of bias. As such, an unclear risk of bias does not indicate methodological flaws, but incomplete reporting.

The **participants** domain covers the recruitment and selection of participants to ensure the study population is consistent and representative of the target population. Relevant details include the participant recruitment strategy (when and where participants were recruited), the inclusion criteria, and how many participants were recruited.

The **predictors** domain covers the consistent definition and measurement of predictors, which in this field typically refers to the generation of digital pathology images. This includes methods for fixing, staining, scanning, and digitally processing tissue samples before modelling.

The **outcome** domain covers the appropriate definition and consistent determination of ground-truth labels. This includes the criteria used to determine diagnosis/prognosis, the expertise of any persons determining these labels, and whether labels are determined independently of any model outputs.

The **analysis** domain covers statistical considerations in the evaluation of model performance to ensure valid and not unduly optimistic results. This includes many factors, such as the number of participants in the test set with each outcome, the validation approaches used (cross-validation, external validation, bootstrapping, etc.), the metrics used to assess performance, and methods used to overcome the effects of censoring, competing risks/confounders, and missing data.

Different risk factors may be interrelated, for example, the risk of bias from using a small dataset is somewhat mitigated by cross-validation, which increases the effective size of the test set to match the size of the full dataset and can be used to assess variability, reducing optimism in the results. Further, the risk caused by using a small dataset depends on the type of outcome being predicted, for example, more data is required for a robust analysis of 5-class classification than binary classification. There must also be sufficient data within all relevant patient subgroups, for example, if multiple subtypes of ovarian cancer are included, there must not be a subtype that is only represented by a few patients. Due to these interrelated factors, there are no strict criteria to determine the appropriate size of a dataset, though fewer than 50 samples per class or fewer than 100 samples overall is likely to be considered high-risk, and more than 1000 samples overall is likely to be considered low-risk.

Risks of bias often arise due to inconsistent methodologies. Inconsistency in the participants and predictors domains may cause heterogeneity in the visual properties of digital pathology slides which may lead to spurious correlations, either through random chance or systematic differences between subgroups in the dataset. Varied data may be beneficial during training to improve model generalisability when using large datasets, though this must be closely controlled to avoid introducing systematic confounding. Inconsistent determination of the outcome can mean that the results of a study are unreliable due to spurious correlations in the ground truth labels, or invalid due to incorrect determination of labels.

While PROBAST provides a framework to assess risks of bias, there is some level of subjectivity in the interpretation of signalling questions. As such, each model was analysed by three independent researchers (any of JB, KA, NR, KZ, NMO), with at least one computer scientist and one clinician involved in the assessment of each model. The PROBAST applicability of research analysis was not implemented as it is unsuitable for such a diverse array of possible research questions.

3.4.2 Risk of Bias Results

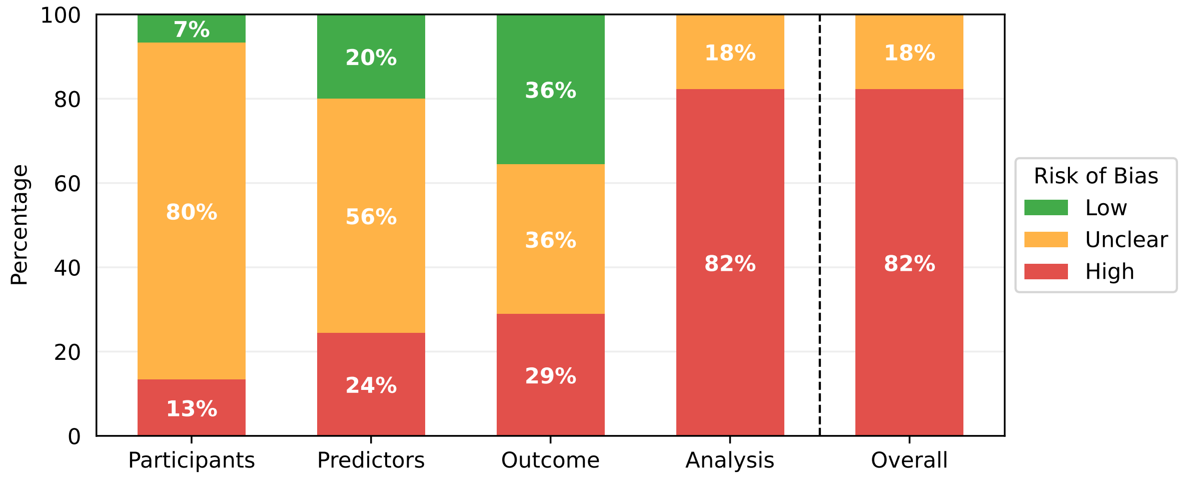


Figure 3.4 PROBAST risk of bias results summarised for the 45 studies included in the systematic literature review.

The results of the PROBAST assessments are shown in Table 3.7. While some studies contained multiple models of interest, none of these contained models with different risk of bias scores for any section of the PROBAST assessment, so one risk of bias analysis is presented per paper. All studies showed either a high overall risk of bias (37/45) or an unclear overall risk of bias (8/45). Every high-risk study had a high-risk score in the analysis section (37/45), with several also being at a high risk of bias in the participants (6/45), predictors (11/45), or outcome (13/45) sections. Less than half of all studies achieved a low risk of bias in any domain (21/45), with most low risks being found in the outcome (16/45) and predictors (9/45) sections. Nearly all of the papers had an unclear risk of bias in at least one domain, most commonly the participants (36/45) and predictors (25/45) domains. Quantitative summaries are presented in Figure 3.4.

Publication	Participants	Predictors	Outcome	Analysis	Overall
Dong 2010(a) [109]	High	High	High	High	**High**
Dong 2010(b) [110]	High	High	High	High	**High**
Signolle 2010 [111]	Unclear	Unclear	High	High	**High**
Janowczyk 2011 [112]	Unclear	Unclear	Low	High	**High**
Janowczyk 2012 [113]	Unclear	High	Unclear	High	**High**
Kothari 2012 [114]	Unclear	Low	Low	Unclear	**Unclear**
Poruthoor 2013 [115]	Unclear	High	High	High	**High**
BenTaieb 2015 [116]	Unclear	Unclear	Low	High	**High**
BenTaieb 2016 [117]	Unclear	High	Unclear	High	**High**
BenTaieb 2017 [118]	Unclear	Unclear	Low	High	**High**
Lorsakul 2017 [119]	Unclear	Unclear	High	High	**High**
Du 2018 [120]	Unclear	Unclear	Unclear	Unclear	**Unclear**
Heindl 2018 [121]	Unclear	Low	Low	High	**High**
Kalra 2020 [122]	Unclear	Low	Low	High	**High**
Levine 2020 [123]	Unclear	Low	Low	Unclear	**Unclear**
Yaar 2020 [124]	Unclear	Unclear	Low	High	**High**
Yu 2020 [125]	Unclear	Low	Low	High	**High**
Gentles 2021 [126]	High	Unclear	High	High	**High**
Ghoniem 2021 [127]	Unclear	Unclear	Unclear	High	**High**
Jiang 2021 [128]	High	High	Unclear	High	**High**
Laury 2021 [129]	Low	High	High	High	**High**
Paijens 2021 [130]	Low	High	Unclear	High	**High**
Shin 2021 [131]	Unclear	Unclear	Unclear	High	**High**
Zeng 2021 [132]	Unclear	Unclear	Low	High	**High**
Boehm 2022 [133]	Unclear	High	Unclear	High	**High**
Boschman 2022 [134]	Unclear	Low	Low	High	**High**
Elie 2022 [135]	Unclear	Low	High	High	**High**
Farahani 2022 [136]	Unclear	Unclear	Low	Unclear	**Unclear**
Hu 2022 [137]	Unclear	Unclear	Unclear	Unclear	**Unclear**
Jiang 2022 [138]	Unclear	Unclear	High	High	**High**
Kasture 2022 [139]	High	High	High	High	**High**
Kowalski 2022 [140]	Unclear	Unclear	Unclear	High	**High**
Lazard 2022 [141]	Unclear	Unclear	Unclear	Unclear	**Unclear**
Liu 2022 [142]	Unclear	Unclear	Unclear	Unclear	**Unclear**
Mayer 2022 [143]	Unclear	Unclear	High	High	**High**
Nero 2022 [144]	Unclear	Low	High	High	**High**
Salguero 2022 [145]	Unclear	Unclear	Low	High	**High**
Wang 2022(a) [146]	Unclear	Unclear	Unclear	High	**High**
Wang 2022(b) [147]	Unclear	Unclear	Low	High	**High**
Yokomizo 2022 [148]	Low	Low	Unclear	Unclear	**Unclear**
Ho 2023 [149]	Unclear	Unclear	Unclear	High	**High**
Meng 2023 [150]	Unclear	Unclear	Low	High	**High**
Ramasamy 2023 [151]	High	High	High	High	**High**
Wang 2023 [152]	Unclear	Unclear	Unclear	High	**High**
Wu 2023 [153]	Unclear	Unclear	Low	High	**High**

Table 3.7 PROBAST risk of bias assessment results for the 45 studies in the review. This is presented as one row per study because every study that contained multiple models of interest was found to have the same risk of bias score for each model.

3.5 Discussion

The vast majority of published research on AI for diagnostic or prognostic purposes in ovarian cancer histopathology was found to be at a high risk of bias due to issues within the analyses performed. Researchers often used a limited quantity of data and conducted analyses on a single train-test data split without using any methods to account for overfitting and model optimism (cross-validation, bootstrapping, external validation). These limitations are common in gynaecological AI research using other data types, with recent reviews pointing to poor clinical utility caused by predominantly retrospective studies using limited data [107, 157] and limited methodologies with weak validation, which risk model performance being overestimated [105, 106].

The more robust analyses included one study in which several relevant metrics were evaluated using 10 repeats of Monte Carlo cross-validation on a set of 406 WSIs, with standard deviations reported for each metric [123]. Other positive examples performed both internal cross-validation and external validation for the same outcome, giving a more rigorous analysis [136, 143, 147]. While external validations were uncommon, those which were conducted offered a real insight into model generalisability, with a clear reduction in performance on all external validation sets except one [136]. The only study which demonstrated high generalisability included the largest training set out of all externally validated approaches, included more extensive data labelling than many similar studies, and implemented a combination of three colour normalisation approaches, indicating that these factors may benefit generalisability.

Studies frequently had an unclear risk of bias within the participants and predictors domains of PROBAST due to incomplete reporting. Frequently missing information included where the patients were recruited, how many patients were included, how many samples/images were used, whether any patients/images were excluded, and the methods by which tissue was processed and digitized. Only three papers were found to be at low risk of bias for participants, with these including clear and reasonable patient recruitment strategies and selection criteria, which can be seen as positive examples for other researchers [129, 130, 148]. Information about the predictors (histopathology images and features derived thereof) was generally better reported, but still often missed key details which meant that it was unclear whether all tissue samples had been processed similarly to avoid risks of bias from visual heterogeneity.

When patient characteristics were reported they often showed a high risk of bias. Many studies included very small quantities of patients with specific differences from the majority (e.g. less than 20 patients with a different cancer subtype to the majority), causing a risk of spurious correlations and results which were not generalisable to the wider population.

Reporting was particularly sparse in studies which only used openly accessible data, possibly indicating that AI-focused researchers were not taking sufficient time to understand these datasets and ensure their research was clinically relevant. For example, many of the researchers who used TCGA data included frozen tissue sections without commenting on whether this was appropriate, even though pathologists do not consider them to be of optimal diagnostic quality. One paper handled TCGA data more appropriately, with a clear explanation of the positives and negatives of the dataset, and entirely separate models for FFPE and frozen slides [122].

Sharing code can help to mitigate the effects of incomplete reporting and drastically improve reproducibility, but only 19 of the 45 papers did this, with some of these appearing to be incomplete or inaccessible. The better code repositories included detailed documentation to aid reproducibility, including environment set-up information [125, 150], overviews of included functions [133, 141, 149], and code examples used to generate reported results [121].

Two papers were found to have major discrepancies between the reported data and the study design, indicating much greater risks of bias than those seen in any other research [139, 151]. In one paper [139], it was reported that TCGA-OV data was used for subtyping with 5 classes, despite this dataset only including high-grade serous and low-grade serous carcinomas. In the other paper [151], it was reported that TCGA-OV data was used for slide-level classification into ovarian cancer and non-ovarian cancer classes using PAS-stained tissue, despite TCGA-OV only containing H&E-stained ovarian cancer slides. In the former paper [139], it was notable that some of the images included in the manuscript and shared data file contained watermarks and copyright information indicating non-TCGA data sources (such as https://www.webpathology.com/), and further analysis showed that many images used in the paper had been cropped to remove copyright symbols which were present in WebPathology. In the latter paper [151], the manuscript contained many figures which were indistinguishable from those in several disparate studies (without references), including a pathology image from a bat ovary [158] and performance graphs from the training of an MRI brain tumour detection model [159]. In both of these cases, concerns were reported to the editors of the journals in which the papers had been published.

3.5.1 Limitations of the Review

While the systematic review protocol was designed to reduce biases and maximise the quantity of relevant research included, there were some limitations. The review was restricted to published literature in the English language, however, AI research may be published in other languages or made available as pre-prints without publication in peer-reviewed journals, making this review incomplete. While most of the review process was completed by multiple independent researchers, duplicate detection was performed by a single researcher, raising the possibility of errors in this step of the review process which would have resulted in incorrect exclusions. Due to the significant time gap between the initial and final literature searches (approximately 12 months), there may have been inconsistencies in interpretations, both for data extraction and risk of bias assessments. Finally, this review focused only on conventional light microscopy images of human histopathology samples relating to ovarian cancer, so may have overlooked useful literature outside of this domain.

3.5.2 Development of the Field

The field of AI in ovarian cancer histopathology diagnosis is rapidly growing, with more research published since the start of 2020 than in all preceding years combined. The earliest research, published in 2010-2013, used hand-crafted features to train classical ML methods such as SVMs. These models were used for segmentation [109–111, 113], malignancy classification [112, 114], grading [115], and overall survival prediction [115]. Most of these early studies focused on IHC-stained tissue (5/7), which would be much less commonly used in subsequent research (6/38).

The field was relatively dormant in the following years, with only 6 papers published between 2014-2019, half of which had the same primary author [116–118]. These models still used traditional ML classifiers, though some used automatically learned features rather than the traditional hand-crafted features. The models developed were used for histological subtyping [116–118] and cellular/tissue classification [119–121].

Since 2020 there has been a much greater volume of research published, most of which has involved the use of deep neural networks for automatic feature extraction and classification. Recent research has investigated a broader array of diagnostic outcomes, including the classification of primary cancer type [122, 151], mutation status [132, 144, 149], homologous recombination deficiency status [141], tumour-stroma reaction level [138], transcriptomic subtypes [125, 132], microsatellite instability [132], and epithelial-mesenchymal transition status [137]. Three additional prognostic outcomes have also been predicted in more recent literature - progression-free survival [129, 133, 148], relapse [144, 148], and treatment response [124, 125, 146, 147, 152].

Despite progress within a few specific outcomes, there was no obvious overall trend in the sizes of datasets used over time, either in terms of the number of slides or the number of participants (Figure 3.5). Similarly, there was no evidence that recent research included more rigorous internal validations, though external validations have been increasing in frequency - no research before 2021 included any external validation with ovarian cancer data, but seven studies published more recently did [131, 132, 134, 136, 143, 147, 150]. While these external validations were typically limited to small quantities of data, the inclusion of any external validation demonstrates progress from previous research. Such validations are essential to the clinical utility of these models as real-world implementation will require robustness to different sources of visual heterogeneity, with variation occurring across different data centres and within data centres over time. As this field continues to mature, researchers must conduct thorough validations with larger, high-quality independent datasets, including clearly reported protocols for patient recruitment and selection, pathology slide creation, and digitization. This will help to reduce the biases, limited reproducibility, and limited generalisability identified in most of the existing research in this domain.

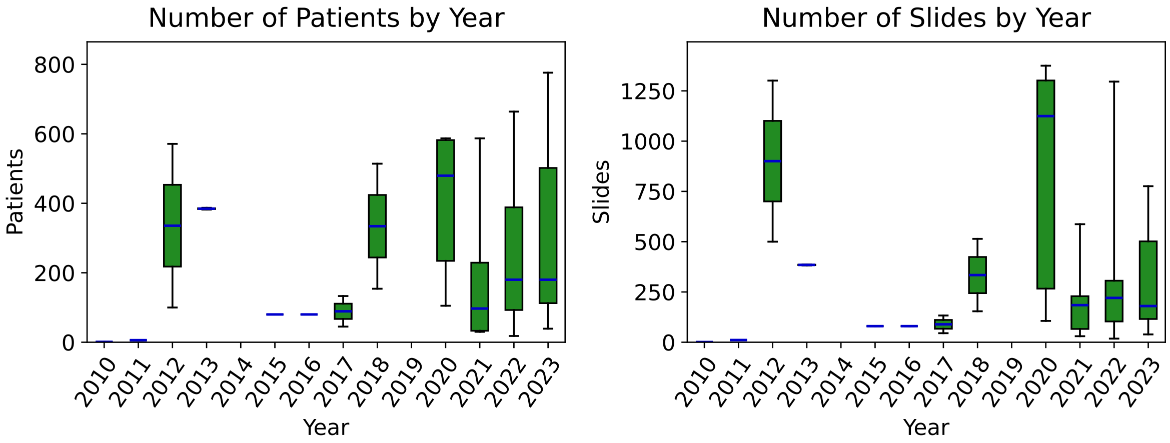


Figure 3.5 Boxplots showing the number of ovarian cancer patients and the number of samples used in the development of each model by publication year. Many of these values are uncertain due to incomplete reporting, as reflected in Table 3.4. There were no accepted studies in 2014 or 2019.

3.5.3 Recent Literature

To find the most recent research, systematic searches were repeated on 25/06/2024, and ad-hoc searches were conducted in Google Scholar, ResearchGate, and the references of other relevant papers. Since the publication of the systematic review, the field has continued to grow, with over 30 new papers identified and two ovarian cancer histopathology challenges conducted.

Histological Subtyping

Histological subtyping has remained one of the most common tasks in recent months. The Ovarian Cancer subtypE clAssification and outlier detectioN (OCEAN) [160] challenge focused on the robustness of ovarian cancer subtyping, with data from over 20 centres, including both WSIs and TMAs, and an *other* class alongside the five most common ovarian carcinoma subtypes. Across training and testing sets, the challenge utilised 1006 WSIs and 1462 TMAs, one of the largest datasets ever used in ovarian cancer subtyping. The greatest performance in this challenge was 66% balanced accuracy, indicating the difficulty in handling such diverse data.

Unfortunately, most of the recently published research on histological subtyping has either used the OCEAN dataset while it was still under embargo [161, 162] or the previously mentioned dataset [139] which had been misrepresented as TCGA-OV data [163–167]. Multiple versions of the latter dataset have been removed from the data hosting website *Mendeley* in response to our concerns, with Mendeley citing *suspected copyright infringement* (https://data.mendeley.com/datasets/kztymsrjx9/1 and https://data.mendeley.com/datasets/w39zgksp6n/1).

Some other recently published studies have also been of questionable value. One study used a particularly small set of ovarian cancer images to validate a subtyping model, with 12 images used for four-class subtyping and 186 images for two-class subtyping [168]. A different study performed patch-level four-class subtyping on a total of 82 WSIs, with it being unclear whether the train-test split was made at the patient level. Another study reported 100% accuracy on the OCEAN challenge dataset [169], but the veracity of this research was questionable due to numerous errors in the manuscript, such as conflicting results, ROC curves made of only a single point, and claims that an uncited 'comprehensive literature review' found only one study in ovarian cancer histopathology. Another study reported 96% balanced accuracy on the OCEAN dataset but with only a very small test set of 15% of the available data [170].

A few recent subtyping studies have been more promising. One used the largest dataset from any previous study, consisting of 948 WSIs [136], to train and validate a multi-scale graph network [82], reporting a cross-validation slide-level balanced accuracy of 73%. Another study used a slightly larger set of 1113 WSIs to investigate a novel domain adaptation approach, achieving optimal balanced accuracies of 81% on internal data and 76% on external data [171]. Two other studies applied histopathology foundation models (see Chapter 7) to ovarian cancer subtyping, with one reporting an 82% balanced accuracy on the OCEAN dataset [172], and the other reporting around 88% balanced accuracy in six-class subtyping with 559 WSIs [173]. These are the highest accuracies reported in any studies to have used such large datasets.

Other Diagnostic and Prognostic Tasks

Prognostic models have remained common in recent literature, with studies focused on predicting survival [174–180], treatment response [181–186], and recurrence [187]. Some of these studies have focused on interpretable features in the tumour microenvironment, using AI to quantify tumour infiltrating lymphocytes [175, 177, 178], segment tumour and stroma [174, 176], or quantify collagen disorder [177], then using these features for prognostication in Cox regression models. Other approaches included MIL for prognostication of survival [179, 188] and treatment response [181, 183, 185]. Three studies used multimodal models, combining histopathology with clinical data for predicting recurrence [187] and survival [180], and with proteomics for predicting treatment response [184].

There was also a challenge based on prognostication. Automated Prediction of Treatment Effectiveness in Ovarian Cancer using Histopathological Images (ATEC23) [189] was built upon previous studies in predicting treatment response [147, 152]. Challenge participants were provided 288 training WSIs from 78 patients and were tasked with classifying whether patients would have six-month progression-free survival after treatment using individual TMA core images. None of the participants were able to achieve an accuracy greater than 70% despite a 57% class imbalance in the test set. This challenge is explored further with respect to our own participation in Appendix F. Both the ATEC23 and OCEAN challenges had a training set composed of mostly WSIs and a test set of mostly TMAs. The tasks in these challenges (treatment response and subtyping) are far from solved at the WSI level, so it is not surprising that participants were not able to achieve great performance when using the vastly less informative TMAs.

Six recent studies created patch-level malignancy classifiers [190–194], with these using datasets derived from TCGA [192, 194], the ATEC23 challenge [190, 191, 193], or an unclear dataset ‘from Kaggle’ [195]. These papers reported high classification performance but exhibited several risks of bias including small validation sets, potential data leakage between training and testing, and a lack of information concerning the development of ground truth labels (with the open access datasets not containing malignancy labels). Other studies included a cell segmentation model [196], and models for the classification of BRCA mutations [197] and ovarian cancer precursor lesions [198], the first AI model for this task.

3.5.4 Current Limitations and Future Recommendations

A large proportion of previously published work did not provide sufficient clinical and pathological information to fully assess the risk of bias. Researchers must thoroughly report data provenance to understand the extent of data heterogeneity, and to understand whether this has been appropriately accounted for in the study design. Modelling and analysis methods must also be thoroughly reported to improve reliability and reproducibility. It may be beneficial to refer to reporting checklists, such as *Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis* (TRIPOD) [199] and the recent AI-focused version, TRIPOD-AI [200], to ensure that all relevant details of the given study are understood and reported. In many studies, it was not clear how AI would fit in the clinical workflow, or whether there were limitations in how the methods could be applied. AI research should be conducted with an understanding of the clinical context of the data and any potential models, ideally with the direct and regular involvement of expert clinicians, such as histopathologists and oncologists.

Many previous studies did not make their data or code available, drastically reducing their reproducibility. It is relatively easy to publish code and generate documentation to enhance usability, and there are few drawbacks to doing so when publishing research. Making data available is more often difficult due to data governance requirements and the potential storage costs, but it can provide benefits beyond the primary research of the original authors. Digital pathology research in ovarian cancer is currently limited by the lack of openly accessible data, leading to over-dependence on TCGA and, more recently, ATEC23 and OCEAN. Many researchers have also been painstakingly collating similar but distinct internal datasets, which often contain little of the heterogeneity seen in multi-centre, multi-scanner data, making it difficult to train robust models or assess generalisability. Where heterogeneous data has been included, it has often included small quantities of data which were different to the majority, introducing risks of bias and confounding rather than helping to overcome these issues. TCGA-based studies are prone to this, with significant differences between TCGA slides originating from different data centres [201], but with many of these centres only providing small quantities of data, leading to a high likelihood of spurious correlations between data subsets. These issues may also be present in ATEC23 and OCEAN-based analyses, with ATEC23 being particularly heterogeneous (Appendix F) and OCEAN not being well-documented (Section 4.2). Improved datasets with detailed protocols describing data creation would allow researchers to conduct more thorough analyses and significantly improve model generalisability and clinical implementability.

For AI to achieve clinical utility in this field, it is essential that more robust validations are performed, especially considering the limitations of the available datasets. This must include thorough analyses, using techniques such as cross-validation, bootstrapping, and external validations to ensure that results are robust and truly reflect the ability of the models to generalise to unseen data, and are not simply caused by chance. The variability of results should be reported (typically in a 95% confidence interval), especially when comparing multiple models to help distinguish whether one model is genuinely better than another or whether any difference is simply due to chance. Statistical tests can also be beneficial for these evaluations.

Current literature in this field can be largely characterised as model prototyping with homogeneous retrospective data. Researchers rarely consider the reality of human-machine interaction, perhaps believing that these models are a drop-in replacement for pathologists. However, these models perform narrow tasks within the pathology pipeline and do not take into consideration the clinical context beyond their limited training datasets and siloed tasks. These models are likely to be more beneficial (and more realistic to implement) as assistive tools for pathologists, providing second opinions or novel ancillary information. While current research is typically focused on assessing model accuracy without any pathologist input, different study designs could be employed to better assess the real-world utility of these models as assistive tools. For example, usability studies could investigate which models are most accessible and most informative to pathologists in practice, and prospective studies could quantify any benefits to diagnostic efficiency and patient outcomes, and investigate the robustness of models in practice. AI-clinician interaction has thus far been a relatively small field of research [202], and we are not aware of any study in which pathologists have systematically compared the merits of different AI implementations for a specific task. Such research would significantly benefit clinical translation.

3.6 Conclusion

In this chapter, we reviewed previous studies in which AI models were used for the diagnosis or prognosis of ovarian cancer from histopathology slides. We identified several weaknesses in previous research and explored how these may be addressed. Many researchers did not sufficiently understand their data and ensure that planned research was clinically relevant with direct oversight from clinicians. Many previous studies had high risks of bias due to using small, heterogeneous datasets with minimal validation techniques, which often did not involve cross-validation, external validation, bootstrapping, or hypothesis testing. Many studies did not include key information about the data used, including how patients were recruited and selected, and how tissue specimens were processed to generate digital pathology images. Finally, few researchers made their code or data accessible, reducing reproducibility and imposing barriers on future researchers. While this field has been growing in recent years, it is not clear that the quality of overall research is improving. The lessons learned through this review have been instrumental in shaping the methodologies and reporting of our primary research. The recommendations provided for future research included using reporting checklists, publishing code, sharing data (where possible), performing deeper validations, reporting variability, and assessing real-world utility.

Chapter 4

Methodology

In this chapter, we detail recurring aspects of the methodology used throughout the following chapters. This includes data acquisition, preprocessing, modelling, hyperparameter tuning, and validation procedures. Some aspects differ in specific chapters, for example, Chapter 5 contains much of the early work which influenced subsequent study design, and Chapter 8 contains an entirely different classification model developed based on the knowledge derived from earlier research.

4.1 Baseline Subtype Classification Model

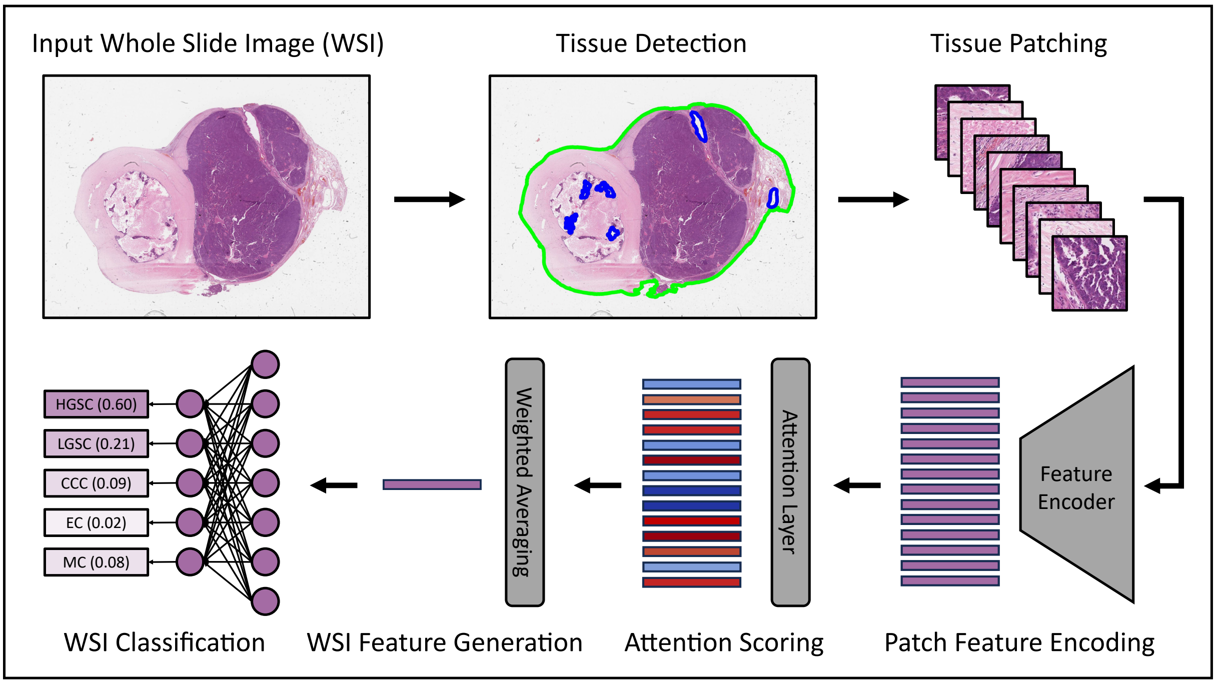


Figure 4.1 Attention-based multiple instance learning (ABMIL) [77] model pipeline for ovarian cancer subtyping.

We adopted the standard ABMIL classifier [77] (Figure 4.1) as a baseline model for our experiments using the implementation from the CLAM repository [78]. This is a commonly used MIL approach which gives consistently strong performance across many tasks, even when compared to more complex MIL approaches [79–81, 136].

The default preprocessing method from the CLAM code repository [78] was implemented to perform tissue segmentation and colour normalisation, with the latter step applied after patch extraction. For tissue segmentation, a downsampled version of each WSI was converted from RGB images into the hue-saturation-value (HSV) colour space and a median blurring was applied to smooth the saturation channel and reduce noise. The segmentation was performed by applying a saturation threshold of 8/255, where all pixels with saturation greater than the threshold were labelled as tissue and all other pixels were labelled as background. Morphological closing was then used to smooth the edges and close small gaps in the detected tissue mask. These tissue masks were defined by their outer contours, and these masks were rescaled to match the size of the WSI.

To extract tissue patches, a rectangular bounding box was placed around each tissue contour and split into 512 x 512 pixel non-overlapping patches [78]. Any patch in which the central point was outside the tissue contour was dropped, leaving the set of viable tissue patches. These patches were downsampled from 512 x 512 pixels at 40x native magnification to 256 x 256 pixels at 20x apparent magnification, an approach which was very common in previous research (Chapter 3) to reduce the effective size of WSIs and hence the computational workload.

The baseline ABMIL model employed an ImageNet-pretrained ResNet50 encoder to extract patch-level features. ImageNet (a set of 1.4 million natural images from 1000 classes) [58] is popular for model pretraining as the quantity and diversity of labelled images enables the creation of a multi-purpose feature set. ResNet50 [59] is a standard CNN which had been trained in a fully supervised manner, achieving a reported 85.1% ImageNet classification accuracy. To use this model as a feature extractor, outputs were taken from the end of the third residual block to give 1024 features per input patch. Before feature extraction, tissue patches were normalised using the ImageNet standard RGB parameters (`mean = (0.485, 0.456, 0.406), std=(0.229, 0.224, 0.225)`) using the standard statistical normalisation procedure $\frac{\text{value}-\text{mean}}{\text{std}}$. The reduction from 256 x 256 x 3 (height x width x colour channels) to 1 x 1024 features compressed the patches by a factor of 192.

ABMIL feature aggregation used a trainable sigmoid gated-attention layer to give each patch feature vector an attention score between 0 and 1, with an attention-weighted average of the patch features taken to generate a 1 x 1024 WSI feature vector. Finally, this WSI feature vector was classified through a fully connected neural network with one output node per class.

4.2 Datasets

The models presented in this thesis were exclusively developed and trained using ovarian carcinoma WSIs from cases treated at Leeds Teaching Hospitals NHS Trust (LTHT), with ethics approval provided by the Wales Research Ethics Committee (reference 18/WA/0222). This dataset was developed by a histopathologist concurrently with the research presented in this thesis, meaning that the full dataset was only available for the most recent work (Chapters 7 and 8). All models were initially validated using the internal dataset, and when performance started to reach reasonable levels, external datasets were sourced to further validate performance.

4.2.1 Internal Ovarian Cancer Dataset

Carcinoma Subtype	Training WSIs (Patients)	Hold-out WSIs (Patients)
High-Grade Serous (HGSC)	1266 (307)	20 (7)
Low-Grade Serous (LGSC)	92 (21)	20 (6)
Clear Cell (CCC)	198 (45)	20 (7)
Endometrioid (EC)	209 (38)	20 (5)
Mucinous (MC)	99 (22)	20 (5)
Total	**1864 (433)**	**100 (30)**

Table 4.1 Dataset breakdown for the training (cross-validation) set and independent internal hold-out test set. Numbers in brackets indicate the number of unique patients.

The LTHT ovarian cancer dataset was retrospectively collected by a histopathologist (Katie Allen) from cases treated between 2008 and 2022. Cases were only included if a gynaecological pathologist had diagnosed them as one of the five most common epithelial ovarian cancer subtypes (HGSC, LGSC, CCC, MC, EC). Tumours were only included if they were carcinomas of tubo-ovarian-primary peritoneal origin with a single epithelial subtype present, and with associated clinical metadata available in the patient health records. The histopathologist independently verified all diagnoses, removing any cases with discrepancies. Several representative H&E-stained adnexal tissue glass slides were selected for each case, with only FFPE samples used. Any mounting artefacts were corrected, and slides were cleaned and anonymised before being digitised at 40x magnification using a single Leica AT2 scanner.

As shown in Table 4.1, the final training dataset consisted of 1864 WSIs from 433 ovarian carcinoma cases, consisting of 1412 primary surgery sample WSIs from 296 cases, and 452 IDS sample WSIs from 137 cases. The population-level class imbalance was reflected in the training set, with the least common subtype (LGSC) represented by only 92 WSIs from 21 cases, compared to 1266 WSIs from 307 cases for the most common subtype (HGSC). This set also reflected the high frequency of stage III, high-grade cancers, particularly driven by the high proportion of HGSC cases (Table 4.2). We aimed to classify primary surgery WSIs as these are generally of better diagnostic quality than IDS specimens (Section 2.2). The training set included both primary surgery samples and IDS samples as we found their inclusion to be beneficial to model training [8].

An independent set was collected following the same protocol, from which 20 WSIs of each carcinoma subtype were taken to form a class-balanced hold-out test set. Rather than representing realistic clinical frequencies, this set focused on quantifying performance across all subtypes equally. It also focused on the goal of accurately classifying the clinical standard primary surgery samples, and so didn’t include any IDS samples. Neoadjuvant chemotherapy usage is much more common in later-stage cancers, so restricting this set to primary surgery samples (without neoadjuvant chemotherapy) led to a much higher proportion of early-stage cancers than in the training set (Table 4.2). Balancing the subtypes also led to the set containing a similar proportion of high-grade and low-grade cancers.

Training Set

Subtype	**HGSC**	**LGSC**	**CCC**	**EC**	**MC**	**Total**
Stage I	23	5	24	24	15	91
Stage II	33	0	5	10	2	50
Stage III	210	12	11	3	4	240
Stage IV	41	4	5	1	1	52
Grade 1 / Low Grade	-	21	-	15	8	23
Grade 2	-	-	-	17	10	27
Grade 3 / High Grade	307	-	45	6	1	359
Ungraded	0	0	0	0	3	3
Primary Surgery	188	11	39	37	21	296
Interval Debulking Surgery (IDS)	119	10	6	1	1	137

Hold-out Test Set

Subtype	**HGSC**	**LGSC**	**CCC**	**EC**	**MC**	**Total**
Stage I	3	3	5	5	3	19
Stage II	2	1	0	0	0	3
Stage III	2	2	2	0	1	7
Stage IV	0	0	0	0	1	1
Grade 1 / Low Grade	-	6	-	2	2	10
Grade 2	-	-	-	3	1	4
Grade 3 / High Grade	7	-	7	0	0	14
Ungraded	0	0	0	0	2	2
Primary Surgery	7	6	7	5	5	30
Interval Debulking Surgery (IDS)	0	0	0	0	0	0

Table 4.2 Clinical features of the 433 training set and 30 hold-out test set patients. Grade 1 is combined with low grade, and grade 3 is combined with high grade, though these terms are not interchangeable. Grades which are incompatible with specific subtypes have been marked with hyphens (-).

4.2.2 External Datasets

Carcinoma Subtype	**Transcanadian [203] WSIs (Patients)**	**OCEAN Challenge [160] WSIs**
High-Grade Serous (HGSC)	30 (30)	217
Low-Grade Serous (LGSC)	9 (9)	42
Clear Cell (CCC)	20 (20)	94
Endometrioid (EC)	11 (11)	119
Mucinous (MC)	10 (10)	41
Total	**80 (80)**	**513**

Table 4.3 Dataset breakdown for the external validation sets. Numbers in brackets indicate the number of unique patients, which was not clear for the OCEAN dataset.

Ovarian cancer has a relative sparsity of available datasets, with the most commonly used set in previous AI research (TCGA-OV [154]) only including serous carcinomas. While there were no suitable external datasets available at the start of this research, two sets recently became available. The first was a dataset from the Transcanadian Study [203]. This set consisted of 80 WSIs from 80 cases digitised using an AperioScope scanner and made available at 20x magnification, alongside subtype labels that had been determined by a gynaecological pathologist. This dataset has previously been used in the training and testing of a subtype classification model [116, 117], but to the best of our knowledge, it has never previously been used as a stand-alone validation set.

The second dataset was only made available for general research usage in April 2024, following the OCEAN Challenge the previous year [160]. The publicly available part of the dataset contained a total of 513 WSIs digitised at 20x magnification (as well as 30 TMAs, which we did not utilise given our focus on classifying WSIs). Details were sparse regarding the data collection, curation, and labelling procedures, though these were likely heterogeneous considering data was sourced from over 20 data centres (including the TMAs and test WSIs which were not publicly available).

With the exception of confirming that the Transcandian Study contained only primary surgery specimens through direct contact with the primary author, we were unable to obtain the clinical characteristics of either external dataset.

4.3 Hyperparameter Tuning and Model Training

An iterative grid search strategy was used to perform hyperparameter tuning for the classification stage of ABMIL, with 2-3 of the hyperparameters selected to be adjusted and evaluated at a time, and all other hyperparameters frozen at their previous best values. The performance of each hyperparameter configuration was evaluated using the average loss of each validation set in five-fold cross-validation. These cross-validation splits were stratified at the patient level to give relatively class-balanced folds while avoiding data leakage. The iterative tuning procedure risked finding a local optimum rather than a global optimum, but it was the most rigorous available method considering that it was not computationally feasible to conduct a full hyperparameter grid search. The standard hyperparameters are defined in this section, though other hyperparameters varied based on the specific requirements of each set of experiments.

The loss function used for model training in the CLAM repository was the cross-entropy loss [78], though given the class imbalance in our dataset we instead used the *balanced* cross-entropy loss:

$$l(\mathbf{y}, c) = -w_i \log \frac{\exp{(y_c)}}{\sum_{i=1}^{m} w_i \exp{(y_i)}}, \tag{4.1}$$

for classes $c \in 1, 2, ..., m$, class predictions $\mathbf{y} \in [0, 1)^m$, and with class weights w_i. These class weights are inversely proportional to the class frequencies n_1 to n_m to balance the relative importance of each class, and are normalised using the average number of slides per class $\frac{1}{m}\sum_{j=1}^{m} n_j$ such that the average weight per WSI equals 1, which reduces the instability caused by the scaled loss values having varied magnitudes:

$$w_i = \frac{\sum_{j=1}^{m} n_j}{m \times n_i}. \tag{4.2}$$

Models were trained with an adaptive moment estimation (Adam) optimiser [204] (as in CLAM [78]), a stochastic gradient descent method in which moving averages estimating the mean and variance of the gradient are used to smooth the optimisation. The Adam update procedure for parameters θ_t with gradients g_t at timestep t is as follows, where m_t and v_t are the rolling estimates of the first and second moments (the mean and variance) of the gradient, and α, β_1, β_2, ϵ are hyperparameters:

$$m_t = \frac{\beta_1 \cdot m_{t-1} + g_t(1-\beta_1)}{1-\beta_1}, \tag{4.3}$$

$$v_t = \frac{\beta_2 \cdot v_{t-1} + g_t^2(1-\beta_2)}{1-\beta_2}, \tag{4.4}$$

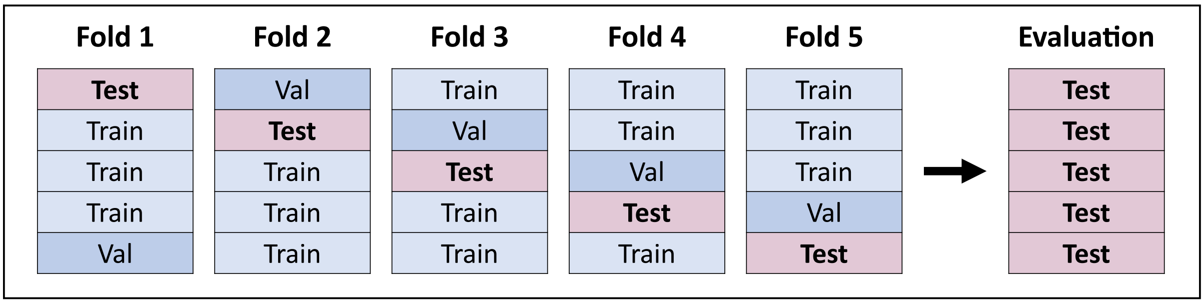
$$\theta_t = \theta_{t-1} - \alpha \frac{m_t}{\sqrt{v_t}+\epsilon}. \tag{4.5}$$

These hyperparameters are the **learning rate** (α), **first moment decay** (β_1), **second moment decay** (β_2), and numerical **stability parameter** (ϵ). As well as the initial learning rate, hyperparameters controlled the rate of decay of the learning rate, with one setting the **LR decay patience** (the number of training epochs without an improvement before the learning rate was decreased), and another controlling the **LR decay factor** (the factor by which the learning rate was multiplied during decay).

To attempt to reduce the effects of overfitting and hence improve model robustness, three types of regularisation were employed during model training, each with a related hyperparameter. The **weight decay** hyperparameter controlled the relative strength of an imposed L2 regularisation penalty, which penalised the sum of squares of the weights when calculating the loss function to incentivise small model weights and hence a parsimonious model. The **dropout rate** controlled the proportion of model parameters that were dropped (set to zero) before the final classification step in the model. The **max patches** hyperparameter applied a similar procedure to the data, with patches randomly selected from a slide and the other patches dropped. These dropout procedures reduced overfitting by acting as efficient data and feature augmentation techniques, making it more difficult for the model to take shortcuts and learn spurious relationships [205].

4.4 Model Validation

The best overall hyperparameter configuration from tuning was used to train a model for each cross-validation fold, and these models were evaluated on the withheld test splits, as shown in Figure 4.2. The cross-validation test sets were somewhat exposed by their usage in the training/validation sets of other folds when determining the optimal hyperparameters, so model performance was also evaluated using a hold-out test set and external validation sets where possible. For such validations, the five cross-validation model predictions were ensembled by taking the average of the softmaxed classifier outputs. Classification performance was evaluated using the balanced accuracy, AUROC, and F1 score (Section 2.5). Bootstrapping was used to evaluate the variability of results, with all model predictions resampled 10,000 times and the mean and 95% percentile confidence interval of each metric calculated.



Fold 1	Fold 2	Fold 3	Fold 4	Fold 5		Evaluation
Test	Val	Train	Train	Train		**Test**
Train	**Test**	Val	Train	Train		**Test**
Train	Train	**Test**	Val	Train	→	**Test**
Train	Train	Train	**Test**	Val		**Test**
Val	Train	Train	Train	**Test**		**Test**

Figure 4.2 Five-fold cross-validation procedure for model training and evaluation. Splits were made at the patient level and kept consistent between the different models.

We measured model efficiency both in terms of the number of model parameters and the runtime in training and inference. We excluded the feature extraction requirements from training time since this step is frozen, so only needs to be run once before any tuning experiments. Feature extraction times were included when measuring inference times to best represent the computational burden of a deployed model.

4.4.1 Hypothesis Testing

In the most rigorous experiments, paired *t*-tests were used to statistically compare model performance for each metric and each validation set across the cross-validation folds. Where multiple models were compared against the same baseline, p-values were adjusted for multiple testing using the Benjamini-Hochberg false discovery rate correction [206]. Results were considered *statistically significant* given an adjusted p-value < 0.05.

4.5 Software and Hardware

Experiments were conducted on two distinct devices which we refer to as the personal computer (PC) and the high-performance computer (HPC). The PC was a standard consumer desktop computer with a single NVIDIA GTX 1660 GPU with 6GB of VRAM, an Intel i5-4460 central processing unit (CPU) @ 3.2GHz, and 16 GB of RAM. The HPC was an NVIDIA DGX A100 server with 8 NVIDIA A100 GPUs and 256 AMD EPYC 7742 CPUs @ 3.4GHz. The HPC GPUs were each segmented into seven instances, with only one instance used for all experiments except those using the largest feature extraction models in Chapter 7.

All experiments were conducted using a PyTorch [207] code base developed as an extension to the CLAM model pipeline [78]. WSIs were loaded from `.svs` files using the Python `openslide` library and the `cv2` package was used for many chromatic and morphological operations in preprocessing and patch extractions. Hyperparameter tuning was implemented using the `Ray Tune` Python library, with the `TrialPlateauStopper` function used to stop training for any configuration where no improvement had been found for a given number of epochs. Hypothesis testing was implemented with the `ttest_rel` function in the `scipy.stats` Python module and p-values were adjusted using the `statsmodels` function `multipletests`. The graph models in Chapter 8 were implemented in Pytorch Geometric. For each set of experiments, a public GitHub repository was published at https://github.com/scjjb, with code examples and further guidance provided to aid reproducibility.

Chapter 5

Active Sampling for Efficient Multiple Instance Learning

In this chapter, we describe our initial experiments in ovarian cancer subtyping. We use an ABMIL backbone and propose an active sampling approach for efficient slide inference by leveraging spatial relationships and attention scoring to determine discriminative tissue regions. This approach is validated using the earliest iterations of the Leeds ovarian carcinoma dataset, with these analyses being fundamental in shaping the full dataset and subsequent modelling techniques.

5.1 Introduction

The huge size of digital pathology slides presents a significant computational burden for AI models. In research settings, it is increasingly common to classify WSIs using HPCs [75, 78, 81, 208], but these are unlikely to be available in the clinical setting due to the financial cost of acquisition and maintenance. It may instead be expected that pathology slides or their digitised image files are outsourced to a location with greater hardware access, but there is still an associated financial cost to hire or maintain off-site resources, and doing so increases the logistical complexity of pathology, which may delay diagnosis and present data governance and security issues. It would be beneficial for classification algorithms to instead be made efficient enough to run on standard clinical hardware (either standard clinical computers or digital pathology scanners). Such an approach may also benefit access to AI diagnostic tools in lower-resourced healthcare settings, where the potential benefit of AI models is greatest given the unavailability of subspecialty experts.

The scalability of digital pathology AI models is essential due to the huge quantity of data generated. For example, at LTHT alone, over 290,000 H&E slides are generated each year [209]. Not all slides contain as much tissue as a typical ovarian cancer resection slide, and they may not all be the targets of AI models, though to apply a model to even 20% of them would require processing a gigapixel WSI every 9 minutes, non-stop. This will become an even greater issue as labs start applying several different algorithms to perform tasks such as quality control, tumour segmentation, metastasis detection, prioritisation, diagnostic classification, and prognostic prediction.

As described in Section 2.4.4, it is common for slide-level classification to be performed using MIL methods, where information is learned at the patch level and aggregated to model the WSI. When using MIL methods in a setting with many instances per bag, such as modelling a WSI as a collection of many patches, it may be pertinent to use within-bag instance sampling to effectively reduce the number of instances. Instance sampling approaches are intended to focus on relevant instances, improve robustness to outlier instances, or reduce the overall computational burden. Within-bag sampling can be as simple as randomly selecting instances [210, 211], or only selecting instances within a specific region when using spatially related instances, such as patches from within an image [212]. Multiple magnifications of a histopathology slide can be efficiently leveraged by random instance selection across magnifications [118], or by performing discriminative region detection on the lower-magnification (smaller) image to guide instance selection on the higher-magnification (larger) image [213].

Within-bag sampling has previously been integrated with ABMIL [77] by splitting each bag into a group of *mini-bags* - overlapping subsets of the original bag [214]. The ABMIL model is trained with these mini-bags, and the slide-level classes are determined by the majority vote of classified mini-bags. This approach reduces memory requirements, but the duplication of instances across multiple mini-bags is likely to increase inference time. Further, as the key instance detection is based upon the ABMIL attention weights, all instances are passed through the feature extractor, which is typically the part of the model with the greatest computational burden. Subsequent work [215] showed this approach to be less accurate than conventional single instance learning for cytological data, but it has not been evaluated for whole slide histopathological data, where single instance learning is not feasible due to significantly larger image sizes.

Some MIL sampling approaches identify relevant patches using patch classification scores rather than attention scores, following an instance classification MIL approach. For example, in top-k sampling [75] all patches are evaluated and those with the highest patch classification scores for the positive class are used for slide-level classification. Monte Carlo sampling [216] instead takes an initial random sample, then iteratively replaces the patches with the lowest individual classification scores with new random patches to improve the overall discriminative power of the sample. Patches can also be sampled using expectation maximisation [70]. It is not clear that any of these classification approaches offer efficiency improvements, with top-k sampling and expectation maximisation requiring all patches to be processed through a CNN before sampling, and Monte Carlo sampling reported to be slower than whole slide processing. While these classification approaches have not demonstrated an increase in efficiency, similar approaches have been shown to benefit WSI segmentation speed without sacrificing accuracy [217, 218].

In this chapter, we present a novel patch sampling approach for use during the inference step of ABMIL. This is an iterative approach in which the attention scores of previously sampled patches are used to assign sampling weights to neighbouring patches, aiming to sample the most diagnostically relevant tissue without fully processing the entire slide and thus achieve high accuracy with improved algorithmic efficiency.

5.2 Methods

To investigate instance sampling for WSI inference, we proposed the approach of *Discriminative Region Active Sampling for Multiple Instance Learning (DRAS-MIL)*. This method uses the trained baseline ABMIL model described in Section 4.1 with an adjusted methodology for instance selection during inference. Where ABMIL uses all available tissue patches, DRAS-MIL aims to find a discriminative subset of patches at a much lower computational cost. Initially, DRAS-MIL takes a random sample of patches and passes these through the trained ABMIL model to calculate their attention scores. These attention scores are then used to generate sampling weights for the remaining patches, with higher sampling weights given to patches in close spatial proximity to high-attention patches. This process is repeated for a fixed number of iterations, and finally, the sampled patches are used to classify the WSI through the trained ABMIL model. This approach leverages the spatial relationships inherent to ovarian carcinoma pathology slides, with diagnostically important tissue areas likely to be clustered together rather than evenly dispersed throughout the entire slide.

We developed and validated the DRAS-MIL model in two stages, with the initial development performed using the earliest iteration of the LTHT ovarian cancer dataset, and the final validation performed on a slightly larger and more refined dataset. As shown in Table 5.1, model prototyping was conducted using a set of 655 WSIs from 127 patients, and final validations were conducted using 714 WSIs from 147 patients. Many slides in this iteration of the dataset contained common non-adnexal tissue types such as omentum and lymph nodes. The initial prototyping dataset also included several WSIs containing rare metastases that had been erroneously included, and so were removed for the final dataset. Despite this being one of the largest ovarian cancer subtyping datasets at the time, there were only 11-16 patients for each of the non-HGSC subtypes, so we focused on the binary classification of HGSC against all other subtypes. The minority non-HGSC was class taken as the positive class.

Carcinoma Subtype	**Initial WSIs (Patients)**	**Final WSIs (Patients)**
High-Grade Serous (HGSC)	416 (75)	455 (92)
Low-Grade Serous (LGSC)	64 (13)	75 (14)
Clear Cell (CCC)	47 (12)	60 (15)
Endometrioid (EC)	78 (16)	76 (15)
Mucinous (MC)	50 (11)	48 (11)
Total	**655 (127)**	**714 (147)**

Table 5.1 Dataset breakdown for the initial and final modelling. Numbers in brackets indicate the number of unique patients.

5.2.1 Initial Prototyping

For initial prototyping, a baseline ABMIL classifier was trained through a 10-fold cross-validation for the binary classification of HGSC using the default hyperparameters from the CLAM repository [78], except for the learning rate which was set to 5e-5 as the default value of 1e-4 gave divergent behaviour. This model served as a backbone for prototyping instance sampling techniques. The baseline version of the proposed DRAS-MIL method started with a completely random sample of 100 patches and used the attention scores of these patches to assign initial sampling weights to the 20 nearest neighbours of each sampled patch. Any patch within the spatial neighbourhood of multiple sampled patches was given the maximum attention-based sampling weight. There were nine resampling iterations, with 100 extra patches sampled per iteration (half at random, half through weighted sampling), each followed by recalculating the sampling weights. A final weighted sample of 200 patches was then taken to give a total of 1200 sampled patches per WSI. Finally, these 1200 patches were used to classify the WSI through the trained ABMIL model.

Prototype models were compared with different numbers of nearest neighbours for assigning sampling weights (20, 50, 80, 100), different numbers of sampling epochs (4, 6, 10, 20), different final sampling strategies for classification (using all sampled patches vs taking a smaller final sample), different random sampling strategies (truly random vs spatially distributed in a grid), different sample weight combination strategies (average vs maximum), and different proportions of random samples to active samples (30%, 50%, 80%). A higher proportion of completely random sampling would allow for more exploration, and a lower proportion would allow for greater exploitation of the high-attention regions found in previous sampling iterations. The baseline model, then, used 20 nearest neighbours, 10 sampling epochs, all previous samples in the final sample, 50% truly random sampling, and maximum sample weights.

Model prototypes were primarily evaluated using the AUROC across the ten cross-validation folds, and this was repeated 10 times to measure the variability inherent to sampling approaches. The best-performing active sampling model was also evaluated with reduced quantities of patches per slide (specifically 250 and 500, where the original approach used 1200), and these sampling approaches were compared to fully random sampling.

5.2.2 Final Validations

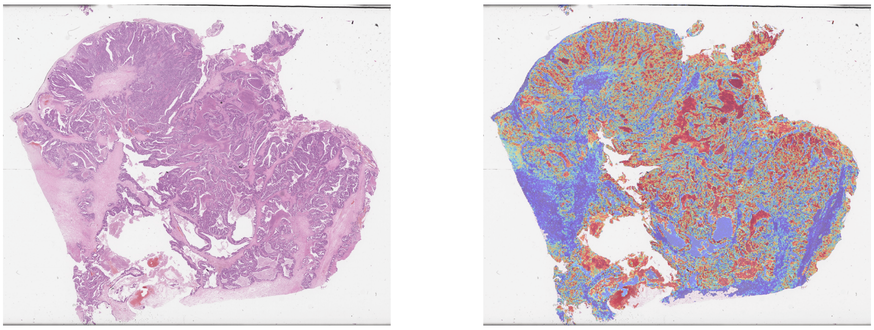
Following the prototyping experiments, the most promising models were fully tuned and validated using the refined set of 714 WSIs (Table 5.1). To make hyperparameter tuning computationally feasible, the final validations were conducted using a 3-fold cross-validation rather than the previous 10-fold. Three hyperparameters were tuned during the training of the baseline ABMIL model and four were separately tuned during inference with the DRAS-MIL sampling method, all using a random hyperparameter search (Table 5.2) to minimise the unbalanced cross-entropy validation loss on a single cross-validation fold. A total of 100 configurations were evaluated for the training hyperparameters and 500 configurations for the inference hyperparameters, with this being a practical limit due to the computational complexity of the models. The hyperparameters that were tuned during model training were the learning rate, weight decay, and dropout rate (as described in Section 4.3).

The hyperparameters that were tuned during inference controlled the number of **resampling iterations**, the number of **nearest neighbours** assigned sampling weights around each previous sample, the initial completely **random sampling proportion**, and the **random sampling decay** which reduced the random sampling proportion each iteration. The options for the hyperparameters were influenced by the results of initial prototyping experiments, which also influenced the decisions to sample a total of 800 patches per WSI, to use non-grid-based random samples, and to take the average sampling weight rather than the maximum when a patch was spatially close to multiple previous samples. The 800 patches per WSI were composed of 640 patches from the iterative procedure (with the number of samples per iteration being 640 divided by the number of resampling iterations), before a final set of 160 non-random samples.

	Hyperparameter	Function	Distribution	Best
Training	Learning Rate	The initial rate of change for model parameters	Log-Uniform (1e-5, 1e-2)	0.0038
	Weight Decay	The relative strength of L2 regularisation on loss function	Log-Uniform (1e-10, 1e-2)	0.00079
	Dropout Rate	The proportion of model weights dropped to reduce overfitting	Uniform (0.00, 0.99)	0.020
Sampling	Resampling Iterations	The number of resampling iterations	Choice [2,4,6,8,10,12,16]	16
	Nearest Neighbours	The number of nearest neighbours assigned weights	Choice [4,8,16,32,48,64]	64
	Random Sampling Proportion	The proportion of samples which are randomly sampled	Uniform (0.00, 0.75)	0.29
	Random Sampling Decay	The reduction in Random Sampling Proportion each iteration	Log-Uniform (0.0001, 0.5)	0.36

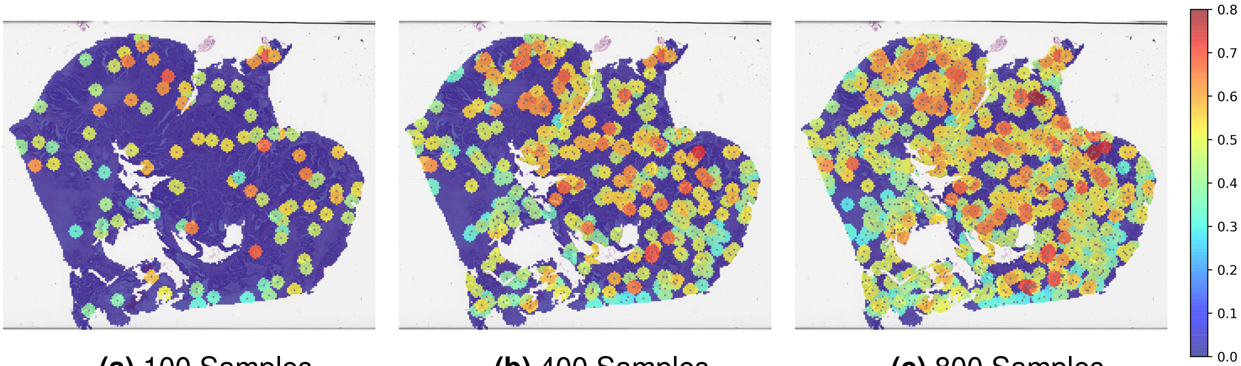
Table 5.2 Hyperparameters tuned using a random search. The first three were tuned during baseline model training, and the subsequent four during inference.

The 800 patches sampled per WSI represented approximately 5% of the tissue area in a typical slide (the slides had 740-33961 tissue patches each, with a mean average of



(a) Unprocessed WSI **(b)** ABMIL attention scores

Figure 5.1 Attention scores from ABMIL whole slide processing.



(a) 100 Samples **(b)** 400 Samples **(c)** 800 Samples

Figure 5.2 Illustrative example of sampling weights generated through spatial sampling at different stages of the active sampling process. The closest 50 patches to each previous sample are assigned the corresponding sampling weight, and the proportion of random samples taken is 0.5, giving a relatively high level of exploration.

15990 and a median of 16230). Any slide with fewer than 800 patches was evaluated with whole slide processing, though this only applied to one slide in the dataset. 800 samples were sufficient to generate sampling weights for the majority of patches in a slide, as shown in Figure 5.2. We compared the performance of DRAS-MIL to completely random sampling with the same number of patches, and to the baseline ABMIL approach using all available patches. To account for randomness and the relatively small available dataset, we repeated each sampling approach 50 times and performed 100,000 iteration bootstrapping, where each slide was represented exactly once per iteration by one of the 50 predictions made for the slide. While the AUROC was sufficient to give a holistic measure of performance when comparing models during initial development, this would not provide sufficient understanding of clinical utility for final validations, which were evaluated using balanced accuracy, AUROC, and F1 score.

We compared the efficiency of the proposed DRAS-MIL model to the baseline model by measuring the average inference time and the maximum GPU memory requirements using a fixed test set of 50 randomly selected WSIs. There was a trade-off between inference time and memory requirements, so we measured both of these metrics for different batch sizes (1, 4, 8, 16, 32, and 64), which represented the maximum number of patches processed concurrently. Efficiency experiments alternated between active sampling and the default whole slide processing of ABMIL, running each three times and taking the median value for each batch size as the true value.

Model training and hyperparameter tuning were conducted on the HPC (Section 4.5). Efficiency experiments were conducted on the PC with the same CPU as found in the computers in the LTHT pathology lab. The code for this chapter was made available at https://github.com/scjjb/DRAS-MIL.

5.3 Results

5.3.1 Prototyping Results

Figure 5.3 shows the results from 13 prototyping configurations alongside completely random sampling. Compared to the default ABMIL score of 0.790, completely random sampling had a median AUROC of 0.788 across 10 repeats, though it was highly variable, with a range of 0.028. The baseline DRAS-MIL method had a lower median AUROC of 0.784, but was much less variable, with a range of 0.014.

The greatest overall performance from a DRAS-MIL prototype model was a median AUROC of 0.789 and range of 0.016, a marginal improvement over random sampling, though still behind default whole-slide processing. In this optimal prototype, the initial sample of 200 patches was selected randomly (not spatially distributed), and then there were 10 resampling iterations of 100 patches each, with 50% selected at random. Sampling weights were propagated to the 50 nearest neighbours of previous samples, with the maximum weight (rather than the average weight) applied for any patch in the receptive field of multiple previous samples. Finally, all previously sampled patches were retained when making the final classification.

Only the aforementioned active sampling prototype model outperformed completely random sampling overall, though two other prototypes gave comparable performance. Both of these also used 50 nearest neighbours to propagate sampling weights, with one starting with 80% random samples and reducing this by 8% on each resampling iteration, and the other taking 50 patches per resampling iteration and taking the average sampling weight for any patch in multiple receptive fields. These prototypes gave median AUROCs of 0.788 and 0.787, respectively. All other prototypes had median AUROC scores between 0.779 and 0.784.

As shown in Figure 5.4, both random and active sampling gave worse performance with a smaller sample size. From a median AUROC of 0.789 with 1200 patches, the active sampling prototype performance fell to 0.773 with 500 patches and 0.767 with 250 patches. Similarly, from a median AUROC of 0.788 with 1200 patches, the random sampling performance fell to 0.777 with 500 patches and 0.771 with 250 patches. As such, random sampling performed better than the active sampling approach when using smaller sample sizes.

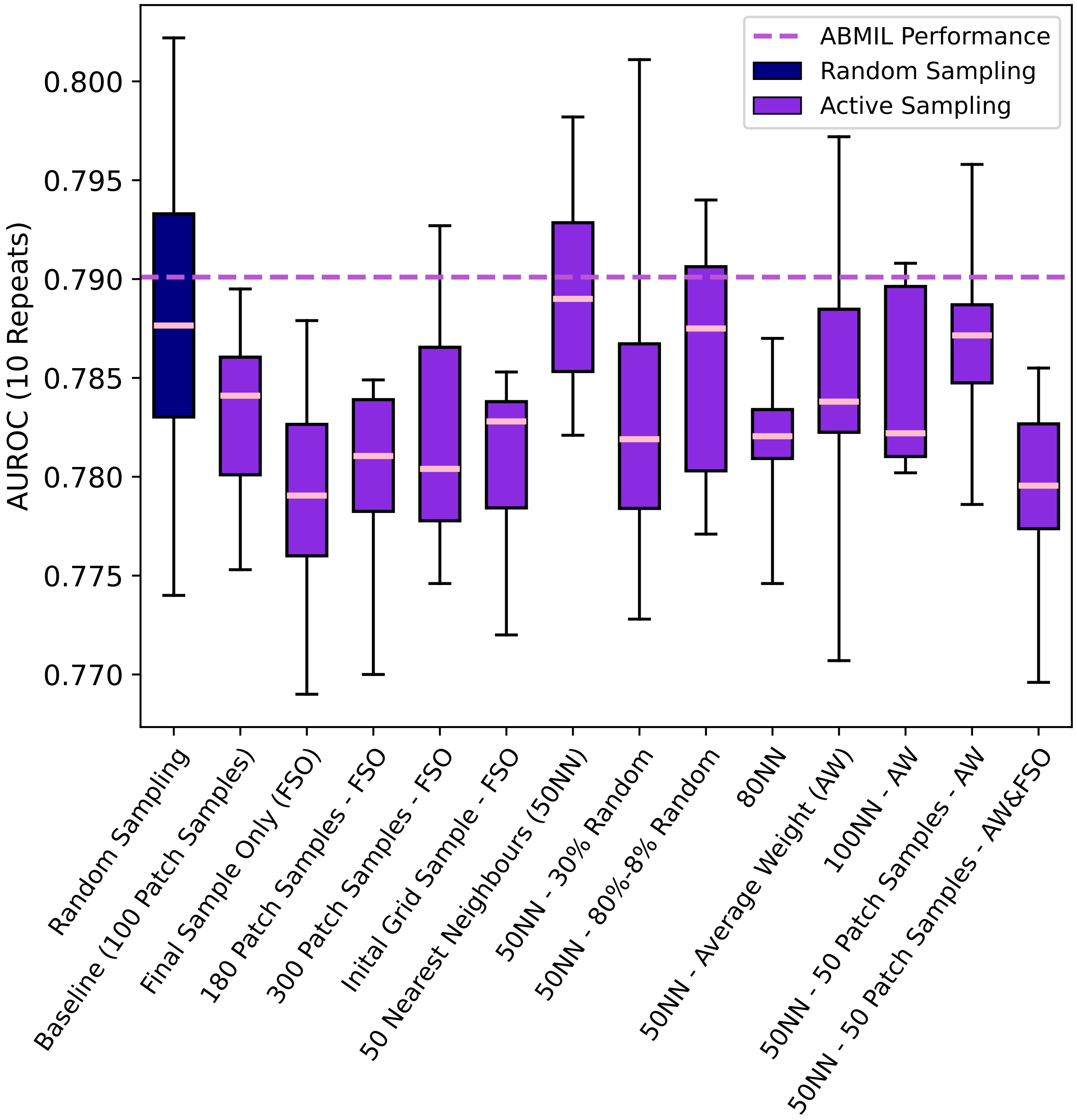


Figure 5.3 AUROC scores from 10 repeats of 10-fold cross-validation for binary classification of HGSC in the initial prototyping set of 655 WSIs, with sampling methods evaluated using 1200 total patches per slide.

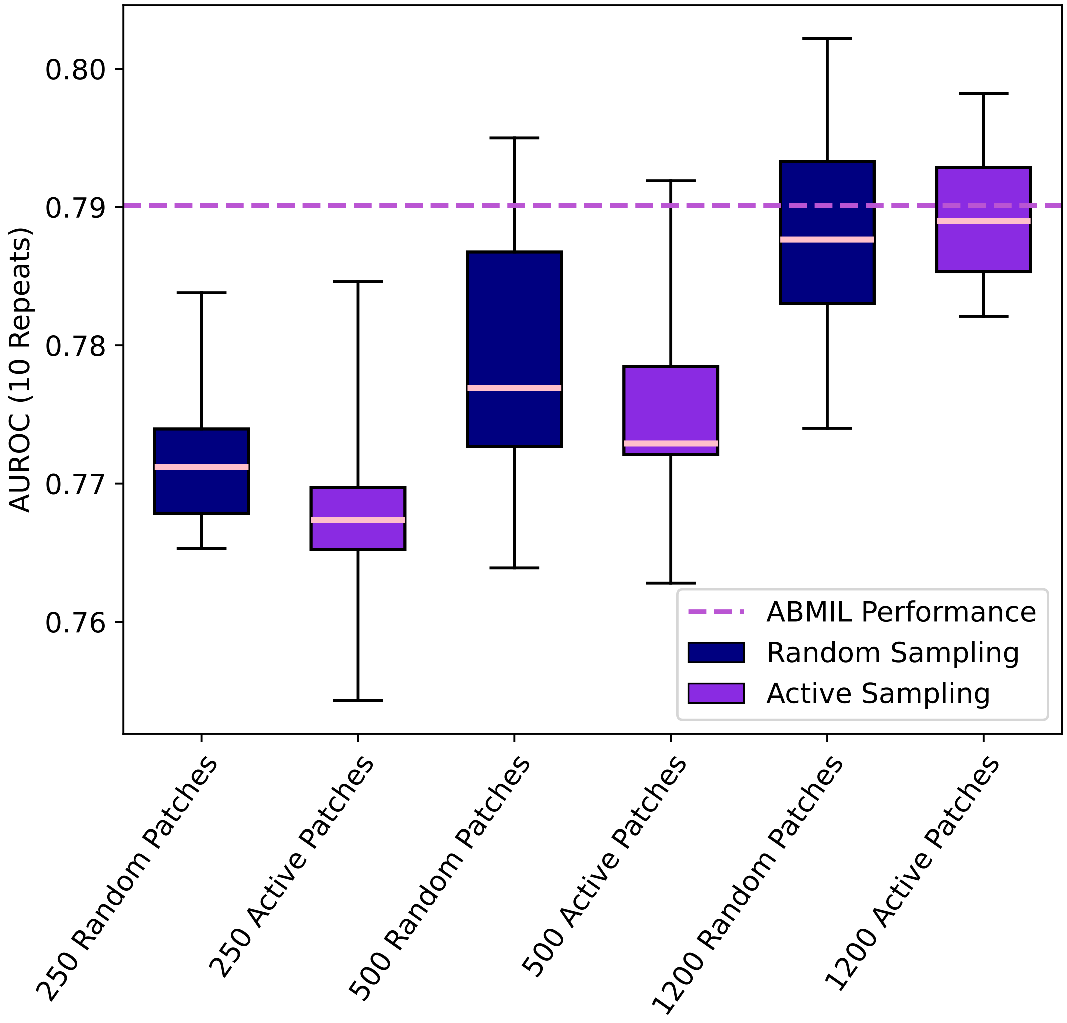


Figure 5.4 AUROC scores from 10 repeats of 10-fold cross-validation using different sample sizes with the optimal active sampling prototype and random sampling.

5.3.2 Final Validation Results

Hyperparameter Tuning Results

The best hyperparameters found for training the baseline ABMIL model were a learning rate of 0.0038, weight decay of 0.00079, and dropout rate of 0.020 (Table 5.2). For active sampling, the best hyperparameters were 16 resampling iterations (40 patches per iteration before the final sample of 160 patches), 64 sampling nearest neighbours, and a random sampling proportion of 29% with this reduced to 0% after the first iteration. The number of sampling neighbours and sampling iterations each took the greatest values available, with no greater options tested given the increased computational requirements.

Subtyping Results

Evaluation Method	Balanced Accuracy	AUROC	F1 Score
Full ABMIL Evaluation	**80.08%**	**0.8781**	**0.7472**
Active Sampling	79.07%±0.69%	0.8679±0.0035	0.7337±0.0093
Random Sampling	78.94%±0.66%	0.8659±0.0034	0.7320±0.0089

Table 5.3 3-fold cross-validation binary classification results using different evaluation approaches with the same ABMIL baseline model (mean ± one standard deviation from 100,000 iteration bootstrapping of 50 repeats). The best results are indicated in **bold**.

The baseline method of full ABMIL evaluation without sampling gave the best classification performance, outperforming active sampling by approximately 1% for each metric (Table 5.3). Active sampling slightly outperformed random sampling for each metric, though these differences were not significant and performance was not consistently better across folds. The median AUROCs for DRAS-MIL across the three folds were 0.806, 0.932, and 0.895, compared to the random sampling scores of 0.804, 0.926, and 0.896 (Figure 5.5). Random sampling marginally outperformed active sampling in the third fold, though the difference was small enough that this may be attributed to chance.

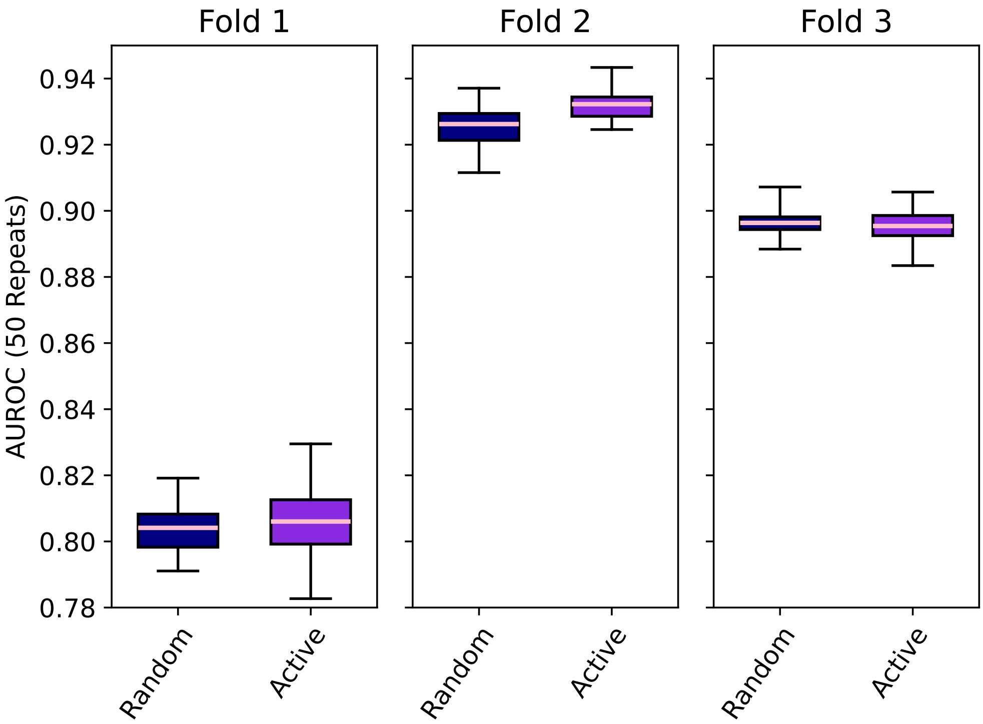


Figure 5.5 Boxplots comparing the AUROC scores from 50 repeats of each cross-validation fold for random and active sampling with 800 patches per slide.

Efficiency Results

On the computational benchmarking dataset of 50 randomly selected WSIs, active sampling reduced GPU memory utilisation from a maximum of 340MB to 60MB (Table 5.4). The best total run time for active sampling was 47 minutes compared to 140 minutes for default ABMIL processing. This represents approximately 56s per WSI for active sampling and 168s per WSI for default ABMIL classification. The difference in run times was much greater when evaluating only using a CPU, with total times of 4h 8m for active sampling and 30h 32m for full MIL evaluation, representing approximately 5 minutes and 37 minutes per slide, respectively. Overall, DRAS-MIL reduced GPU memory requirements by at least 82% and inference time by 67% when using a GPU, and by 86% when using the CPU alone.

Inference Method	Batch Size	Total Inference Time for 50 WSIs	Average Inference Time Per WSI	Maximum GPU Memory Utilisation
ABMIL	1	4h 37m	332.0s	**340MB**
	4	2h 46m	198.9s	342MB
	8	2h 33m	183.8s	356MB
	16	2h 24m	173.3s	471MB
	32	**2h 20m**	**167.8s**	702MB
	64	2h 21m	169.3s	1163MB
DRAS-MIL	1	**47m**	**56.0s**	**60MB**
	4	47m	56.2s	103MB
	8	49m	58.6s	161MB
	16	53m	63.5s	275MB
	32	58m	69.4s	506MB
	64	1h 1m	73.6s	967MB
ABMIL (CPU)	32	30h 32m	2198.1s (36m 38s)	0MB
DRAS-MIL (CPU)	1	4h 8m	298.0s (4m 58s)	0MB

Table 5.4 Inference efficiency on a subset of 50 WSIs. Each experiment was repeated three times and the median value was taken. All experiments used a GPU except those labelled *CPU*. The best results are indicated in **bold**.

5.4 Discussion

The results indicated that active sampling could drastically reduce the computational requirements for WSI inference with minimal impact on the classification accuracy. Completely random sampling also retained classification performance, with active sampling only performing marginally better. While random sampling appears to be a viable approach, DRAS-MIL has the advantage that its sampling maps improve model interpretability as an efficient proxy to the ABMIL attention heatmap (Figure 5.6).

The relatively high classification performance of the sampling approaches compared to whole slide processing may have been influenced by the slides in the dataset containing a relatively high proportion of tumour tissue, making it likely that tumour tissue would be sampled by chance. This is supported by the relatively low level of random sampling used in the tuned DRAS-MIL model, which may indicate that it was possible to find diagnostically relevant tissue without extensive exploration. Such approaches are likely to be inappropriate in highly heterogeneous samples, or those with relatively small regions of interest, where the sampling approach may not find sufficient relevant tissue. It remains to be seen whether pathologists will support the use of models that do not thoroughly analyse all available tissue, and further investigations using varied datasets will be required to understand whether the sampling procedure is sufficiently robust.

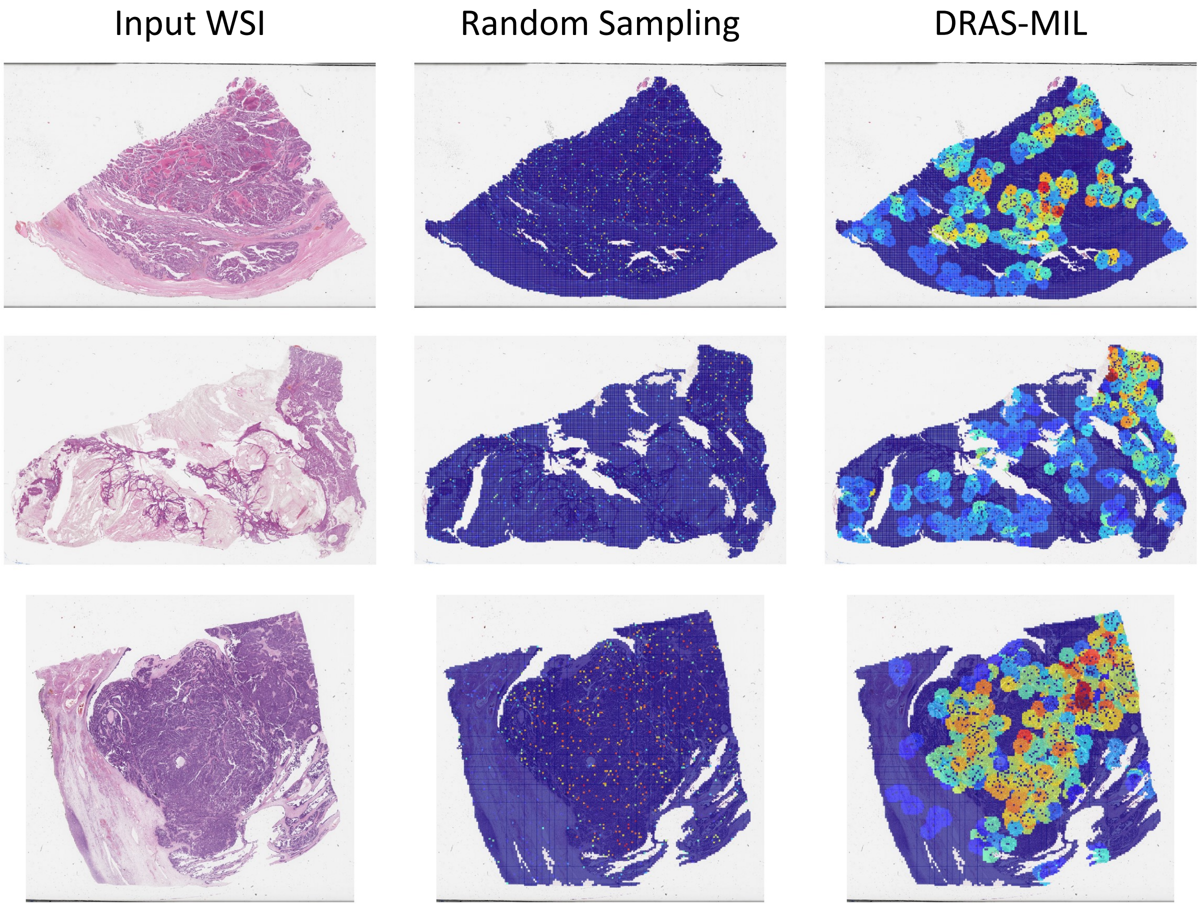


Figure 5.6 Example WSIs with corresponding random sampling attention maps and DRAS-MIL sampling weight maps. Each method takes 800 sample patches per WSI. Red indicates a higher attention/weight, and blue indicates a lower attention/weight.

The main limitation of this analysis was the dataset. The inclusion of non-adnexal tissue was a potential source of bias as the model may have learned shortcuts based on the heterogeneous tissue types, rather than based on the morphological tumour subtypes. This was a particularly limiting factor given the relatively small dataset used, with only a handful of examples given for some tissue types. As such, the non-adnexal tissue WSIs were removed from the dataset for all subsequent research. Without any hold-out or external validations, it was unclear whether the sampling method was robust to different sources of variability. The models in this chapter were also limited to performing binary classification of the most common subtype. Considering HGSC composes around 70% of all ovarian cancers, a binary classifier may be beneficial as an ancillary tool to rapidly confirm the majority of diagnoses, though this utility is limited compared to the target five-class classification, which would cover over 90%

of all ovarian cancers. To truly achieve utility, these models would likely also need to indicate when a WSI does not fit one of these subtypes.

Sampling during inference could be useful in the clinical utility of classifiers, with a model being trained on an HPC in a research setting, then then deployed to the clinic for slide inference. If these models could be feasibly run on standard desktop computers or integrated within the slide scanner hardware it would reduce the investment burden and avoid the need to share data outside of the pathology lab. It is important to focus on reducing barriers to clinical implementation of computer-aided diagnostic tools as the underlying models are increasingly being shown to work at an expert level of accuracy in research settings and are receiving regulatory approval [52], but they are not being widely adopted. In particularly under-resourced settings, it may be worthwhile to trade a small level of diagnostic accuracy for drastically improved efficiency, as this may be the only feasible way to access an expert-level second opinion. However, such low-resource settings may struggle to access digital pathology infrastructure.

5.5 Conclusion

In this chapter, we proposed an active patch sampling approach called DRAS-MIL. This utilised the attention mechanism of ABMIL to generate sampling weights for tissue regions, allowing diagnostically relevant tissue regions to be discovered and leveraged without processing the whole slide. This drastically reduced inference time with only a marginal impact on the discriminative ability of the classifier. Completely random sampling was found to give a discriminative performance almost as great as DRAS-MIL, though it reduced the interpretability of the corresponding attention heatmaps. Improvements in the efficiency of slide classification inference will be essential if models are to be deployed directly to the pathology clinic given the limited computational resources and, in this chapter, we have found that active patch sampling is a promising approach. However, the underlying models did not offer sufficient performance to be considered for clinical deployment, with performance limited by a relatively small, heterogeneous dataset.

Chapter 6

Analysis of Tissue Magnifications

In this chapter we describe our first five-class subtype classification model, using the same ABMIL backbone from the previous chapter with a larger training dataset and a hold-out test set. We investigate how the tissue magnification used in these models impacts performance, both in terms of classification accuracy and computational efficiency. Through this chapter, we formalise the iterative hyperparameter tuning and validation procedures used in subsequent research.

6.1 Introduction

While MIL approaches have become increasingly common for WSI-level classification tasks, it is not clear which tissue magnification is optimal for computational analysis. Higher magnifications provide more cellular-level detail, whereas lower magnifications offer greater architectural context at the tissue level. Pathologists typically assess a slide at multiple magnifications, with slides scanned at 40x magnification to facilitate the highest required cellular resolution. However, the requirements of an AI model are likely to differ from those of a human pathologist since a model can thoroughly process all available information at the pixel level.

The optimal magnification likely depends upon the given task. Comparative studies have often found either 10x or 20x magnification to be best, with 10x reported as best for bladder cancer subtyping [82], cervical lymph nodes metastasis classification [219], cervical cancer prognostication [219], and melanoma immune subtyping [81], and 20x reported to be best for lymphoma subtyping [220], breast cancer lymph node metastasis classification [219], and lung cancer subtyping [219]. However, most comparisons were made between only two different magnifications, which often appeared to be arbitrarily selected.

The optimal magnification may also depend on the specific models used. For example, in the aforementioned melanoma immune subtyping study, 10x was determined to be the best overall, but each of the other evaluated magnifications (2.5x, 5x, 20x, 40x) achieved the greatest performance in at least one experiment using different feature extraction and MIL classification models [81]. Further, in two studies classifying the primary cancer type in TCGA WSIs, one study found 10x to be best [221] and the other found 20x to be best [222]. Both of these studies found 20x to be best specifically for distinguishing ovarian cancer from other primary cancers, though both studies used small ovarian cancer cohorts (<100 WSIs).

Previous MIL models for ovarian cancer subtyping have primarily used tissue patches at 20x magnification [5, 134, 136, 160, 171, 173], with relatively few studies using different magnifications [118, 123, 136]. The effects of the different magnifications on classification performance have not been directly compared, though one study did find that a multi-scale attention model applied different levels of attention to different magnifications (5x, 10x, 20x) for different subtypes [82], reflecting the inherent differences between magnifications.

A key limitation in previous research has been the lack of separate hyperparameter tuning at each tissue magnification, making it likely that some models underperformed due to using sub-optimal hyperparameters for the given magnification. In this chapter, we present the most extensive analysis of the effects of tissue magnification on ovarian cancer subtyping to date. A separate ABMIL classifier was trained for five-class subtyping at each of six magnifications from 1.25x to 40x. Hyperparameter tuning was performed separately at each magnification and both the classification performance and model efficiency were evaluated using cross-validation and a hold-out test set.

6.2 Methods

6.2.1 Classification Methodology

To compare the effects of tissue magnifications on subtype classification performance, we used the baseline model training and hyperparameter tuning protocols (Section 4.1) for the original 40x WSIs, as well as for downsampled WSIs at 20x, 10x, 5x, 2.5x, and 1.25x magnification. Patches were extracted such that the resulting patch size was 256 x 256 pixels after downsampling (for example, 8192 x 8192 pixel patches were downsampled to 256 x 256 for the 1.25x experiments given the 32x downsampling factor). Downsampling by a factor of two reduced the overall slide

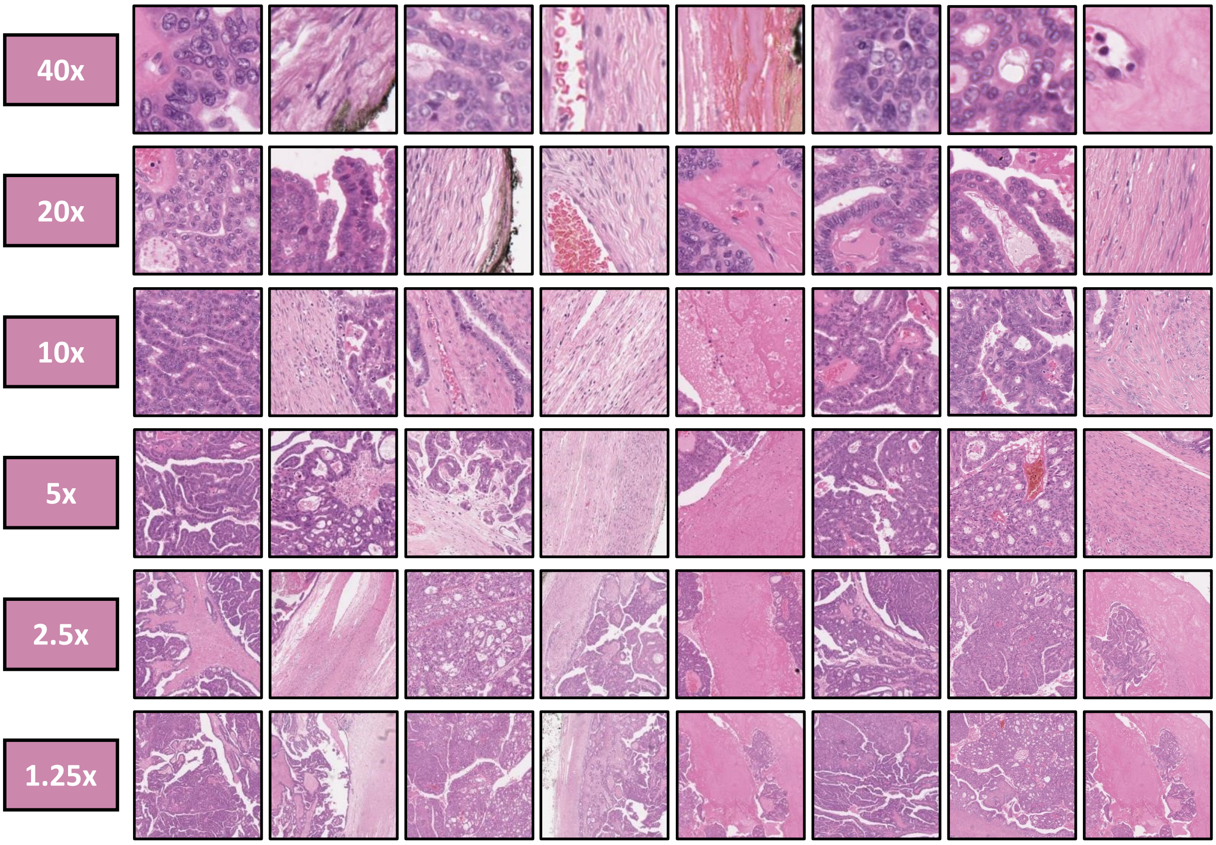


Figure 6.1 Examples of 256 x 256 pixel tissue patches from a single WSI (shown in Figure 4.1) at six different apparent magnifications after downsampling from the 40x native magnification.

area by a factor of four $\left(\frac{\text{height}}{2} \times \frac{\text{width}}{2}\right)$, and so also reduced the number of patches by an approximate factor of four. On average there were 68,913 patches per slide at the highest magnification (40x), compared to only 81 at the lowest (1.25x). As in the standard baseline model, an ImageNet-pretrained ResNet50 encoder was used to extract 1 x 1024 feature vectors from the downsampled 256 x 256 pixel tissue patches, and these features were used to train ABMIL classifiers [77] for ovarian cancer subtyping.

6.2.2 Tuning and Validation

Iterative grid hyperparameter tuning was used, with eight hyperparameters tuned until the validation loss stopped improving (Table 6.1). The first three stages of tuning were limited to 30 epochs of model training, the subsequent six stages to 100 epochs, and the final four to 150 epochs. The tuned hyperparameters were the standard Adam optimizer hyperparameters (learning rate, first and second moment decay, stability parameter) and regularisation hyperparameters (weight decay, dropout rate, max patches per slide) described in Section 4.3, as well as a **model size** hyperparameter controlling the dimensions of the attention layer and subsequent fully connected layer in the classifier. Initial hyperparameters were influenced by the CLAM default hyperparameters [78] and by the experiments in Chapter 5. Approximately 80 unique configurations were evaluated for classification at each tissue magnification.

	Tuning Iteration												
Hyperparameter	1	2	3	4	5	6	7	8	9	10	11	12	13
Learning Rate	✓			✓		✓		✓		✓			
First Moment Decay		✓											✓
Second Moment Decay		✓											✓
Stability Parameter							✓						
Weight Decay			✓								✓		
Dropout Rate			✓	✓		✓			✓	✓			
Max Patches	✓			✓		✓				✓		✓	
Model Size					✓								

Table 6.1 Iterative hyperparameter tuning procedure. Tick marks indicate which hyperparameters were adjusted at each stage of tuning, with all other hyperparameters frozen at their previous best values.

Carcinoma Subtype	**Training WSIs (Patients)**	**Hold-out WSIs (Patients)**
High-Grade Serous (HGSC)	484 (107)	20 (7)
Low-Grade Serous (LGSC)	23 (5)	20 (6)
Clear Cell (CCC)	156 (33)	20 (7)
Endometrioid (EC)	205 (36)	20 (5)
Mucinous (MC)	95 (20)	20 (5)
Total	**963 (201)**	**100 (30)**

Table 6.2 Dataset breakdown showing the number of primary resection WSIs per ovarian carcinoma subtype in the training (cross-validation) and independent hold-out test sets for the analysis of tissue magnification. Numbers in brackets indicate the number of unique patients.

Five-fold cross-validation was conducted with a training set of 963 primary resection adnexal specimen WSIs from 201 patients (Table 6.2), stratified at the patient level. An average ensemble of the five-fold classifiers at each magnification was evaluated using a class-balanced hold-out test set of 100 WSIs from 30 patients. Both datasets were part of the internal LTHT dataset (Section 4.2).

As well as using the standard metrics to measure discriminative power (balanced accuracy, AUROC, F1 score), we measured the classification efficiency through the average training and inference times on the HPC, and the average inference time on the PC (Section 4.5). Training times were measured as the average time to train a classifier on a single cross-validation fold, whereas inference times were measured as the average time to classify a slide using a class-balanced subset of 20 WSIs from the hold-out test set. Preprocessing and feature extraction times were excluded from training but included in inference times to represent the computational burden of a deployed model. PyTorch-based code was made available at https://github.com/scjjb/Ovarian_Subtype_Mags.

6.3 Results

6.3.1 Hyperparameter Tuning Results

	Magnification					
Hyperparameter	**40x**	**20x**	**10x**	**5x**	**2.5x**	**1.25x**
Learning Rate	1e-3	5e-4	5e-4	5e-4	1e-3	5e-4
First Moment Decay	0.95	0.99	0.8	0.95	0.9	0.9
Second Moment Decay	0.99	0.99	0.99	0.999	0.9999	0.999
Stability Parameter	1e-10	1e-8	1e-4	1e-14	1e-4	1e-14
Weight Decay	1e-4	1e-4	1e-4	1e-6	1e-5	1e-5
Dropout Rate	0.6	0.7	0.6	0.6	0.7	0.5
Max Patches	50000	8000	1000	400	40	7
Model Size	512, 256	256, 64	256, 64	128, 32	256, 64	256, 64

Table 6.3 The hyperparameters used to train the final model at each resolution, determined through the iterative hyperparameter tuning procedure to minimise the average balanced cross-entropy validation loss in 5-fold cross-validation.

The optimal hyperparameters found through hyperparameter tuning are shown in Table 6.3 for each magnification. The clearest trend was that the optimal number of patches used in training at lower magnifications typically covered a lower proportion of the entire slide area, with the 7 patches at 1.25x covering only 9% of an average slide, compared to the 50,000 patches at 40x covering 73% of an average slide. The size of the optimal model weakly corresponded to the magnification, with the 40x model being the largest (795,408 parameters), and the 5x model being the smallest (141,424 parameters). Other hyperparameters did not exhibit clear trends across magnifications.

6.3.2 Magnification Validation Results

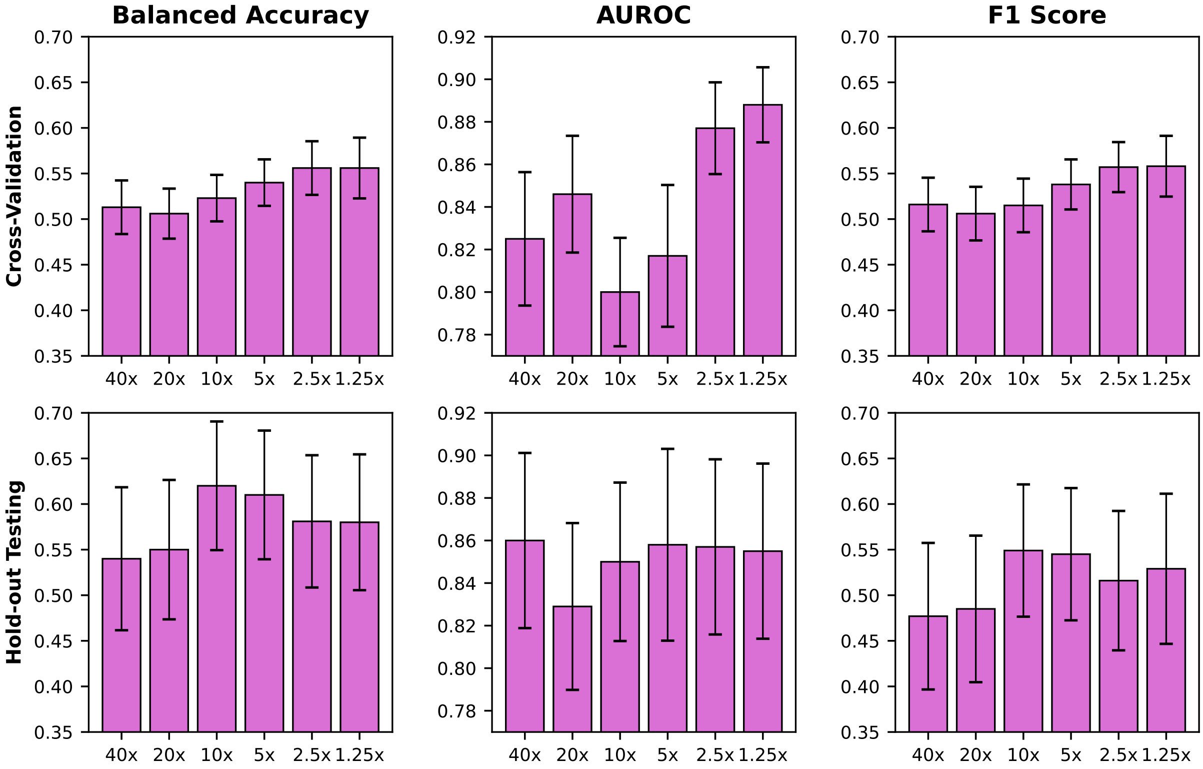


Figure 6.2 Classification performance at each magnification in cross-validation and hold-out testing, with error bars indicating the 95% confidence intervals from 10,000 iteration bootstrapping.

No single magnification gave the best classification results across all metrics in both validations (Figure 6.2). In cross-validation (Table 6.4), the greatest classification performance across the five folds was achieved by the 1.25x magnification model, with 55.6% balanced accuracy, 0.888 AUROC, and 0.558 F1 score. Performance did not vary greatly between magnifications, with the lowest balanced accuracy being 50.6% (at 20x), the lowest AUROC being 0.800 (at 10x), and the lowest F1 being

0.506 (at 20x). 95% confidence intervals for the balanced accuracy and F1 score were overlapping for all models, and 95% confidence intervals for AUROC were also overlapping for most models.

In hold-out testing (Table 6.5), the greatest classification performance was achieved by the 10x model, with 62.0% balanced accuracy, 0.850 AUROC, and 0.549 F1 score. The 5x model performed similarly, with a marginally lower balanced accuracy and F1 score, but a marginally higher AUROC. Performance varied more across magnifications than in cross-validation, with the lowest balanced accuracy being 54.0% (at 40x), the lowest AUROC being 0.829 (at 20x), and the lowest F1 score being 0.477 (at 40x). However, the 95% confidence intervals were all overlapping for each metric (Figure 6.2).

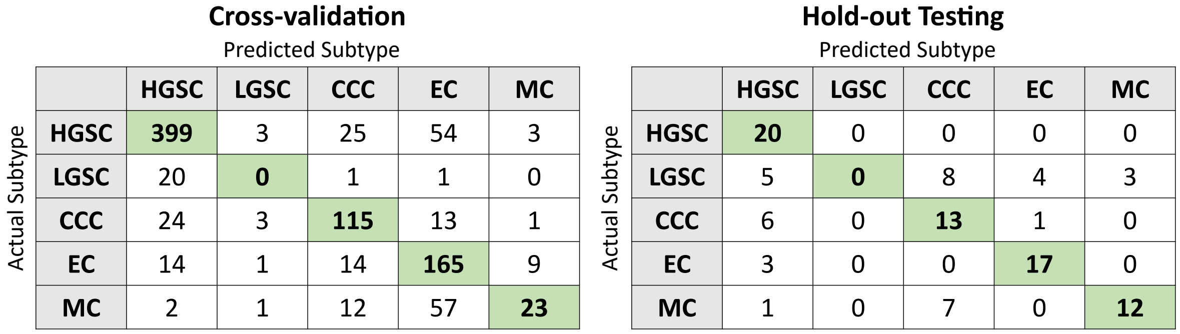
Magnif.	**Balanced Accuracy**	**AUROC**	**F1 Score**
40x	51.3% (48.4-54.2%)	0.825 (0.794-0.856)	0.516 (0.487-0.545)
20x	50.6% (47.9-53.3%)	0.846 (0.819-0.873)	0.506 (0.477-0.535)
10x	52.3% (49.8-54.8%)	0.800 (0.775-0.825)	0.515 (0.486-0.544)
5x	54.0% (51.5-56.5%)	0.817 (0.784-0.850)	0.538 (0.511-0.565)
2.5x	**55.6%** (52.7-58.5%)	0.877 (0.855-0.899)	0.557 (0.530-0.584)
1.25x	**55.6%** (52.3-58.9%)	**0.888** (0.870-0.906)	**0.558** (0.525-0.591)

Table 6.4 Classification results from five-fold cross-validation at each magnification. Results are reported as the mean and 95% confidence intervals (in brackets) from 10,000 iterations of bootstrapping. The greatest results are shown in **bold**.

Magnif.	**Balanced Accuracy**	**AUROC**	**F1 Score**
40x	54.0% (46.2-61.8%)	**0.860** (0.819-0.901)	0.477 (0.397-0.557)
20x	55.0% (47.4-62.6%)	0.829 (0.790-0.868)	0.485 (0.405-0.565)
10x	**62.0%** (54.9-69.1%)	0.850 (0.813-0.887)	**0.549** (0.476-0.622)
5x	61.0% (53.9-68.1%)	0.858 (0.813-0.903)	0.545 (0.472-0.618)
2.5x	58.1% (50.8-65.4%)	0.857 (0.816-0.898)	0.516 (0.440-0.592)
1.25x	58.0% (50.6-65.4%)	0.855 (0.814-0.896)	0.529 (0.447-0.611)

Table 6.5 Classification resulting from hold-out testing at each magnification, with predictions generated by an ensemble of the five-fold classification models. Results are reported as the mean and 95% confidence intervals (in brackets) from 10,000 iterations of bootstrapping. The greatest results are shown in **bold**.

Performance varied drastically between subtypes in both validations (Figure 6.3). The optimal model from hold-out testing (10x magnification) failed to correctly classify the least common subtype (LGSC) a single time in either validation, whereas it classified the most common subtype (HGSC) with F1 scores of 0.846 and 0.727 in cross-validation and hold-out testing, respectively.



Cross-validation

Predicted Subtype

Actual Subtype

	HGSC	LGSC	CCC	EC	MC
HGSC	**399**	3	25	54	3
LGSC	20	**0**	1	1	0
CCC	24	3	**115**	13	1
EC	14	1	14	**165**	9
MC	2	1	12	57	**23**

Hold-out Testing

Predicted Subtype

Actual Subtype

	HGSC	LGSC	CCC	EC	MC
HGSC	**20**	0	0	0	0
LGSC	5	**0**	8	4	3
CCC	6	0	**13**	1	0
EC	3	0	0	**17**	0
MC	1	0	7	0	**12**

Figure 6.3 Confusion matrices of the optimal ABMIL model from hold-out testing (10x magnification). Correct classifications are indicated in **bold**.

6.3.3 Efficiency Evaluations

The fastest models to train were the 5x (10m 37s) and 1.25x (10m 49s) magnification models (Table 6.6). Model training was always faster at lower magnifications, with the exception of the 5x model, which trained faster than the 2.5x or 1.25x models due to having fewer parameters. The fastest models in inference were the 10x model (1m 9s per slide on HPC, 3m 8s on PC), and the 5x model (1m 11s per slide on HPC, 3m 15s on PC). The 40x model was the slowest by a wide margin for both training (3h 45m) and inference (3m 51s per slide on HPC, 7m 29s on PC). The highest magnification models were slowed by the slides containing more patches, and the lowest magnification models by the larger size of patches before downsampling, which took slightly longer to process. The large patch size would not be a factor for slides with a lower native magnification, where the lower magnifications would be expected to always be faster during inference. Such lower magnifications may be accessible at lower levels of the pyramidal WSI file, though we chose to downsample from 40x for consistency as not all WSI had the same lower magnifications available.

Magnif.	**Training Time per Fold**	**Inference Time per Slide (HPC)**	**Inference Time per Slide (PC)**
40x	3h 45m 13s	3m 51s	7m 29s
20x	44m 23s	1m 20s	3m 56s
10x	12m 51s	**1m 9s**	**3m 8s**
5x	**10m 37s**	1m 11s	3m 15s
2.5x	12m 29s	1m 24s	4m 28s
1.25x	10m 49s	1m 58s	5m 42s

Table 6.6 Average training and inference times. Inference times were analysed using a balanced subset of 20 WSIs from the hold-out test set on a high-performance computer (HPC) and a personal computer (PC). The best times are indicated in **bold**.

6.4 Discussion

Our results indicate that the standard 40x and 20x magnifications used in the clinical setting may not be optimal for computational ovarian cancer subtyping, with 10x and 5x each giving greater balanced accuracies in cross-validation and hold-out testing while also running much faster. The balanced accuracy and F1 scores reported in this chapter were lower than those in other studies using similar methodologies [82, 123, 136], likely due to the few available LGSC cases leading to very poor performance at classifying this specific subtype. It is, however, worth noting that the models have demonstrated discriminative power, with the optimal balanced accuracies of 56% and 62% in cross-validation and hold-out testing being much greater than the 20% baseline from random 5-class classification.

The differences between results in the cross-validation and hold-out test sets were likely influenced by the imbalance in the cross-validation set, with AUROC particularly affected by class imbalance. The small size of the test sets caused a degree of uncertainty in the results, reflected in the wide confidence intervals (Figure 6.2). As such, most differences between results were not statistically significant, meaning we cannot be confident that the improved classification at lower magnifications was not caused by random chance. Differences in the efficiency of models were much clearer, with the most efficient models running over twice as fast as the slowest across all evaluations. While classification accuracies are highly dependent on specific datasets and modelling decisions, the improved efficiency is likely to hold across a range of datasets and models.

Given the black-box nature of the deep learning models used in this analysis, it is not entirely clear why performance was best at lower magnifications. The 40x diagnostic standard allows pathologists to thoroughly assess aspects of cellular and nuclear morphology that are less readily identifiable at lower magnifications, though it may be that these features can be sufficiently analysed at a lower magnification by computer vision models which can interrogate images down to a single pixel. Wider histoarchitectural patterns and wider chromatic patterns may be possible to assess at lower magnifications, and these lower magnifications may benefit the ABMIL model by giving a wider context window within each patch. It may be the case that no single magnification is optimal for the computational analysis of all subtypes. A recent study [82] suggested that higher magnifications may be best for classifying LGSC, where cellular features are necessary to determine the low grade, and lower magnifications may be best for classifying EC, where the characteristic glandular architecture may be readily assessed. The 10x magnification, which performed best in our hold-out testing, may have offered the best trade-off between relevant features from higher and lower magnifications, not so low as to lose relevant cellular and nuclear features, and not so high as to miss the spatial context.

This investigation was limited to a single model type (ABMIL) which processed tissue at a single magnification. Different models, such as graphs and transformers, may perform differently due to their ability to model patch relationships (Section 2.4.5). Multi-magnification models [82, 118] may be able to improve performance by combining the cellular and histoarchitectural information from different magnifications, akin to how pathologists interpret morphological features from multiple magnifications to obtain overall diagnostic insight. This analysis was also limited to a single data centre, so it is unclear how model generalisability may be affected by tissue magnification. Implementation of these models will require improved accuracy and more extensive validation to ensure model robustness.

While the routine collection of data at a lower native magnification would offer efficiency benefits (reduced storage requirements and cost), this is unlikely to be clinically viable as pathologists require higher magnifications for manual review. The benefit of lower magnification modelling (aside from improved classification performance) is in the reduced computational requirements, allowing models to be deployed directly to the pathology lab, avoiding any diagnostic delay and financial expense from exporting samples off-site.

6.5 Conclusion

In this chapter, we reported the most extensive evaluation of tissue magnifications for ovarian cancer subtyping conducted to date. We tuned ABMIL classifiers at six different magnifications from 1.25x to 40x and found that downsampling to reduce the apparent magnification from the standard 40x did not significantly degrade discriminative performance, and in many cases, slightly improved it. The 10x and 5x magnifications gave the greatest balanced accuracy in hold-out testing while also being the fastest models in training and inference, classifying a slide in little over 3 minutes on a desktop computer with a consumer-grade GPU. The classification performance of even lower magnification models was also impressive, with 1.25x magnification outperforming 40x and 20x by most metrics. The vastly reduced computational burden of lower magnification models may allow them to be deployed directly to the clinic, but classification accuracy and validation must first be improved.

Chapter 7

Analysis of Histopathology Feature Extraction Models

In this chapter we present our most comprehensive analysis of ovarian cancer subtype classifiers, with ABMIL models trained using seventeen different feature extractors on the final version of the Leeds ovarian carcinoma dataset, then validated through hold-out testing and two external validations. Where previous classifiers used ImageNet-pretrained CNNs for feature extraction, here we assess fourteen different histopathology foundation models and an ImageNet-pretrained ViT to determine the extent of the benefits that can be obtained through the newer architecture and domain-specific pretraining. We perform an ablation study to quantify any benefits from hyperparameter tuning, and we investigate the relationship between classification performance and model efficiency. Finally, pathologists review the best models, assessing the ABMIL attention heatmaps, and determining the potential causes of misclassified slides.

7.1 Introduction

Considering the vast size of histopathology WSIs, it is often impractical to train slide-level classifiers end-to-end. It is common for these classifiers to instead be built using frozen pre-trained patch encoders. As such, any limitation in the pretrained feature extractor can limit the final classification performance. In applying MIL to WSI-level classification, many researchers have used ImageNet-pretrained ResNet CNNs [59] for patch feature extraction [5, 71, 78, 81, 134, 223]. ImageNet [58] is a huge set of labelled natural images, making it very popular for model pretraining (Section 2.4.2). However, the resulting generic features are likely to be suboptimal and computationally inefficient when applied to histopathology images, which contain a relatively homogeneous and restricted set of shapes and colours, with subtle differences being relevant to diagnostic decisions [28, 203].

Recently, many researchers have attempted to create histopathology *foundation models*, using self-supervised learning (SSL) techniques to generate broad histopathological feature sets which are not specific to a single organ/cancer type. The scale of these approaches has grown rapidly, from tens of thousands of WSIs used to train models with tens of millions of parameters in 2022 and early 2023 [83, 224–228] to millions of WSIs [47, 229, 230] and billions of parameters more recently [173, 231]. Foundation models have typically been based on ViTs, utilizing the impressive scalability of transformers seen across many fields, most notably with large language models [232, 233].

Histopathology foundation models have exhibited impressive performance across diverse tasks [208, 226, 234, 235] including ovarian cancer subtyping [172, 173], although analyses have been relatively shallow, without thorough hyperparameter tuning and rigorous statistical comparison of the resulting classifiers. Consequently, it is unclear whether models were applied optimally (especially those exhibiting sub-optimal performance), and whether the differences between them were significant. Furthermore, many analyses have been conducted using single-centre data, limiting the assessment of models' generalisability.

In this chapter, we present the most comprehensive validation conducted to date comparing feature extraction methods for ovarian cancer subtyping, including three ImageNet-pretrained feature extractors and fourteen histopathology foundation models. The analysis includes rigorous hyperparameter tuning and evaluations through five-fold cross-validation, hold-out testing, and external validations, and was conducted with the largest collection of ovarian cancer WSIs used in any AI validation to date. We aim to quantify any benefit of foundation models for this task and to find which feature extractors give the best trade-off between diagnostic accuracy and computational efficiency. We further investigate whether the classification performance of the ImageNet-pretrained ResNet50 features can match those of the foundation models through stain normalisation, tissue augmentation, or different tissue detection techniques.

7.2 Methods

7.2.1 Slide Classification Pipeline

Slide classification was performed using the baseline ABMIL classification pipeline (Section 4.1), with different frozen patch feature extractors applied to 256 x 256 pixel patches at 10x apparent magnification since this gave the best accuracy and efficiency in Chapter 6. The patches were preprocessed following the specific procedure of each feature extraction model, which typically involved first applying the standard normalisation to the RGB colour channels (Section 4.1), and for ViT-based models typically also involved resizing or cropping patches to 224 x 224 pixels. Patch features were then used to train an ABMIL classifier for each feature extractor.

Analyses were conducted using the full LTHT training set of 1864 ovarian carcinoma WSIs of adnexal tissue from 433 cases. Internal validations were performed using a five-fold cross-validation on the training set, as well as a five-model average ensemble on a hold-out test set of 100 WSIs from 30 patients. External validations were performed using the same ensembling approach on a set of 80 WSIs from 80 patients in the Transcanadian Study, and a set of 513 WSIs from an unknown number of patients in the OCEAN Challenge. These datasets are described further in Section 4.2.

7.2.2 Feature Extraction Models

A total of seventeen patch feature extractors were compared (Table 7.1), three of which had been trained through the traditional approach of supervised classification on ImageNet data [58], and the other fourteen had been trained using histopathology images through various SSL approaches. All feature extractors were available online, with some requiring approval before they could be accessed.

The ImageNet-pretrained models were a ResNet50 [59], ResNet18 [59], and a large vision transformer (ViT-L) [62]. The ResNet50 outputs were taken from the end of the third residual block (as in CLAM [78]) to give 1024 features per input patch. The ResNet18 does not have a layer this large, so 512 features were extracted from the end of the fourth residual block instead. ViT-L was applied without a final fully connected layer to give 1024 features per patch. ImageNet-pretraining for ResNet models had

Feature Extractor	Backbone	Data Type	Data Source	Pretraining Algorithm	Pretraining Images	Pretraining Magnification(s)	Parameters	Patch Features
RN50 [59]	ResNet50	Natural	ImageNet-1k	Supervised	1,431,167	NA	8,543,296	1024
RN18 [59]	ResNet18	Natural	ImageNet-1k	Supervised	1,431,167	NA	11,176,512	512
ViT-L [62]	ViT-L	Natural	ImageNet-21k	Supervised	14,197,122	NA	303,301,632	1024
RN18-Histo [224]	ResNet18	Histology	57 Open Sets	SimCLR	>25,000 WSIs	10x,20x,40x,100x	11,176,512	512
Lunit [226]	ViT-S	Histology	TCGA + Internal	DINO	36,666 WSIs	20x,40x	21,670,272	384
RN50-Histo [226]	ResNet50	Histology	TCGA + Internal	Barlow Twins	36,666 WSIs	20x,40x	23,508,032	2048
CTransPath [225]	CNN + SwinT	Histology	TCGA + PAIP	Novel SSL	32,220 WSIs	20x	27,520,038	768
Hibou-B [229]	ViT-B	Histology	Internal	DINOv2	1,141,581 WSIs	Unclear	85,741,056	768
Phikon [227]	ViT-B	Histology	TCGA	iBOT	6,093 WSIs	20x	85,798,656	768
Kaiko-B8 [236]	ViT-B	Histology	TCGA	DINO	∼29,000 WSIs	5x,10x,20x,40x	85,807,872	768
GPFM [172]	ViT-L	Histology	47 Open Sets	Novel Distillation	72,280 WSIs	Unclear	303,228,928	1024
UNI [208]	ViT-L	Histology	Internal + GTEx	DINOv2	100,426 WSIs	20x	303,350,784	1024
Hibou-L [229]	ViT-L	Histology	Internal	DINOv2	1,141,581 WSIs	Unclear	303,659,264	1024
Virchow [230]	ViT-H	Histology	Internal	DINOv2	1,488,550 WSIs	20x	631,229,184	2560
Virchow2-CLS [47]	ViT-H	Histology	Internal	DINOv2	3,134,922 WSIs	5x,10x,20x,40x	631,239,424	1280
H-optimus-0 [231]	ViT-g	Histology	Internal	DINOv2	>500,000 WSIs	20x	1,134,774,272	1536
Prov-GigaPath [173]	ViT-g	Histology	Internal	DINOv2	171,189 WSIs	20x	1,134,953,984	1536

Table 7.1 Summary of the seventeen feature extraction models (grouped by the pretraining data type and ordered by model size).

been conducted using the original 1,000 class ImageNet dataset alone, whereas the ViT-L was first trained on the much larger set of nearly 22,000 classes, and then fine-tuned to the same set of 1,000 classes. The reported ImageNet classification accuracies were 80.9%, 69.8%, and 85.1% for ResNet50 [237], ResNet18 [238], and ViT-L [239], respectively.

The SSL pretraining of the foundation models allowed large quantities of diverse data to be leveraged without the need for extensive labelling. One of the earliest histopathology foundation models was a ResNet18 trained through the SimCLR contrastive learning strategy [240] with 57 open datasets in 2021 [224], which we refer to as ‘RN18-Histo’. A similar approach was taken in a subsequent study to pre-train a ResNet50 with a combination of TCGA and proprietary data using Barlow Twins [241], which we refer to as ‘RN50-Histo’ [226]. Another early approach, CTransPath [225], used a novel backbone which combined a CNN with a Swin Transformer, and pretrained these through a novel SSL strategy using multiple open datasets.

Newer histopathology foundation models have typically used vision transformer backbones. The smallest such model, Lunit [226], used DINO distillation [242] to train a small vision transformer (ViT-S) to create a model of a similar size as RN50-Histo which had been pretrained with the same dataset. Three of the foundation models were built using the base vision transformer (ViT-B) backbone with different pretraining procedures, with Phikon [227] trained using iBOT [243] on a small subset of TCGA data, Kaiko-B8 [236] on a much larger set of TCGA data using DINO [242], and Hibou-B [229] on a huge proprietary dataset using DINOv2 [244]. The authors of Kaiko-B8 also made their model available with four other backbone sizes, though the B8 variation gave the best overall performance in their evaluations [236]. Hibou-B was included as it was the best-available version of this model when initial validations were conducted, although the authors reported their larger model, Hibou-L, to have given better performance [229].

The largest histopathology foundation models (all published in 2024) have typically been vision transformers trained with proprietary datasets of over 50,000 WSIs using DINOv2. GPFM [172], UNI [208], and Hibou-L [229] are large vision transformers (ViT-L) trained with 72,280 WSIs, 100,426 WSIs, and 1,141,581 WSIs, respectively. GPFM was the largest foundation model to not be trained using DINOv2, with a novel distillation method used instead. Virchow [230] and its recent update, Virchow2 [47], are huge vision transformers (ViT-H) trained with the largest dataset for any histopathology foundation model to date, with nearly 1.5m WSIs in the first version and over 3m WSIs in the second version. Virchow also has the largest feature space as the class tokens are concatenated with the average patch tokens from the ViT, where typically only the class tokens would be used. As Virchow2 was reported by the original authors to give better results using just the class tokens [47], we adopted this version as ‘Virchow2-CLS’.

Prov-GigaPath [173] and H-optimus-0 [231] were the largest accessible histopathology foundation models by far, with the ViT-g backbone giving over one billion parameters, nearly twice as many as the next largest model (Virchow2-CLS), and over 100x as many parameters as the smallest foundation model (RN18-Histo). These models had also been trained with hundreds of thousands of WSIs using DINOv2. Prov-GigaPath includes a patch-to-slide aggregator, though we focused only on the patch feature extractor.

7.2.3 Normalisation and Augmentation Analysis

Previous studies have often used normalisations and augmentations to attempt to improve the robustness of models based on ImageNet-pretrained CNNs [245], including models for ovarian cancer subtyping [134, 136]. To investigate whether the baseline ImageNet-pretrained ResNet50 encoder could be made competitive with the modern alternatives, we applied this feature extractor with a variety of data preprocessing techniques, including normalisation, augmentation, and automated saturation thresholding. These were compared to the default preprocessing approaches (Section 4.1).

Otsu thresholding [246] is applied during tissue segmentation to automatically determine the saturation threshold for each image by minimising the variance within the separated high-saturation and low-saturation groups. Saturation thresholding is a computationally efficient tissue segmentation approach, but risks including artefacts such as bubbles, pen marks, and coverslip edges in the foreground region. While more robust (and complex) tissue segmentation techniques exist [247, 248], we focused on simple approaches as the attention mechanism in the classification models should learn to ignore any remaining artefacts. We compared the CLAM [78] default static saturation threshold (8/255) to Otsu thresholding with parameters manually adjusted to qualitatively improve the segmentation (specifically by reducing the strength of median blurring and increasing the strength of morphological closing to reduce separation between small tissue segments).

Normalisation and augmentation techniques control data variability, which is particularly important for generalisability in histopathology, where varied staining and scanning procedures between labs result in chromatic heterogeneity [4]. Normalisation reduces variability, adjusting images into a consistent colour space to allow models to learn general features. We investigated two commonly used [9] stain normalisation techniques - Reinhard normalisation [249] and Macenko normalisation [250]. These approaches work in logarithmic colour spaces, where stains behave linearly, making them easier to separate and manipulate. Reinhard normalisation is a standard normalisation technique applied in $l\alpha\beta$ space (radiance l, blue-yellow α, red-green β). Macenko normalisation uses singular value decomposition in logarithmic RGB space to separate stain and saturation values, before scaling the stain values. Basic RGB normalisations were also applied to all images (after any other colour adjustments) to match the ImageNet pretraining procedure. The normalisation approaches were implemented using the `torchstain` default hyperparameters. A target stain profile was already provided for Macenko normalisation but not for Reinhard normalisation, for which we manually selected a single target image with a standard apparent colour profile. These targets were fixed to give a consistent standardisation across both the training and the validation sets. While many more sophisticated stain normalisation techniques have been developed, it remains unclear whether any such approach is better than Macenko normalisation overall [9].

Augmentation techniques conversely increase the variability of the training data to allow the model to learn a more general domain. For such large images, training models end-to-end to allow for online data augmentation (adjustments during training) is extremely computationally intensive [251]. Some researchers have attempted to apply online augmentations in the embedding space using generative models [223, 252], though this adds an extra layer of complexity to an already resource-intensive model pipeline. Instead, offline augmentation creates a finite set of augmented versions of the original data, artificially increasing the diversity of training data to a lesser extent than online augmentation. We investigated colour augmentations which adjusted the brightness, contrast, saturation and hue of each patch using parameters from a previous study [253], which we found to create plausibly altered colours (Figure 7.1).

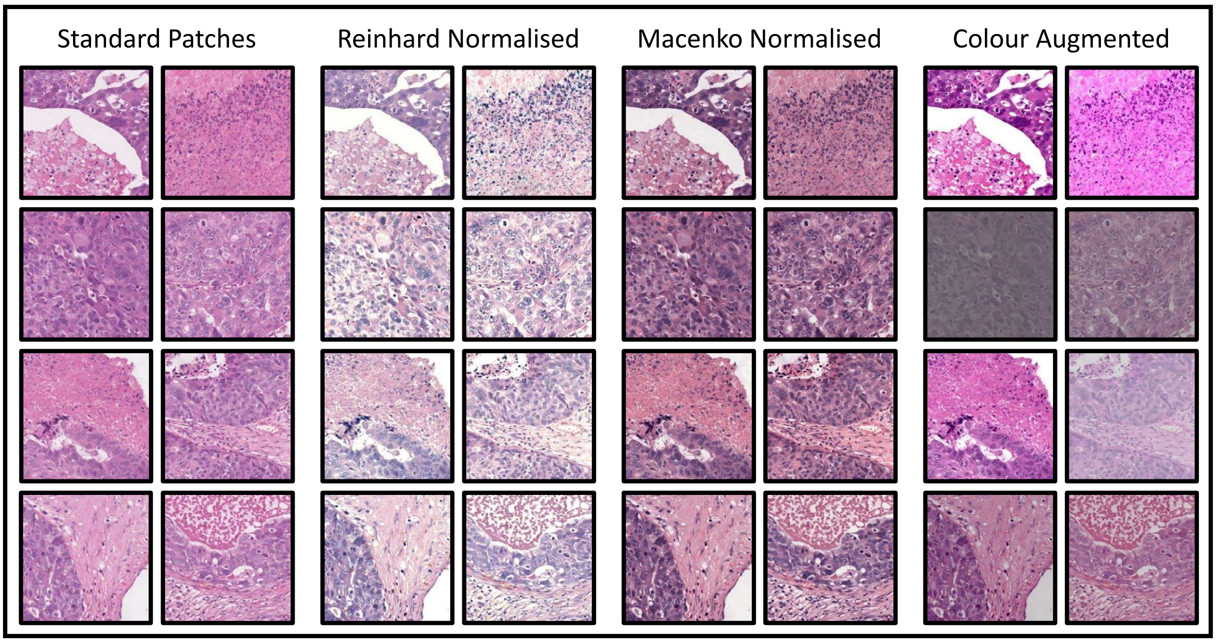


Figure 7.1 Tissue normalisation and augmentation procedures illustrated using 256 x 256 pixel patches from a single whole slide image at 10x magnification.

7.2.4 Hyperparameter Tuning and Validation Procedures

ABMIL classifiers were tuned using the iterative grid search procedure detailed in Section 4.3. Ten hyperparameters were tuned, including the eight hyperparameters tuned in Chapter 6, as well as the learning rate (LR) decay factor and LR decay patience. The initial hyperparameters were taken from the tuned 10x magnification ABMIL model with the ImageNet-pretrained ResNet50 encoder. Through 17 tuning iterations (Table 7.2), over 150 unique hyperparameter configurations were evaluated for each classifier. An ablation study was also conducted to investigate whether hyperparameter tuning improved model performance, with the performance of the tuned models compared to those using the default hyperparameters.

Tuning Iteration	Learning Rate (LR)	LR Decay Patience	LR Decay Factor	First Moment Decay	Second Moment Decay	Stability Parameter	Weight Decay	Dropout Rate	Max Patches	Model Size
1	✓									✓
2								✓	✓	
3				✓	✓					
4	✓						✓			
5				✓		✓				
6									✓	✓
7		✓	✓							
8	✓							✓		
9										✓
10	✓	✓								✓
11								✓	✓	
12		✓	✓							
13	✓									✓
14							✓		✓	
15										✓
16				✓	✓					
17	✓			✓				✓	✓	✓

Table 7.2 Iterative hyperparameter tuning procedure, with check marks (✓) indicating the hyperparameters that were adjusted at each stage of tuning, with all others frozen. These are grouped into hyperparameters relating to the learning rate, Adam optimizer, regularisation, and model architecture.

Paired *t*-tests were used to test for statistically significant differences in the discriminative performance of each model compared to the baseline ResNet50 across the five cross-validation folds, with p-values adjusted for multiple testing using a false discovery rate correction [206]. Results were considered *statistically significant* given an adjusted p-value < 0.05. Paired *t*-tests were also used in the hyperparameter tuning ablation to determine whether tuning the ABMIL classifiers had a statistically significant effect on the final results. Model efficiency was evaluated as the average time to preprocess and classify a WSI using a consistent class-balanced set of 20 WSIs from the internal hold-out test set, with the evaluation repeated three times for each model and the median result used to account for variability.

This work was reported following the TRIPOD+AI checklist [200] to ensure thorough reporting, with the completed checklist available in Appendix C. Experiments were conducted using the HPC (Section 4.5), and the PyTorch-based code was made available at https://github.com/scjjb/Ovarian_Features.

7.3 Results

7.3.1 Foundation Model Validation Results

No single model gave the greatest results in every validation (Figure 7.2). Virchow2-CLS gave the greatest performance in cross-validation (Table 7.3), H-optimus-0 in hold-out testing (Table 7.4), GPFM in the Transcanadian Study external validation (Table 7.5), and Virchow in the OCEAN Challenge external validation (Table 7.6). RN18-Histo had the worst performance of any foundation model in all validations and was the only foundation model to perform worse than any ImageNet-pretrained encoder overall (Table 7.7).

The H-optimus-0 model achieved the greatest averaged performance across all validations (Table 7.7), with 83.0% average balanced accuracy, 0.965 average AUROC, and 0.822 average F1 score. This performance very was closely followed by that of UNI and Virchow2-CLS. The worst averaged performances were given by CNN-based feature extraction models (RN50, RN18, RN18-Histo), followed by the ImageNet-pretrained vision transformer. Confusion matrices for the optimal H-optimus-0 model (Figure 7.3) show that no single class was the best (or worst) classified across all validations. The worst F1 scores were found for the classification of LGSC in cross-validation (0.443) and the OCEAN Challenge validation (0.582), and for EC in the OCEAN Challenge validation (0.606). In these validations, LGSC was often confused with HGSC and there was a moderate level of confusion between EC and MC. Further class-level results are provided in Table 7.8.

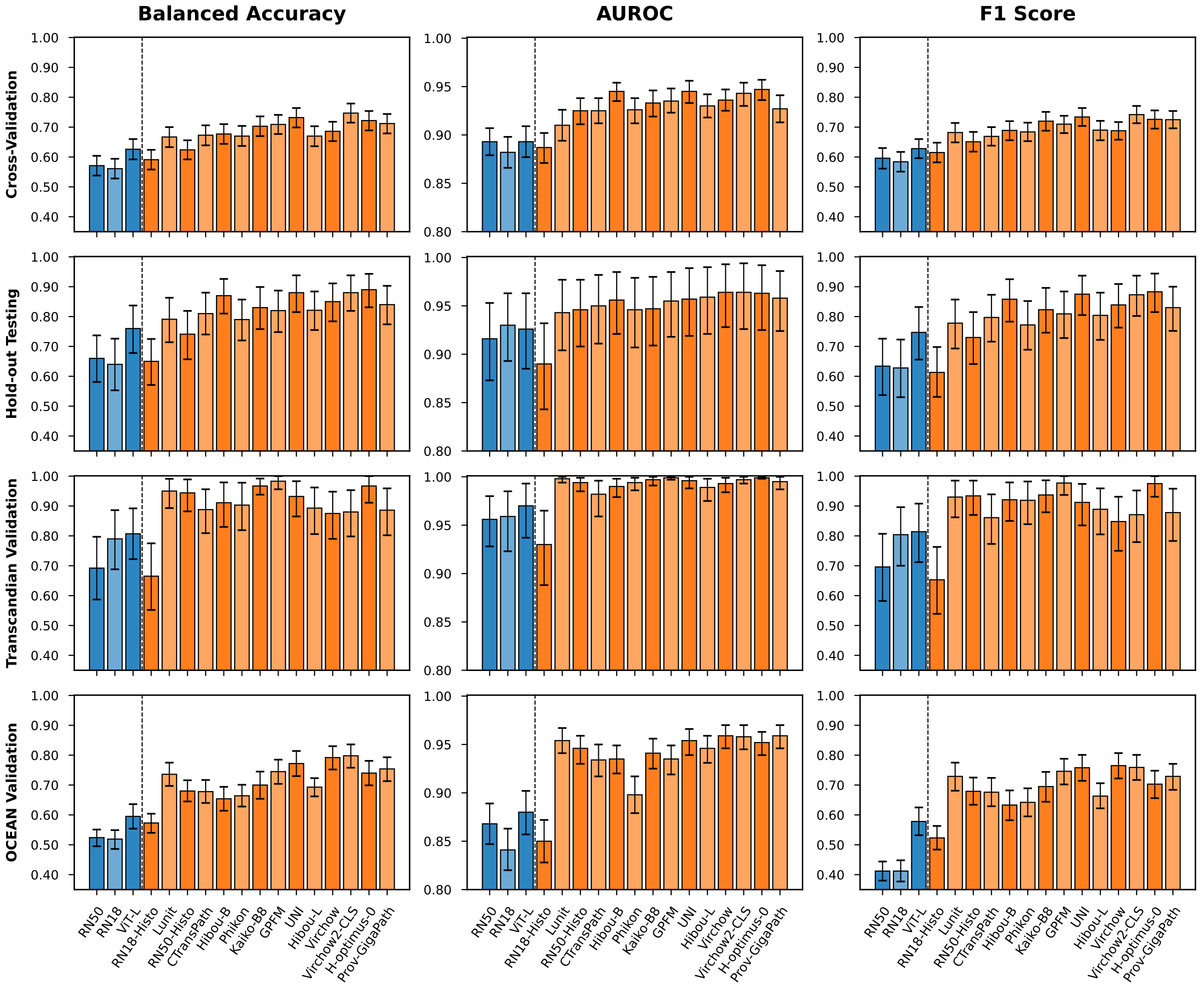


Figure 7.2 Ovarian cancer subtyping results for each feature extractor and validation (mean and 95% confidence interval generated by 10,000 iterations of bootstrapping). Blue indicates ImageNet-pretrained feature extractors, orange indicates histopathology foundation models. Hold-out testing and external validation results are based on an ensemble of five cross-validation models.

Confusion matrices for the optimal H-optimus-0 model (Figure 7.3) show that the model did not completely fail at classifying any one subtype (as in Chapter 6), though there was still variability in class-wise performance, especially in validations that included IDS samples (cross-validation and the OCEAN Challenge). The least common class in the training set (LGSC) was poorly classified in these validations (F1 scores of 0.443 and 0.582) but was much better classified in the other validations (F1 scores of 0.865 and 0.941). The most consistently classified subtype was the most common in the training dataset (HGSC), with F1 scores of at least 0.807 in all validations.

Feature Extractor	Balanced Accuracy	AUROC	F1 Score
RN50	57.1% (53.8-60.4%)	0.893 (0.879-0.907)	0.596 (0.561-0.630)
RN18	56.1% (52.8-59.4%)	0.882 (0.866-0.898)	0.584 (0.551-0.617)
ViT-L	62.6% (59.2-66.0%)	0.893 (0.877-0.909)	0.628 (0.596-0.660)
RN18-Histo	59.1% (55.8-62.4%)	0.887 (0.871-0.902)	0.615 (0.582-0.648)
Lunit	66.6% (63.3-70.0%)	0.910 (0.894-0.926)	0.682 (0.649-0.714)
RN50-Histo	62.4% (59.2-65.6%)	0.925 (0.911-0.938)	0.651 (0.618-0.684)
CTransPath	67.3% (63.9-70.6%)	0.925 (0.912-0.938)	0.669 (0.638-0.700)
Hibou-B	67.7% (64.4-71.0%)	0.945 (0.935-0.954)	0.689 (0.656-0.720)
Phikon	67.0% (63.7-70.4%)	0.926 (0.912-0.938)	0.684 (0.653-0.715)
Kaiko-B8	70.3% (67.0-73.6%)	0.933 (0.919-0.946)	0.720 (0.688-0.751)
GPFM	70.9% (67.7-74.1%)	0.935 (0.923-0.948)	0.710 (0.680-0.739)
UNI	73.2% (69.9-76.4%)	0.945 (0.933-0.956)	0.734 (0.704-0.764)
Hibou-L	67.0% (63.6-70.3%)	0.930 (0.918-0.942)	0.690 (0.656-0.721)
Virchow	68.6% (65.3-71.8%)	0.936 (0.925-0.947)	0.688 (0.658-0.717)
Virchow2-CLS	**74.7%** (71.5-77.9%)	0.943 (0.930-0.954)	**0.742** (0.713-0.771)
H-optimus-0	72.2% (68.9-75.4%)	**0.947** (0.936-0.957)	0.726 (0.695-0.756)
Prov-GigaPath	71.2% (67.9-74.4%)	0.927 (0.913-0.941)	0.725 (0.696-0.754)

Table 7.3 Results of five-fold cross-validation. Results are reported as the mean and 95% confidence intervals (in brackets) from 10,000 iterations of bootstrapping. The greatest results are shown in **bold**.

Feature Extractor	**Balanced Accuracy**	**AUROC**	**F1 Score**
RN50	66.0% (58.1-73.7%)	0.916 (0.873-0.953)	0.634 (0.537-0.726)
RN18	64.0% (55.3-72.6%)	0.930 (0.893-0.963)	0.628 (0.530-0.723)
ViT-L	76.0% (67.8-83.7%)	0.926 (0.885-0.963)	0.747 (0.656-0.832)
RN18-Histo	65.0% (57.1-72.5%)	0.890 (0.843-0.932)	0.613 (0.531-0.698)
Lunit	79.1% (71.4-86.3%)	0.943 (0.904-0.977)	0.778 (0.693-0.857)
RN50-Histo	74.1% (65.7-81.9%)	0.946 (0.908-0.977)	0.730 (0.641-0.815)
CTransPath	81.0% (74.0-88.0%)	0.950 (0.911-0.982)	0.797 (0.716-0.873)
Hibou-B	87.0% (81.0-92.6%)	0.956 (0.921-0.985)	0.858 (0.783-0.925)
Phikon	79.0% (72.0-85.7%)	0.946 (0.907-0.979)	0.772 (0.689-0.852)
Kaiko-B8	83.0% (75.8-89.9%)	0.947 (0.909-0.980)	0.823 (0.746-0.896)
GPFM	82.0% (74.8-88.7%)	0.955 (0.918-0.985)	0.809 (0.728-0.884)
UNI	88.0% (81.5-93.8%)	0.957 (0.919-0.989)	0.875 (0.805-0.937)
Hibou-L	82.1% (75.5-88.4%)	0.959 (0.921-0.990)	0.804 (0.722-0.880)
Virchow	85.0% (78.4-91.1%)	**0.964** (0.928-0.993)	0.839 (0.763-0.909)
Virchow2-CLS	88.0% (81.9-93.8%)	**0.964** (0.926-0.994)	0.873 (0.802-0.937)
H-optimus-0	**89.0%** (83.1-94.3%)	0.963 (0.925-0.992)	**0.883** (0.815-0.944)
Prov-GigaPath	84.0% (77.4-90.3%)	0.958 (0.924-0.986)	0.830 (0.752-0.900)

Table 7.4 Results of hold-out testing, with predictions generated by an ensemble of the five-fold classification models. Results are reported as the mean and 95% confidence intervals (in brackets) from 10,000 iterations of bootstrapping. The greatest results are shown in **bold**.

Feature Extractor	Balanced Accuracy	AUROC	F1 Score
RN50	69.2% (58.7-79.7%)	0.956 (0.928-0.980)	0.696 (0.582-0.807)
RN18	79.0% (68.8-88.6%)	0.959 (0.923-0.985)	0.804 (0.700-0.896)
ViT-L	80.7% (72.2-89.2%)	0.970 (0.937-0.993)	0.814 (0.712-0.908)
RN18-Histo	66.5% (55.2-77.5%)	0.930 (0.888-0.965)	0.653 (0.539-0.763)
Lunit	95.0% (89.3-99.1%)	0.998 (0.994-1.000)	0.930 (0.862-0.985)
RN50-Histo	94.4% (88.2-98.9%)	0.994 (0.985-0.999)	0.934 (0.870-0.985)
CTransPath	88.8% (80.9-95.6%)	0.982 (0.959-0.996)	0.861 (0.773-0.939)
Hibou-B	91.1% (83.0-97.9%)	0.990 (0.979-0.998)	0.921 (0.850-0.979)
Phikon	90.3% (81.9-97.8%)	0.994 (0.986-0.999)	0.919 (0.839-0.982)
Kaiko-B8	96.7% (93.8-99.2%)	0.997 (0.991-1.000)	0.937 (0.879-0.986)
GPFM	**98.3%** (95.6-100.0%)	**0.999** (0.997-1.000)	**0.977** (0.937-1.000)
UNI	93.2% (86.5-98.3%)	0.996 (0.988-1.000)	0.912 (0.835-0.974)
Hibou-L	89.3% (80.6-96.2%)	0.989 (0.975-0.998)	0.889 (0.805-0.959)
Virchow	87.5% (79.0-94.8%)	0.993 (0.984-0.999)	0.848 (0.750-0.931)
Virchow2-CLS	88.0% (79.8-95.3%)	0.997 (0.993-1.000)	0.871 (0.779-0.952)
H-optimus-0	96.7% (91.1-100.0%)	**0.999** (0.998-1.000)	0.975 (0.931-1.000)
Prov-GigaPath	88.6% (80.2-95.9%)	0.995 (0.987-1.000)	0.878 (0.783-0.958)

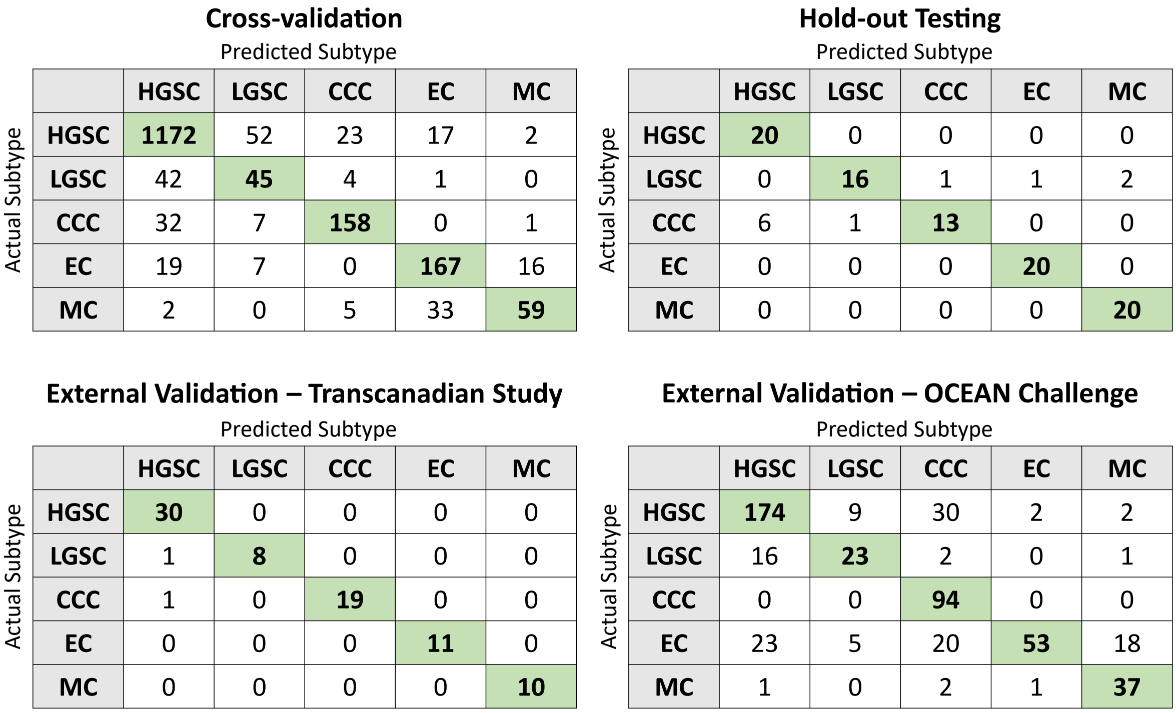
Table 7.5 Results of external validation on the Transcanadian Study dataset, with predictions generated by an ensemble of the five-fold classification models. Results are reported as the mean and 95% confidence intervals (in brackets) from 10,000 iterations of bootstrapping. The greatest results are shown in **bold**.

Feature Extractor	**Balanced Accuracy**	**AUROC**	**F1 Score**
RN50	52.4% (49.5-55.1%)	0.868 (0.847-0.889)	0.412 (0.380-0.444)
RN18	51.9% (48.6-54.9%)	0.841 (0.820-0.863)	0.412 (0.377-0.448)
ViT-L	59.5% (55.4-63.6%)	0.880 (0.857-0.902)	0.578 (0.532-0.625)
RN18-Histo	57.3% (54.0-60.4%)	0.850 (0.828-0.872)	0.523 (0.484-0.563)
Lunit	73.6% (69.7-77.5%)	0.954 (0.941-0.967)	0.729 (0.681-0.775)
RN50-Histo	68.0% (64.5-71.6%)	0.946 (0.930-0.959)	0.679 (0.634-0.725)
CTransPath	67.8% (64.0-71.7%)	0.934 (0.917-0.950)	0.676 (0.629-0.724)
Hibou-B	65.4% (61.4-69.4%)	0.935 (0.920-0.949)	0.633 (0.582-0.682)
Phikon	66.4% (62.8-70.1%)	0.898 (0.879-0.917)	0.642 (0.595-0.689)
Kaiko-B8	70.0% (65.4-74.5%)	0.941 (0.925-0.956)	0.695 (0.644-0.744)
GPFM	74.5% (70.4-78.5%)	0.935 (0.919-0.949)	0.746 (0.702-0.788)
UNI	77.2% (73.0-81.4%)	0.954 (0.939-0.966)	0.758 (0.714-0.801)
Hibou-L	69.3% (66.2-72.3%)	0.946 (0.931-0.959)	0.663 (0.622-0.706)
Virchow	79.2% (75.2-83.0%)	**0.959** (0.946-0.970)	**0.765** (0.722-0.807)
Virchow2-CLS	**79.8%** (75.8-83.6%)	0.958 (0.945-0.970)	0.759 (0.717-0.801)
H-optimus-0	74.0% (69.9-78.1%)	0.952 (0.939-0.963)	0.703 (0.656-0.748)
Prov-GigaPath	75.4% (71.3-79.3%)	**0.959** (0.946-0.970)	0.729 (0.684-0.771)

Table 7.6 Results of external validation on the OCEAN dataset, with predictions generated by an ensemble of the five-fold classification models. Results are reported as the mean and 95% confidence intervals (in brackets) from 10,000 iterations of bootstrapping. The greatest results are shown in **bold**.

	Feature Extractor	**Balanced Accuracy**	**AUROC**	**F1 Score**	**Avg Inference Time (s/WSI)**
ImageNet-Pretrained Models	RN50	61.2%	0.908	0.585	75.6
	RN18	62.8%	0.903	0.607	75.4
	ViT-L	69.7%	0.917	0.692	99.3
Histopathology Foundation Models	RN18-Histo	62.0%	0.889	0.601	76.1
	Lunit	78.6%	0.951	0.780	76.4
	RN50-Histo	74.7%	0.953	0.749	75.1
	CTransPath	76.2%	0.948	0.751	75.7
	Hibou-B	77.9%	0.957	0.775	76.9
	Phikon	75.7%	0.941	0.754	76.9
	Kaiko-B8	80.0%	0.955	0.794	129.0
	GPFM	81.4%	0.956	0.811	125.1
	UNI	82.9%	0.963	0.820	99.9
	Hibou-L	76.9%	0.956	0.762	130.4
	Virchow	80.1%	0.963	0.785	243.1
	Virchow2-CLS	82.6%	**0.966**	0.811	245.8
	H-optimus-0	**83.0%**	0.965	**0.822**	425.0
	Prov-GigaPath	79.8%	0.960	0.791	319.8

Table 7.7 Averaged results across the four validations. The average inference times were measured on a subset of the internal hold-out test set. The greatest result for each metric is shown in **bold**.



Cross-validation

Predicted Subtype

Actual Subtype

	HGSC	LGSC	CCC	EC	MC
HGSC	**1172**	52	23	17	2
LGSC	42	**45**	4	1	0
CCC	32	7	**158**	0	1
EC	19	7	0	**167**	16
MC	2	0	5	33	**59**

Hold-out Testing

Predicted Subtype

Actual Subtype

	HGSC	LGSC	CCC	EC	MC
HGSC	**20**	0	0	0	0
LGSC	0	**16**	1	1	2
CCC	6	1	**13**	0	0
EC	0	0	0	**20**	0
MC	0	0	0	0	**20**

External Validation – Transcanadian Study

Predicted Subtype

Actual Subtype

	HGSC	LGSC	CCC	EC	MC
HGSC	**30**	0	0	0	0
LGSC	1	**8**	0	0	0
CCC	1	0	**19**	0	0
EC	0	0	0	**11**	0
MC	0	0	0	0	**10**

External Validation – OCEAN Challenge

Predicted Subtype

Actual Subtype

	HGSC	LGSC	CCC	EC	MC
HGSC	**174**	9	30	2	2
LGSC	16	**23**	2	0	1
CCC	0	0	**94**	0	0
EC	23	5	20	**53**	18
MC	1	0	2	1	**37**

Figure 7.3 Confusion matrices for the optimal ABMIL classifier with features from the H-optimus-0 foundation model. Correct classifications are indicated in **bold**.

The difference in performance between each foundation model (except RN18-Histo) and the baseline ImageNet-pretrained ResNet50 was found to be significant by all metrics in all validations (Table 7.9), except the AUROC in cross-validation (for nine foundation models), RN50-Histo in internal validations, and Hibou-B in the external validation on the OCEAN Challenge dataset. There was no significant difference between the performance of the baseline model and either the RN18 or the RN18-Histo model in most validations. The difference between the baseline ResNet50 and the ViT-L feature extractor was statistically significant in most validations for the balanced accuracy and F1 score, but not the AUROC.

There was a strong positive relationship ($R^2 = 0.93$) between the size of feature extraction models and the runtime (Figure 7.4). The most computationally efficient models were typically the smallest, with an average inference time per WSI between 75 and 77 seconds for each of the ResNets, Lunit, CTransPath, Hibou-B, and Phikon models (Table 7.7). Feature encoding was the slowest step of slide inference, taking over 90% of the total computational time for all models, with the remaining time divided between the initial tissue patch extraction and the subsequent forward pass of patch features through the trained ABMIL classifiers. The average inference times did not vary greatly for any model over the three repeats, with a maximum range of 1.7s (75.3 - 77.0s) per WSI from the CTransPath model. The largest models were the slowest overall, with Prov-GigaPath averaging 320 seconds and H-optimus-0 averaging 425 seconds per WSI, over 5 times as long as the fastest models. These largest feature extractors also required much greater computational resources (particularly VRAM) as they were each over 4GB in size, whereas the smallest models were each under 100MB (RN50, RN18, RN18-Histo, Lunit, RN50-Histo).

	Subtype	**F1 Score**	**Precision**	**Recall / Sensitivity**	**Specificity**	**Balanced Accuracy**
Cross-Validation	HGSC	0.925	0.925	0.926	0.841	0.883
	LGSC	0.443	0.405	0.489	0.963	0.726
	CCC	0.814	0.832	0.798	0.981	0.889
	EC	0.782	0.766	0.799	0.969	0.884
	MC	0.667	0.756	0.596	0.989	0.793
Hold-out Testing	HGSC	0.870	0.769	1.000	0.925	0.963
	LGSC	0.865	0.941	0.800	0.988	0.894
	CCC	0.765	0.929	0.650	0.988	0.819
	EC	0.976	0.952	1.000	0.988	0.994
	MC	0.952	0.909	1.000	0.975	0.988
Transcanadian Study	HGSC	0.968	0.938	1.000	0.960	0.980
	LGSC	0.941	1.000	0.889	1.000	0.944
	CCC	0.974	1.000	0.950	1.000	0.975
	EC	1.000	1.000	1.000	1.000	1.000
	MC	1.000	1.000	1.000	1.000	1.000
OCEAN Challenge	HGSC	0.807	0.813	0.802	0.865	0.833
	LGSC	0.582	0.622	0.548	0.970	0.759
	CCC	0.777	0.635	1.000	0.871	0.936
	EC	0.606	0.946	0.445	0.992	0.719
	MC	0.747	0.638	0.902	0.956	0.929

Table 7.8 Additional classwise classification metrics for the optimal H-optimus-0 ABMIL classifier.

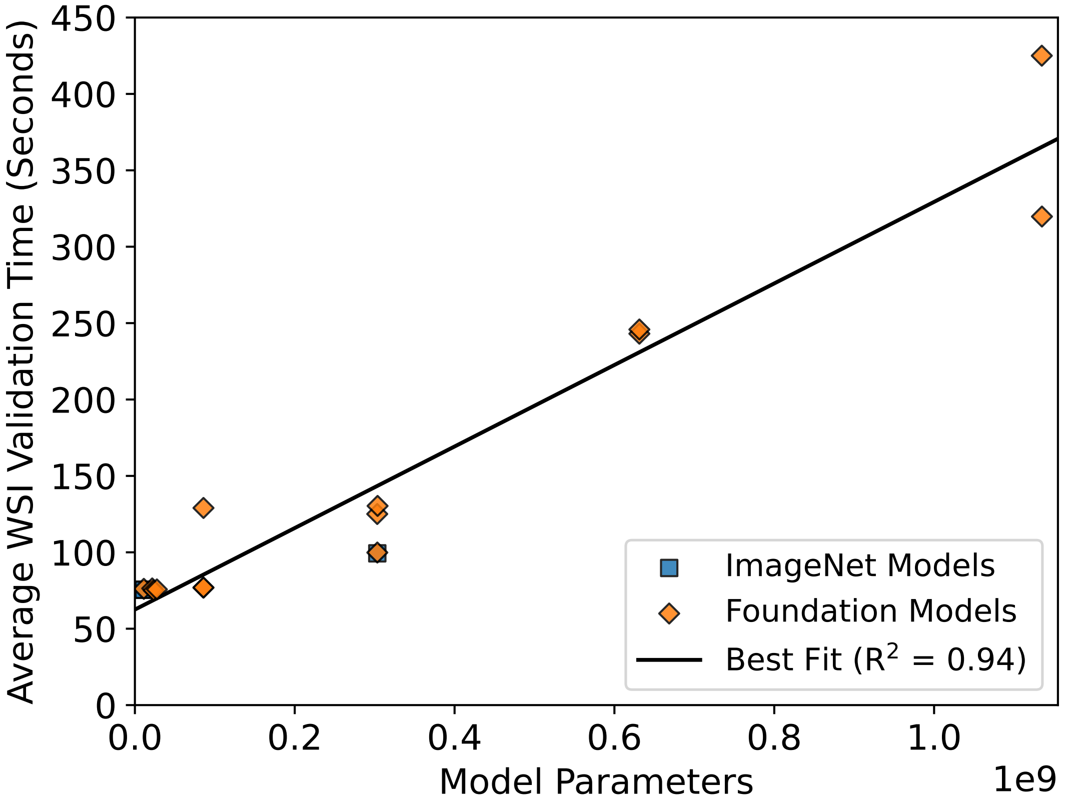


Figure 7.4 Model inference times. The average inference time per WSI for each model, including tissue patch extraction, feature encoding, and ABMIL classification time.

	Cross-Validation p-values			**Hold-out Testing p-values**		
Model	**Balanced Accuracy**	**AUROC**	**F1 Score**	**Balanced Accuracy**	**AUROC**	**F1 Score**
RN18	0.736	0.557	0.736	0.601	**0.028**	0.317
ViT-L	**0.033**	0.824	0.074	**0.003**	0.051	**0.002**
RN18-Histo	0.365	0.964	0.613	0.870	0.192	0.967
Lunit	**0.019**	0.183	**0.009**	**0.010**	**0.035**	**0.008**
RN50-Histo	0.232	0.072	0.152	0.214	**0.009**	0.189
CTransPath	**0.009**	0.082	**0.007**	**0.003**	**0.012**	**0.003**
Hibou-B	**0.019**	**0.029**	**0.012**	**0.003**	**0.006**	**0.003**
Phikon	**0.009**	0.149	**0.007**	**0.003**	**0.012**	**0.003**
Kaiko-B8	**0.013**	0.063	**0.010**	**0.003**	**0.011**	**0.002**
GPFM	**0.007**	0.063	**0.007**	**0.003**	**0.006**	**0.003**
UNI	**0.015**	**0.020**	**0.009**	**0.003**	**0.006**	**0.002**
Hibou-L	**0.007**	0.072	**0.007**	**0.003**	**0.009**	**0.003**
Virchow	**0.011**	**0.020**	**0.006**	**0.003**	**0.006**	**0.003**
Virchow2-CLS	**0.011**	0.063	**0.009**	**0.002**	**0.006**	**0.002**
H-Optimus-0	**0.005**	**0.020**	**0.001**	**0.003**	**0.006**	**0.002**
Prov-GigaPath	**0.013**	0.063	**0.007**	**0.008**	**0.006**	**0.008**

	Transcanadian Study p-values			**OCEAN Challenge p-values**		
Model	**Balanced Accuracy**	**AUROC**	**F1 Score**	**Balanced Accuracy**	**AUROC**	**F1 Score**
RN18	0.446	0.773	0.399	0.237	**0.034**	0.541
ViT-L	**0.021**	0.090	**0.022**	0.170	0.265	**0.019**
RN18-Histo	0.490	0.211	0.403	0.235	0.987	**0.002**
Lunit	**0.003**	**0.011**	**0.006**	**0.002**	**0.004**	**0.001**
RN50-Histo	**0.015**	**0.011**	**0.018**	**0.018**	**0.004**	**0.003**
CTransPath	**0.007**	**0.023**	**0.018**	**0.002**	**0.005**	**0.001**
Hibou-B	**0.008**	**0.024**	**0.007**	0.107	**0.009**	**0.019**
Phikon	**0.002**	**0.011**	$<$**0.001**	**0.001**	**0.013**	**0.001**
Kaiko-B8	**0.007**	**0.011**	**0.022**	**0.004**	**0.006**	**0.003**
GPFM	**0.003**	**0.011**	**0.006**	**0.001**	**0.004**	**0.001**
UNI	**0.006**	**0.011**	**0.015**	**0.001**	**0.004**	**0.001**
Hibou-L	**0.003**	**0.013**	**0.003**	**0.002**	**0.005**	**0.002**
Virchow	**0.006**	**0.018**	**0.015**	**0.002**	**0.004**	**0.001**
Virchow2-CLS	**0.007**	**0.011**	**0.015**	**0.001**	**0.004**	**0.001**
H-Optimus-0	**0.003**	**0.011**	**0.004**	**0.002**	**0.004**	**0.001**
Prov-GigaPath	**0.019**	**0.017**	**0.035**	**0.005**	**0.004**	**0.002**

Table 7.9 Resulting p-values from paired *t*-tests comparing the subtype classification results with each feature extractor to the ImageNet-pretrained ResNet50 baseline. False discovery rate p-value adjustments were applied to account for multiple testing [206]. Values below 0.05 are indicated in **bold**.

7.3.2 Normalisation and Augmentation Results

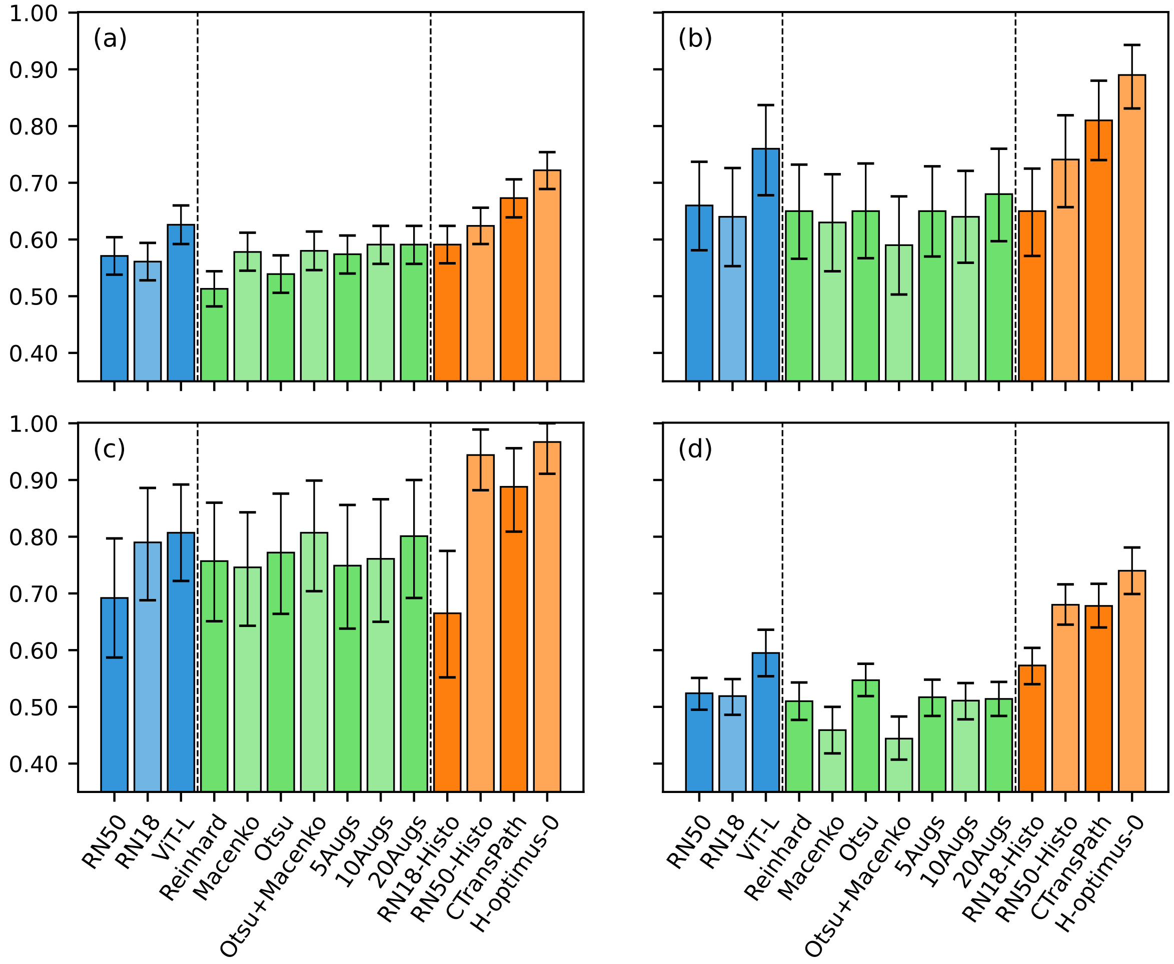


Figure 7.5 Balanced accuracy (mean and 95% confidence interval from 10,000 iterations of bootstrapping) for each standard ImageNet-pretrained feature extractor (blue), the seven ResNet50 models with varied preprocessing techniques (green), as well as the three worst-performing (RN18-Histo, RN50-Histo, and CTransPath) and the single best-performing foundation models (H-optimus-0) in (a) cross-validation, (b) hold-out testing, (c) external validation on the Transcanadian Study dataset, (d) external validation on the OCEAN Challenge dataset. For validations (b)-(d), predictions were ensembled from the five cross-validation models.

Different preprocessing techniques had inconsistent effects on the ImageNet-pretrained ResNet50 feature extractor (Figure 7.5), with some modest benefits in internal validations, and variable effects in external validations. In cross-validation (Table 7.10), no pre-processing method improved the balanced accuracy or F1 score by more than 0.02, and no improvement was seen in AUROC with any method. In hold-out testing (Table 7.11), only the 20x augmentation improved performance, increasing F1 by 0.023 and balanced accuracy by 0.020, but reducing AUROC by 0.012. However, in the external validation on the Transcanadian Study dataset (Table 7.12), every preprocessing method improved performance compared to the baseline by over 0.05 balanced accuracy and F1 score, and 0.002 AUROC. The greatest performances in this validation were found by combining Otsu thresholding with Macenko normalisation and by 20x colour augmentations, which each increased the F1 score and balanced accuracy above baseline performance by over 0.1, and AUROC by over 0.016. For the OCEAN Challenge dataset (Table 7.13), most preprocessing methods gave worse results than the baseline approach, with only Otsu thresholding providing any benefit over the baseline performance.

Preprocessing Approach	Balanced Accuracy	AUROC	F1 Score
Baseline	57.1% (53.8-60.4%)	**0.893** (0.879-0.907)	0.596 (0.561-0.630)
Reinhard Normalisation	51.3% (48.2-54.4%)	0.872 (0.856-0.887)	0.520 (0.488-0.553)
Macenko Normalisation	57.8% (54.5-61.2%)	0.882 (0.867-0.896)	0.601 (0.567-0.635)
Otsu Thresholding	53.9% (50.6-57.2%)	0.888 (0.873-0.903)	0.566 (0.532-0.600)
Otsu + Macenko	58.0% (54.6-61.4%)	0.882 (0.865-0.898)	0.605 (0.571-0.638)
5x Colour Augmentation	57.4% (54.0-60.7%)	0.888 (0.873-0.902)	0.592 (0.560-0.625)
10x Colour Augmentation	**59.1%** (55.7-62.4%)	0.891 (0.877-0.905)	**0.615** (0.581-0.649)
20x Colour Augmentation	**59.1%** (55.7-62.4%)	0.892 (0.877-0.905)	0.596 (0.564-0.627)

Table 7.10 Results of five-fold cross-validation for the ImageNet-pretrained ResNet50 with varied preprocessing approaches. Results are reported as the mean and 95% confidence intervals (in brackets) from 10,000 iterations of bootstrapping. The greatest results are shown in **bold**.

Preprocessing Approach	Balanced Accuracy	AUROC	F1 Score
Baseline	66.0% (58.1-73.7%)	0.916 (0.873-0.953)	0.634 (0.537-0.726)
Reinhard Normalisation	65.0% (56.6-73.2%)	**0.923** (0.881-0.961)	0.632 (0.534-0.727)
Macenko Normalisation	63.0% (54.4-71.5%)	0.915 (0.873-0.951)	0.620 (0.521-0.715)
Otsu Thresholding	65.0% (56.7-73.4%)	0.916 (0.872-0.955)	0.637 (0.542-0.732)
Otsu + Macenko	59.0% (50.3-67.6%)	0.918 (0.878-0.952)	0.577 (0.475-0.674)
5x Colour Augmentation	65.0% (57.0-72.9%)	0.916 (0.876-0.951)	0.630 (0.536-0.725)
10x Colour Augmentation	64.0% (55.9-72.1%)	0.906 (0.864-0.944)	0.616 (0.522-0.710)
20x Colour Augmentation	**68.0%** (59.7-76.0%)	0.904 (0.861-0.942)	**0.657** (0.563-0.750)

Table 7.11 Results of hold-out testing for the ImageNet-pretrained ResNet50 with varied preprocessing approaches, with predictions generated by an ensemble of the five-fold classification models. Results are reported as the mean and 95% confidence intervals (in brackets) from 10,000 iterations of bootstrapping. The greatest results are shown in **bold**.

Preprocessing Approach	**Balanced Accuracy**	**AUROC**	**F1 Score**
Baseline	69.2% (58.7-79.7%)	0.956 (0.928-0.980)	0.696 (0.582-0.807)
Reinhard Normalisation	75.8% (65.1-86.0%)	0.968 (0.943-0.986)	0.761 (0.647-0.861)
Macenko Normalisation	74.5% (64.3-84.3%)	0.959 (0.933-0.980)	0.756 (0.648-0.857)
Otsu Thresholding	77.2% (66.4-87.6%)	0.963 (0.937-0.985)	0.797 (0.685-0.895)
Otsu + Macenko	**80.5%** (70.4-89.9%)	**0.983** (0.967-0.995)	**0.834** (0.730-0.921)
5x Colour Augmentation	74.9% (63.8-85.6%)	0.966 (0.941-0.986)	0.762 (0.647-0.866)
10x Colour Augmentation	76.1% (65.0-86.6%)	0.962 (0.935-0.983)	0.768 (0.659-0.869)
20x Colour Augmentation	80.0% (69.2-90.0%)	0.973 (0.953-0.989)	0.806 (0.706-0.897)

Table 7.12 Results of external validation on the Transcanadian Study dataset for the ImageNet-pretrained ResNet50 with varied preprocessing approaches, with predictions generated by an ensemble of the five-fold classification models. Results are reported as the mean and 95% confidence intervals (in brackets) from 10,000 iterations of bootstrapping. The greatest results are shown in **bold**.

Preprocessing Approach	**Balanced Accuracy**	**AUROC**	**F1 Score**
Baseline	52.4% (49.5-55.1%)	0.868 (0.847-0.889)	0.412 (0.380-0.444)
Reinhard Normalisation	51.0% (47.7-54.3%)	0.870 (0.850-0.888)	0.392 (0.350-0.437)
Macenko Normalisation	45.9% (41.8-50.0%)	0.837 (0.814-0.860)	0.407 (0.360-0.455)
Otsu Thresholding	**54.7%** (51.9-57.6%)	**0.883** (0.864-0.901)	**0.440** (0.401-0.482)
Otsu + Macenko	44.4% (40.7-48.3%)	0.840 (0.816-0.862)	0.388 (0.347-0.432)
5x Colour Augmentation	51.7% (48.4-54.8%)	0.867 (0.845-0.887)	0.401 (0.363-0.441)
10x Colour Augmentation	51.1% (47.8-54.2%)	0.877 (0.856-0.897)	0.404 (0.367-0.443)
20x Colour Augmentation	51.4% (48.4-54.4%)	0.874 (0.853-0.893)	0.391 (0.352-0.433)

Table 7.13 Results of external validation on the OCEAN Challenge dataset for the ImageNet-pretrained ResNet50 with varied preprocessing approaches, with predictions generated by an ensemble of the five-fold classification models. Results are reported as the mean and 95% confidence intervals (in brackets) from 10,000 iterations of bootstrapping. The greatest results are shown in **bold**.

Despite some modest improvements offered by different preprocessing techniques, particularly in the Transcanadian Study external validation, the best-performing model based on the ImageNet-pretrained ResNet50 backbone was still outperformed by every foundation model (except RN18-Histo) in every validation (Figure 7.5). Furthermore, none of the different preprocessing methods gave statistically significant differences in performance compared to the baseline approach in any validation.

7.3.3 Hyperparameter Tuning Ablation Results

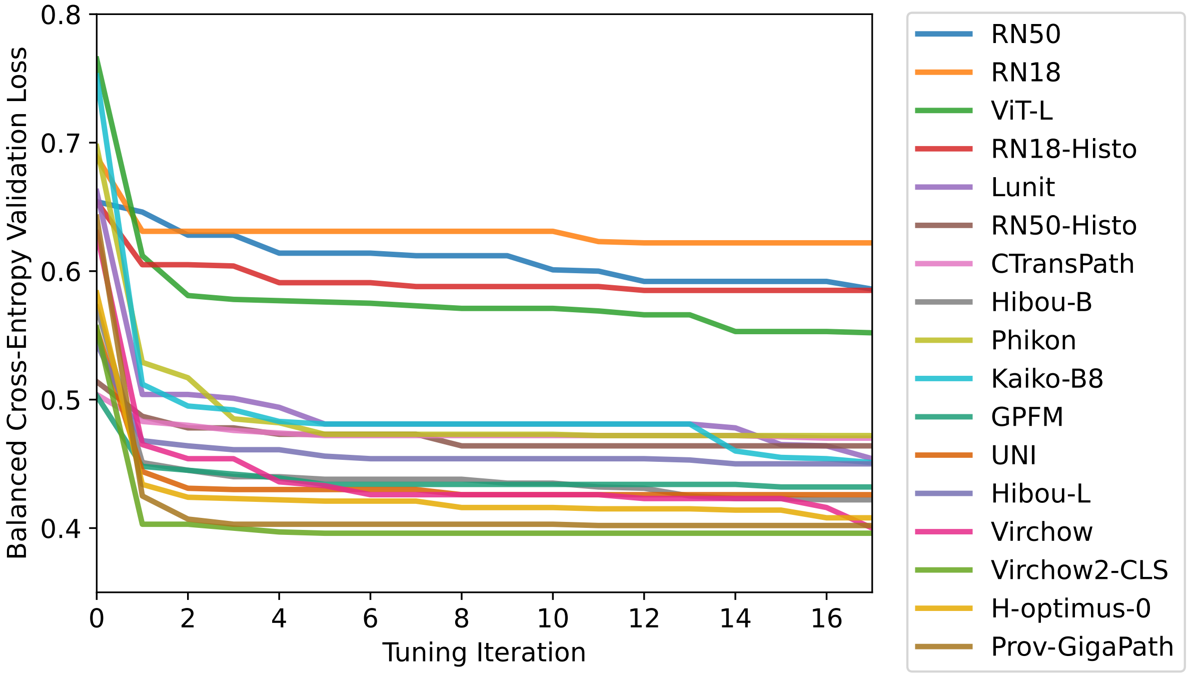


Figure 7.6 Average validation loss from five-fold cross-validation for each model and each hyperparameter tuning iteration.

Hyperparameter tuning improved the average validation loss for every model by at least 0.034 (CTransPath from 0.504 to 0.470), with a median improvement of 0.150, and a maximum of 0.301 (Kaiko-B8 from 0.752 to 0.451). As shown in Figure 7.6, the majority of this benefit was found within the first iteration of hyperparameter tuning for every model (except the ImageNet-pretrained ResNet50), with a median improvement of 0.121 validation loss from tuning only the learning rate and ABMIL model size.

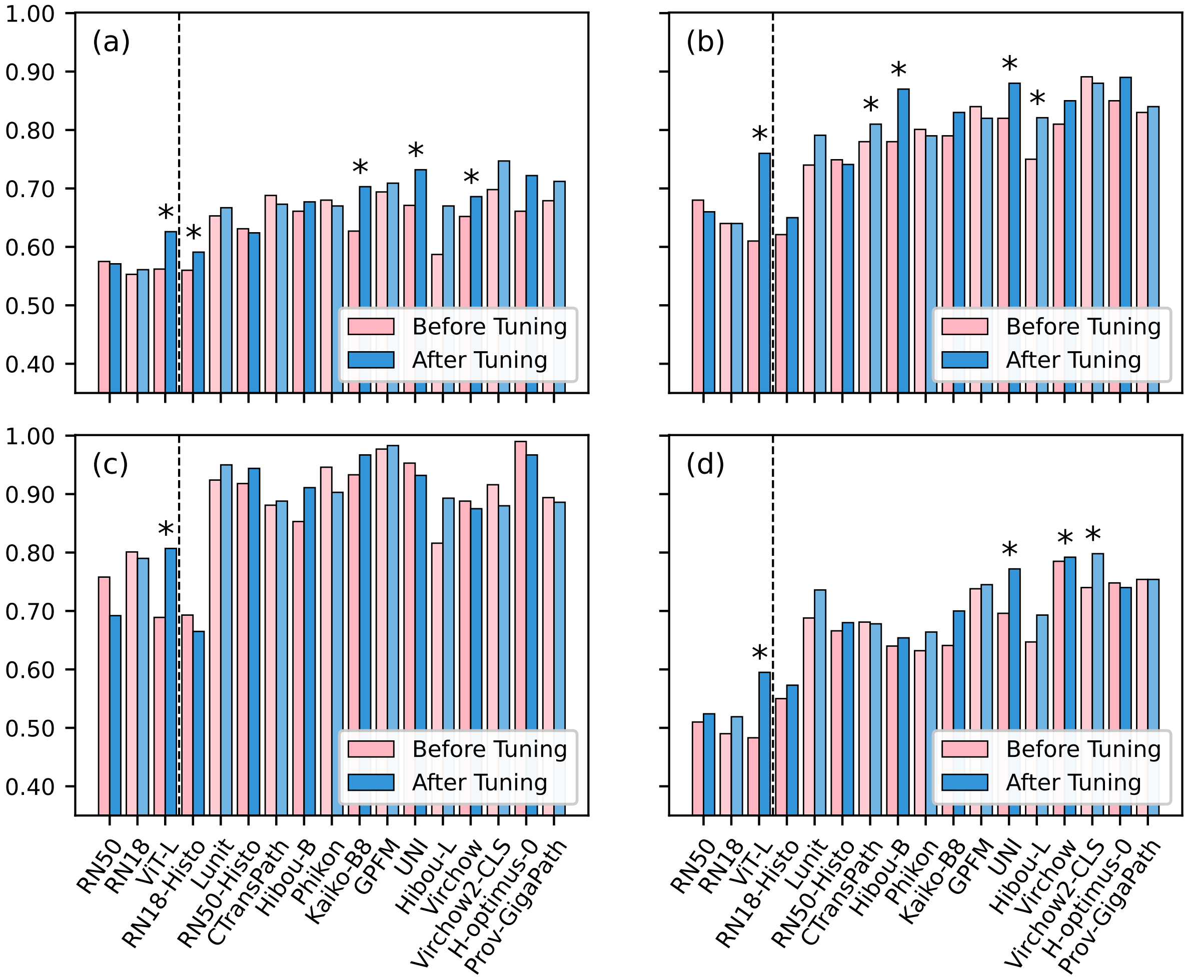


Figure 7.7 Balanced accuracy results for each model compared with the ABMIL classifier trained using the default hyperparameters (pink) and the tuned hyperparameters (blue) for (a) cross-validation, (b) hold-out testing, (c) external validation on the Transcanadian Study dataset, (d) external validation on the OCEAN Challenge dataset. For validations (b)-(d), predictions were ensembled from the five cross-validation models. *Indicates a significant difference in the paired *t*-test at the 5% significance level.

The balanced accuracies of the tuned ABMIL classifiers are compared to the untuned models (using default hyperparameters) in Figure 7.7. The median impact of hyperparameter tuning across all models and validations was an improvement of 1.9% balanced accuracy, 0.005 AUROC, and 0.025 F1 score, though the effect on any specific model in any given validation was variable, with balanced accuracies changed by -6.6% to $+15.0\%$, AUROCs by -0.013 to $+0.041$, and F1 scores by -0.073 to $+0.146$. The only models which did not benefit from hyperparameter tuning were those using the ResNet50, ResNet18, Phikon, and H-optimus-0 feature extractors. All of the other models had a statistically significant difference between tuned and untuned results in at least one validation (Tables 7.14 and 7.15), with these significant differences only occurring in cases where tuning improved performance. The extent of the benefits varied across validations, with a median change in balanced accuracy of $+3.1\%$ in cross-validation, $+3.0\%$ in hold-out testing, -0.8% in the Transcanadian Study external validation, and $+1.9\%$ in the OCEAN Challenge external validation. The only models to significantly benefit in every validation were the ImageNet-pretrained ViT-L and Hibou-L, though these benefits were not present for every metric. Exact p-values are provided in (Appendix D)

	Cross-Validation						Hold-out Testing					
Feature Extractor	Balanced Accuracy		AUROC		F1 Score		Balanced Accuracy		AUROC		F1 Score	
RN50	57.5%	-	0.877	↓	0.593	-	68.0%	↑	0.923	-	0.670	↑↑
RN18	55.3%	-	0.857	↓	0.561	↓	64.0%	-	0.927	-	0.626	-
ViT-L	56.2%	↓↓↓*	0.857	↓↓*	0.580	↓↓*	61.0%	↓↓↓*	0.917	-	0.601	↓↓↓*
RN18-Histo	56.0%	↓↓*	0.879	-	0.574	↓↓*	62.1%	↓	0.889	-	0.586	↓
Lunit	65.3%	↓	0.891	↓*	0.646	↓↓	74.0%	↓↓↓	0.932	↓	0.727	↓↓↓
RN50-Histo	63.1%	-	0.915	↓	0.656	-	74.9%	-	0.943	-	0.739	-
CTransPath	68.8%	↑	0.927	-	0.690	↑	78.0%	↓↓*	0.941	-	0.768	↓*
Hibou-B	66.1%	↓	0.911	↓↓	0.667	↓	78.0%	↓↓↓*	0.958	-	0.765	↓↓↓*
Phikon	68.0%	↑	0.912	↓	0.672	↓	80.1%	↑	0.941	-	0.792	↑
Kaiko-B8	62.7%	↓↓↓*	0.907	↓	0.633	↓↓↓*	79.0%	↓↓	0.949	-	0.786	↓↓
GPFM	69.4%	↓	0.912	↓*	0.690	↓	84.0%	↑	0.953	-	0.831	↑
UNI	67.1%	↓↓↓*	0.915	↓↓*	0.684	↓↓↓*	82.0%	↓↓↓*	0.962	-	0.806	↓↓↓*
Hibou-L	58.7%	↓↓↓	0.889	↓↓*	0.622	↓↓↓	75.0%	↓↓↓*	0.959	-	0.730	↓↓↓*
Virchow	65.2%	↓↓*	0.896	↓↓	0.658	↓↓	81.0%	↓↓	0.956	-	0.801	↓↓
Virchow2-CLS	69.8%	↓↓	0.917	↓*	0.681	↓↓↓	89.1%	↑	0.963	-	0.883	↑
H-optimus-0	66.1%	↓↓↓	0.916	↓↓	0.678	↓↓	85.0%	↓↓	0.965	-	0.843	↓↓
Prov-GigaPath	67.9%	↓↓	0.919	-	0.675	↓↓↓	83.0%	↓	0.949	-*	0.820	↓

Table 7.14 Results of internal validations without hyperparameter tuning. Arrows indicate the absolute difference in performance compared to the tuned models, with one arrow (↑) for difference a of at least 1%, two arrows (↑↑) for a difference of at least 3%, and three arrows (↑↑↑) for a difference of at least 5%. *Indicates a p-value less than 0.05 when compared to the tuned model. While the Prov-GigaPath AUROC only exhibited a reduction of 0.009 in hold-out testing, this was found to be statistically significant.

	Transcanadian Study			OCEAN Challenge		
Feature Extractor	**Balanced Accuracy**	**AUROC**	**F1 Score**	**Balanced Accuracy**	**AUROC**	**F1 Score**
RN50	75.8% ↑↑↑	0.969 ↑↑	0.769 ↑↑↑	51.0% ↓	0.857 ↓	0.411 -
RN18	80.1% ↑	0.946 ↓	0.807 -	49.0% ↓	0.837 -	0.370 ↓↓
ViT-L	68.9% ↓↓↓*	0.960 ↓	0.702 ↓↓↓*	48.3% ↓↓↓*	0.843 ↓↓*	0.506 ↓↓↓*
RN18-Histo	69.3% ↑	0.942 ↑	0.689 ↑↑	55.0% ↓	0.849 -	0.504 ↓
Lunit	92.4% ↓	0.989 -	0.879 ↓↓↓	68.8% ↓↓	0.935 ↓*	0.661 ↓↓↓
RN50-Histo	91.8% ↓	0.995 -	0.902 ↓↓	66.6% ↓	0.934 ↓*	0.684 -
CTransPath	88.1% -	0.978 -	0.847 ↓	68.1% -	0.934 -	0.686 ↑
Hibou-B	85.3% ↓↓↓	0.987 -	0.871 ↓↓↓	64.0% ↓	0.928 -	0.604 ↓
Phikon	94.6% ↑↑	0.996 -	0.944 ↑	63.2% ↓↓	0.903 -	0.598 ↓↓
Kaiko-B8	93.3% ↓↓	0.996 -	0.926 ↓	64.1% ↓↓↓	0.929 ↓*	0.596 ↓↓*
GPFM	97.7% -	0.998 -	0.964 ↓	73.8% -	0.937 -	0.725 ↓
UNI	95.3% ↑	0.996 -	0.952 ↑↑	69.6% ↓↓↓*	0.948 -*	0.693 ↓↓↓*
Hibou-L	81.7% ↓↓↓	0.988 -*	0.813 ↓↓↓*	64.7% ↓↓	0.936 ↓*	0.615 ↓↓
Virchow	88.8% ↑	0.991 -	0.864 ↑	78.5% -*	0.948 ↓*	0.766 -
Virchow2-CLS	91.6% ↑↑	1.000 -	0.915 ↑↑	74.0% ↓↓↓*	0.956 -	0.719 ↓↓*
H-optimus-0	99.0% ↑	1.000 -	0.991 ↑	74.8% -	0.951 -	0.747 ↑↑
Prov-GigaPath	89.4% -	0.993 -	0.871 -	75.4% -	0.957 -	0.720 -

Table 7.15 Results of external validations without hyperparameter tuning. Arrows indicate the absolute difference in performance compared to the tuned models, with one arrow (↑) for difference a of at least 1%, two arrows (↑↑) for a difference of at least 3%, and three arrows (↑↑↑) for a difference of at least 5%. *Indicates a p-value less than 0.05 when compared to the tuned model. In the OCEAN Challenge validation, the UNI AUROC only exhibited a reduction of 0.006, and Virchow balanced accuracy only exhibited a reduction of 0.7%, but these differences were found to be statistically significant.

Final Hyperparameters

Feature Extractor	Learning Rate (LR)	LR Decay Patience	LR Decay Factor	First Moment Decay	Second Moment Decay	Stability Parameter	Weight Decay	Dropout Rate	Max Patches	Model Size
RN50	2e-3	20	0.75	0.75	0.95	1e-2	1e-3	0.4	800	[512,128]
RN18	1e-4	20	0.9	0.8	0.99	1e-4	1e-5	0.5	700	[1024,256]
ViT-L	5e-5	10	0.35	0.85	0.999	1e-3	1e-1	0.0	800	[512,384]
RN18-Histo	2e-4	20	0.9	0.9	0.99	1e-4	1e-4	0.6	1000	[512,512]
Lunit	1e-4	10	0.75	0.99	0.9999	1e-5	1e-1	0.6	900	[1024,512]
RN50-Histo	2e-4	25	0.75	0.8	0.99	1e-4	1e-3	0.6	700	[512,384]
CTransPath	1e-4	25	0.9	0.7	0.99999	1e-3	1e-3	0.4	1000	[256,128]
Hibou-B	4e-5	10	0.9	0.99	0.9999	1e-3	1e-2	0.3	1600	[256,128]
Phikon	5e-5	25	0.75	0.99	0.999	1e-5	1e-5	0.8	1200	[512,256]
Kaiko-B8	2e-5	10	0.75	0.95	0.9999	1e-5	1e-1	0.2	600	[512,128]
GPFM	1e-4	25	0.9	0.95	0.99	1e-4	1e-6	0.8	1000	[512,128]
UNI	1e-5	10	0.75	0.9	0.999	1e-5	1e-3	0.0	1000	[512,256]
Hibou-L	5e-5	25	0.75	0.75	0.99999	1e-4	1e-7	0.6	400	[256,128]
Virchow	2e-4	20	0.9	0.95	0.99	1e-3	1e-2	0.8	1100	[512,256]
Virchow2-CLS	2e-5	10	0.75	0.55	0.999	1e-4	1e-4	0.6	1000	[512,256]
H-optimus-0	2.5e-5	5	0.75	0.5	0.9999	1e-4	1e-2	0.4	1000	[128,32]
Prov-GigaPath	5e-5	15	0.75	0.7	0.99	1e-4	1e-4	0.7	1300	[512,256]
RN50 Reinhard	2e-3	25	0.75	0.75	0.95	1e-2	1e-3	0.4	400	[512,256]
RN50 Macenko	2e-3	15	0.75	0.85	0.95	1e-2	1e-3	0.3	400	[512,128]
RN50 Otsu	2e-3	15	0.9	0.75	0.95	1e-2	1e-3	0.1	600	[512,256]
RN50 Otsu+Macenko	2e-3	25	0.9	0.75	0.99	1e-3	1e-4	0.3	1000	[512,256]
RN50 5Augs	1e-3	25	0.6	0.8	0.99	1e-4	1e-4	0.4	700	[128,32]
RN50 10Augs	2e-3	20	0.75	0.8	0.99	1e-2	1e-3	0.4	700	[512,256]
RN50 20Augs	1e-3	20	0.75	0.7	0.999	1e-3	1e-4	0.6	1000	[512,128]

Table 7.16 The final hyperparameters of each model determined by an iterative grid search tuning procedure using five cross-validation folds, including the models from the ablation study. These are grouped into hyperparameters relating to the learning rate, Adam optimizer, regularisation, and model architecture. The model size is the number of parameters in the attention layer and subsequent fully connected layer.

The optimal hyperparameters (Table 7.16) typically did not vary greatly for models using the same feature extraction backbone, with a few notable exceptions. The regularisation hyperparameters (weight decay, dropout rate, max patches) varied greatly across all models, including those with the same backbone. The classifier based on the five-times augmented training data was the smallest ResNet50-based classifier by far (and had the smallest stability parameter and LR decay factor), with only 0.1M parameters compared to the next smallest at 0.7M. The ViT-based models had between 0.2M (H-optimus-0) and 1.6M parameters (Virchow). The largest ViT-based encoders typically had smaller values for the first moment decay (0.5-0.75) than the smaller ViT-based encoders (0.9-0.99). Other hyperparameters were relatively stable within a given backbone architecture.

Some hyperparameters varied greatly between model architectures. The learning rate was much smaller for ViT-based models (0.00001-0.0002) than ImageNet-pretrained ResNet50 models (0.001-0.002) and often had a faster rate of decay. The Adam optimiser first and second moment decay parameters were also often higher in ViT-based models than in ResNet50 models. Other hyperparameters were relatively consistent between model architectures.

7.4 Discussion

In this chapter, we have thoroughly compared the effects of different patch feature extractors on the slide-level classification of ovarian carcinoma morphological subtypes. The results clearly indicated that transformer-based histopathology foundation models improved classification performance when compared to non-domain-specific and ResNet-based feature extractors, with 13 out of 14 foundation models outperforming all ImageNet-pretrained models in all evaluations. The only foundation model which did not exceed ImageNet-pretrained model performance was RN18-Histo, which was the single worst-performing model in hold-out testing and external validation on the Transcanadian Study dataset, though it did outperform the ImageNet-pretrained ResNet models in the other two validations. RN18-Histo was the earliest published histopathology foundation model and as such it was one of the few foundation models to not use a transformer-based backbone. In this study, RN18-Histo was also the smallest foundation model, had the second-smallest feature space, and was pretrained with the second-smallest dataset.

As shown in Figure 7.8, in most validations there was a slight positive relationship between performance (specifically, balanced accuracy) and each of the foundation model size and pretraining dataset size. These relationships were fairly weak, with the relationship between performance and foundation model size having R^2 values between 0.02 and 0.36, and the relationship between performance and pretraining dataset size between -0.01 and 0.24 (though the relationship between performance and dataset size was unduly influenced by the particularly large dataset used in Virchow2-CLS, with this causing clear outliers in Figure 7.8). The trends were weakly positive for three validations, but there was no trend found in the Transcanadian validation. It is not clear why this occurred, though as performance was consistently

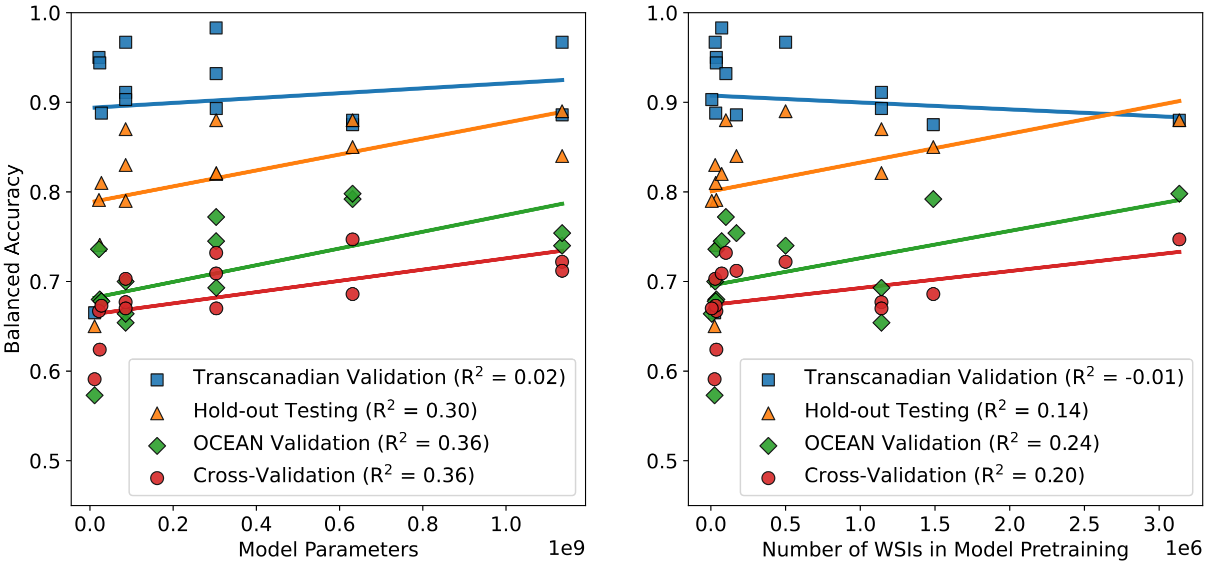


Figure 7.8 Balanced accuracy results for each histopathology foundation model-based classifier in each validation shown in relation to the number of model parameters and number of WSIs used in the pretraining of the foundation model. The line of best fit and the corresponding coefficient of determination (R^2) are provided for each validation.

high on this dataset, it may be that a smaller model was sufficient. The greatest performance in most validations was achieved by one of the largest models (Virchow, Virchow2-CLS, H-optimus-0), though the smaller GPFM model performed best in the Transcanadian Study external validation, and the single largest model (Prov-GigaPath) did not achieve optimal results in any validation. Three models were trained with over one million WSIs, with two being among the best-performing models (Virchow, Virchow2-CLS), and the other being one of the worst-performing ViT-based foundation models overall (Hibou-B).

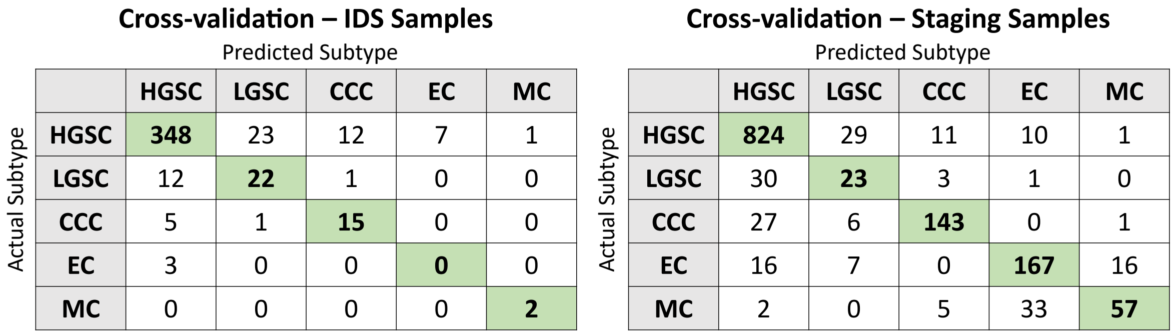
To investigate which foundation models outperformed expectations, we investigated which models had positive residuals of at least 1% when compared to the lines of best fit in Figure 7.8. UNI and Kaiko-B8 consistently performed better than expected given their foundation model size, with GPFM and Virchow2-CLS performing better than expected in three of four validations. The UNI and GPFM models consistently performed better than expected given the pretraining dataset size, with Kaiko-B8, Virchow2-CLS and H-optimus-0 all better than expected in three of four validations. These results indicate that UNI is particularly data-efficient and computationally-efficient for a foundation model of its ability. Where the H-optimus-0 classifier took an average of 425s per WSI, UNI took only 100s (24% as long) with a reduction of only 0.1% average balanced accuracy across the four validations (Table 7.7). It was not clear how UNI outperformed expectations in this way, with similar overall methodologies employed in training models which did not achieve such great results. The proportion of gynaecological WSIs in the UNI training set (5.8%) was exceeded in the training of several other models [172, 224, 227, 229, 230], though for most models it was not clear what proportion of the training set was specifically composed of the five ovarian carcinoma subtypes of interest, so it was not clear whether this was an influential factor.

Different preprocessing techniques often had little impact on internal performance (likely due to the homogeneity of the single-centre dataset) and the OCEAN Challenge validation, but they did aid the generalisability to the Transcanadian Study dataset. There was a modest positive trend between the number of augmentations used and the resulting model performance which may continue beyond the 20 augmentations per image used herein, though this may not be worth the considerable associated computational burden since the normalisation approaches achieved a similar level of performance. No individual normalisation, augmentation, or tissue detection approach consistently improved performance, with each giving worse performance than the baseline in at least one validation, and no statistically significant benefits found. As such, we believe there is much greater value in selecting the optimal feature extractor than there is in applying varied preprocessing techniques in the training of a WSI classifier. This conclusion was also found in a recent study [254] which investigated 14 different feature extractors using ABMIL in the context of breast and colorectal cancers (without hyperparameter tuning).

Hyperparameter tuning the ABMIL classifier had a modest but often significantly beneficial effect on classification performance. This did not necessarily need to be extensive to provide a benefit, with a large proportion of the benefit obtained simply by adjusting the learning rate and model size. It is worth noting that the ABMIL classifiers were orders of magnitude smaller than most of the feature extraction models, making it much more computationally feasible to tune the classifiers rather than the feature extractors. The variability in the benefits may reflect both the fitness of the originally selected hyperparameters and the versatility of the models. The original hyperparameters were taken from our previous work using the ImageNet-pretrained

ResNet50 (Chapter 6), so the hyperparameters were likely better suited to this feature extractor than those which used different architectures and training datasets. Most of the benefit of hyperparameter tuning on the validation loss was achieved by adjusting the learning rate and the size of the ABMIL classifier, so just tuning these may be a more computationally efficient approach to improve model performance and the robustness of validations.

Performance was generally higher in hold-out testing than in cross-validation and was higher still in the external validation with the Transcanadian Study dataset. However, the external validation with the OCEAN dataset gave a similar performance to that of cross-validation. This may be influenced by the diagnostic quality of the data, with the internal cross-validation dataset incorporating post-chemotherapy WSIs and the OCEAN dataset being unclear in this regard. Validations using only staging data achieved optimal balanced accuracies of 89% and 97%, compared to only 75% and 80% in the validations including IDS samples which can pose diagnostic challenges (Section 2.2). In cross-validation, the balanced accuracy for IDS samples was only 64.7% (with all EC slides incorrectly classified), compared to 71.0% for primary surgery samples (Figure 7.9). The challenge posed by neoadjuvant treatment is recognised by pathologists, and it is recommended in these cases that tumour subtyping is performed using pre-treatment biopsies rather than resection specimens [255].



Cross-validation – IDS Samples

Predicted Subtype

Actual Subtype		HGSC	LGSC	CCC	EC	MC
	HGSC	**348**	23	12	7	1
	LGSC	12	**22**	1	0	0
	CCC	5	1	**15**	0	0
	EC	3	0	0	**0**	0
	MC	0	0	0	0	**2**

Cross-validation – Staging Samples

Predicted Subtype

Actual Subtype		HGSC	LGSC	CCC	EC	MC
	HGSC	**824**	29	11	10	1
	LGSC	30	**23**	3	1	0
	CCC	27	6	**143**	0	1
	EC	16	7	0	**167**	16
	MC	2	0	5	33	**57**

Figure 7.9 Confusion matrices for the optimal ABMIL classifier with features from the H-optimus-0 foundation model in cross-validation broken down by treatment status. Correct classifications are indicated in green.

Two pathologists (KA and NMO) reviewed a subset of 100 WSIs in the OCEAN set and found that eight exhibited extensive TMA coring, two were almost entirely necrotic, and one displayed image stitching problems. Furthermore, the staining and colour balance was inconsistent across this cohort, which comprised both biopsies and resection specimens. These characteristics may have contributed to the poorer performance noted on the OCEAN dataset. In contrast, the Transcanadian Study set contained a single representative staging slide of the tumour per patient and the slides were largely devoid of artefacts. This particularly high-quality data may represent a best-case research scenario, rather than a more realistic representation of the variable quality and tumour content of clinical slides, where guidance recommends the sampling of heterogeneous areas of tumour that have the potential to compromise the quality of slide preparation and interpretation, with features such as calcification or necrosis. The hold-out and external validations likely also benefitted from the five-fold ensembled predictions when compared to the five-fold cross-validation. While this is the most comprehensive study of AI ovarian cancer subtyping to date, the relatively small size of the test sets still resulted in a high level of uncertainty, as reflected in the wide confidence intervals. Thus, part of the difference in performance between datasets may be attributed to random chance.

The results in this chapter are similar to those of the only previous studies to use large ovarian cancer subtyping datasets (each with around 1000 WSIs) [82, 136, 171]. One study presented a multi-scale graph model [82] and reported an optimal cross-validation balanced accuracy of 73% and an F1 score of 0.69. Another [136] evaluated four MIL approaches and reported an optimal cross-validation balanced accuracy of 81%, AUROC of 0.95, and F1 score of 0.79. In an external validation using an ensemble of cross-validation models on 60 WSIs, the authors reported a balanced accuracy of 80%, AUROC of 0.96, and F1 score of 0.81. The final study focused on adversarial domain adaptation [171] and achieved optimal internal and external balanced accuracies of 80% and 83% from a CTransPath-based MIL classifier. Other studies applying foundation models to ovarian cancer subtyping have reported optimal balanced accuracies of 82% and $\sim$88% using UNI on the OCEAN dataset and Prov-GigaPath on an internal dataset, respectively [172, 173]. These comparisons are provided for context and should not be considered to be conclusive given the differences in the datasets used. A sparsity of publicly available data has limited external validations in most previous research [1], and for the largest accessible dataset (the OCEAN Challenge set) very little information has been provided about the data provenance.

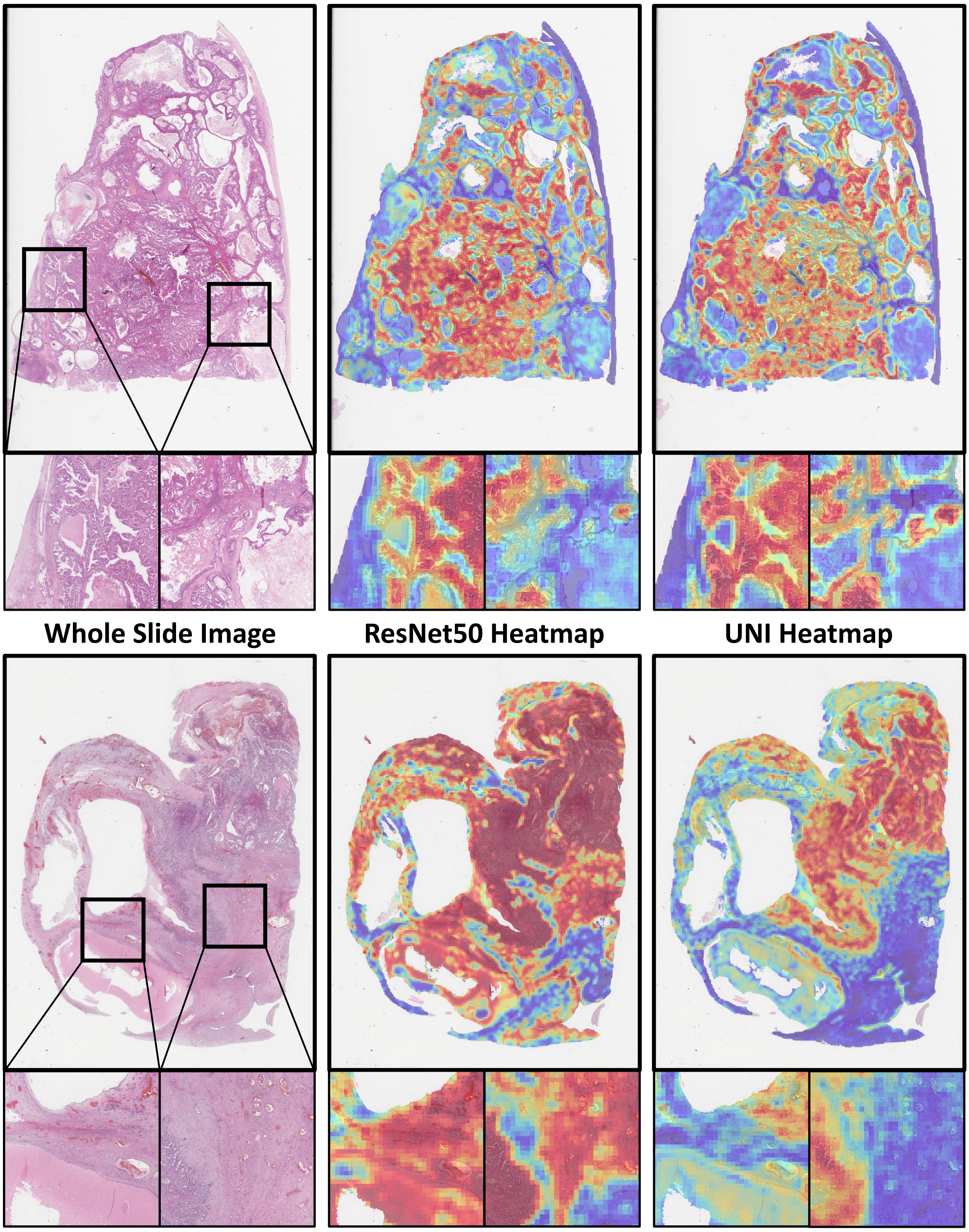


Figure 7.10 Attention heatmaps from the ABMIL classifier using the ImageNet-pretrained ResNet50 and UNI foundation model features. (Upper) A typical difference between heatmaps with different diagnoses. (Lower) The most extreme qualitative difference found between heatmaps in the internal test set. In both examples, the UNI classification was correct (upper - MC, lower - CCC), and the ResNet50 classification was incorrect (upper - EC, lower - MC). These heatmaps are based on 256 x 256 pixel patches with 50% overlap at 10x apparent magnification, with visual differences caused by the variable size of resection samples.

To qualitatively analyse the differences between foundation models and CNNs, two pathologists (KA and NMO) qualitatively compared the ABMIL attention heatmaps generated using the ImageNet-pretrained ResNet50 and the UNI foundation model (Figure 7.10). Most heatmaps were well-focused on tumour and relevant stromal regions for both models, with often only subtle differences between them. The UNI-based heatmaps generally indicated a slightly greater focus on tumour tissue, whereas the ResNet50 model also paid attention to some stromal regions of variable diagnostic relevance (Appendix E). Attention heatmaps can be useful for identifying potential sources of error but should be interpreted with caution since they cannot provide a complete explanation of classification decisions [256].

All of the WSIs which were misclassified by the optimal H-optimus-0 model (Figure 7.3) in hold-out testing were reviewed by the pathologists involved in the study, who found that the majority (6/11) had incorrect ground truth labels, and had been correctly classified by the model. This underscores the value of the model in detecting the human errors which occur in the production of large-scale repositories. A subsequent review to identify any possible further labelling errors affecting internal data did not locate any issues. The five slides that were truly misclassified by the model in hold-out testing (three CCC classified as HGSC, one CCC classified as LGSC, and one LGSC classified as EC) showed the typical morphology (both architectural and cytological) of their true subtypes, making it unclear why these errors occurred.

The pathologists also reviewed a selection of misclassified slides in cross-validation. The 42 EC slides classified as other subtypes all exhibited potentially confusing morphological features that occur within the broad spectrum of EC, including villoglandular and papillary architecture as well as foci of mucinous and squamous metaplastic differentiation, and squamous morule formation. ECs misclassified as HGSCs were of a higher grade and featured both greater nuclear pleomorphism and a more solid growth pattern. It would be interesting to determine whether any of these misclassifications reflect shared genetic features. The most commonly confused subtypes were HGSC and LGSC, which is not surprising considering their similar histoarchitecture. These entities were historically considered a single entity with a three-tier grading system until the characterisation of their distinct molecular alterations and clinical behaviours [257]. Collecting additional training data may help to improve the discrimination of these similar subtypes, with LGSC having only formed 5% of the training set (Section 4.2) due to it being a relatively uncommon subtype.

The strong performance of the foundation models was particularly impressive considering that they were applied here at 10x magnification, despite often only being trained using 20x magnification data. This was a practical computational limitation when performing hyperparameter tuning, as 20x magnification tissue would produce approximately 4 times as many patches per WSI as 20x magnification tissue, thus quadrupling the total runtime. While we previously found 10x magnification to be best when using the ImageNet-pretrained ResNet50 (Chapter 6), it may not have been optimal when using foundation models that had typically been trained at 20x magnification. However, a previous study of foundation models for slide-level classification found no consistent benefit from increasing to 20x magnification [254].

In this chapter, we reported the second-highest ever external validation performance of an AI model for ovarian cancer subtyping (behind only our subsequent graph model in Chapter 8), with 97% balanced accuracy on the Transcanadian Study dataset. However, results were variable across datasets. The improved performance from histopathology foundation models is promising for the potential clinical utility of these AI approaches, though further work is required to ensure that the models generalise to all relevant sources of variation, especially across different histopathology labs and slide scanners. This may require larger, more diverse training datasets. Models should be made robust to the influence of lower-quality data and artefacts to reduce the burden of quality control. Ideally, models should also be able to accurately classify post-treatment tissue, though if this proves infeasible it may be necessary to restrict the scope of the models to the classification of high-quality primary surgery tissue samples, for which these models already excel. Furthermore, it is currently unclear how best to present automatically generated information to pathologists to assist them, rather than to distract, frustrate, or confuse them. This may require improved model interpretability and a measure of model uncertainty, especially considering the existence of rare subtypes which are notoriously difficult to collect sufficient data on outside the context of multi-centre collections.

Ideally, algorithms would be made more computationally efficient for use in the clinic, but the best-performing foundation models are much less computationally efficient than the ResNet. This problem is exacerbated by the limited digitisation of histopathology services, with most pathological diagnoses still made under a microscope. AI adoption will be contingent on it being accessible and beneficial given limited computational infrastructure and users who may not be technological experts. While various issues are inhibiting the clinical translation of ovarian cancer subtyping models, these seem increasingly likely to be overcome in the near future.

7.5 Conclusion

In this chapter, we conducted a rigorous validation of feature extraction methods for ovarian cancer subtyping. We found that the features generated by histopathology foundation models drastically improved classification performance when compared to ImageNet-pretrained feature extractors. Several different data preprocessing techniques were evaluated in an attempt to improve the performance of the ImageNet-pretrained ResNet50 baseline, and while these somewhat improved performance, they were far from sufficient to match the performance of the foundation models. Through a five-fold ensemble of ABMIL classifiers, the best overall foundation model, H-optimus-0, achieved a five-class balanced accuracy of 89% on internal test data and 97% and 80% on external test sets, compared to 68%, 81%, and 55% respectively for the best ImageNet-pretrained ResNet models. This represents the greatest performance for the ovarian carcinoma subtype classification task in any peer-reviewed literature to date. The largest models and those pretrained with the largest datasets generally gave the best performance, though the UNI foundation model was one of the best-performing models despite a relatively moderate model and dataset size, giving an average balanced accuracy only 0.1% lower than H-optimus-0 while running over 4 times as fast. Hyperparameter tuning the classifiers improved classification performance by a median of 1.9% balanced accuracy, although this was variable. While the improved classification performance offered by histopathology foundation models may be sufficient for clinical implementation, the need to address logistical hurdles and conduct larger-scale validations remains.

Chapter 8

Multi-Resolution Histopathology Patch Graph Networks

In this chapter, we describe a novel multi-resolution graph network for slide-level classification. This model uses a histopathology foundation model to extract patch features at multiple magnifications, and we assess several techniques for combining these features into a single graph. We use attention-based graph layers to prioritise patches and share information in spatial neighbourhoods, giving a more complete slide representation for classification. We use the same robust training and validation procedures and datasets as in the previous chapter to determine whether capturing spatial relationships provides a benefit in ovarian carcinoma subtyping.

8.1 Introduction

Many MIL models (including ABMIL) treat all instances as functionally independent of one another. This misses the inherent spatial relationships between neighbouring patches and hence does not model the local tissue context around each patch. GNNs [102] offer an approach to model these spatial relationships. Graphs are composed of nodes and edges. Each node contains some local information, and GNN message-passing layers are used to share information along edges to provide contextual information from connected nodes. A message-passing layer updates node features based on first-order neighbours, and by stacking multiple such layers, information can be passed from distant parts of the graph.

Given the relatively high computational complexity of cell graphs (Section 2.4.5), we focus on patch graphs, where each graph node represents a tissue patch. While pathologists analyse tissue at multiple magnifications, slide-level patch graphs have typically used data at only a single magnification [99, 258–261]. When multi-resolution graphs have been implemented, they have often sampled patches to reduce the computational complexity and balance the relative importance of different magnifications [82, 262], though this discards potentially relevant diagnostic information. This has also been the case for non-graph MIL methods for ovarian cancer subtyping [116–118]. To ensure a rigorous analysis of each slide, we instead follow the strategy of using all available tissue at multiple magnifications using a multi-level grid structure [263–265].

Only one previous study has applied GNNs to ovarian cancer subtyping [82], where it was reported that a novel multi-resolution graph model gave a better balanced accuracy than other MIL methods including ABMIL, TransMIL [71], and single-magnification graph models. This study used only a single set of hyperparameters and a single dataset, making it unclear whether all models were optimally tuned to the given task and data, and whether the models would generalise well to external data.

In this chapter, we present the most thorough evaluation of a GNN for ovarian cancer subtyping to date, including hyperparameter tuning and both hold-out and external validations. To the best of our knowledge, it was also the first multi-resolution graph model implemented using features from the vision transformer (ViT)-based histopathology foundation model, UNI [208].

8.2 Methods

8.2.1 Graph Model Pipeline

The WSI classification pipeline (Figure 8.1) included tissue patch extraction, patch feature extraction, graph modelling for patch aggregation, and slide classification. The tissue patch extraction procedures were the same as in the baseline model (Section 4.1), with 256 x 256 pixel downsampled patches extracted at 5x, 10x, and 20x magnifications. At the native 40x magnification, this required taking 512 x 512 patches before downsampling for 20x, 1024 x 1024 patches for 10x, and 2048 x 2048 pixel patches for 5x, with each doubling of the apparent magnification quadrupling the number of resulting patches. This was the only step that differed for the external datasets, with smaller patches required before downsampling given the lower original magnification. Features were extracted from all downsampled patches using the UNI foundation model [208] (requiring further downsampling from 256 x 256 to 224 x 224 pixels), given its exceptional performance in the previous analysis of feature extractors (Chapter 7). The models were also evaluated using the standard ImageNet-pretrained ResNet50 feature extractor on 256 x 256 pixel patches [58, 59] to better understand

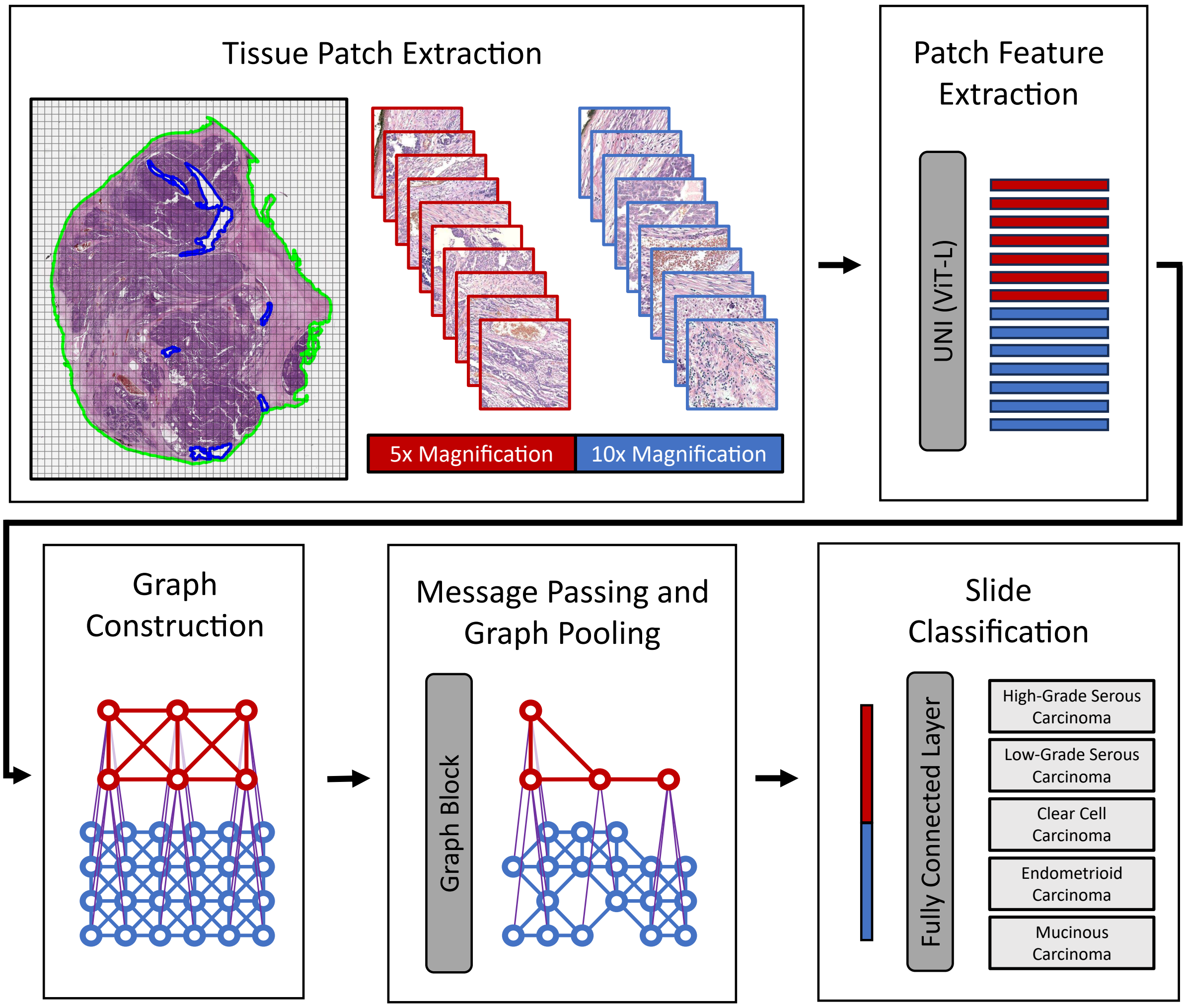


Figure 8.1 Multi-resolution graph model pipeline for slide-level classification, illustrated using 5x and 10x magnification tissue patches. Graph blocks were composed of at least one GATv2 message-passing layer [266] followed by a SAGPool graph-pooling layer [267].

whether the UNI features were truly more discriminative, or whether the benefit was dependent on using the ABMIL classifier.

As shown in Figure 8.2, graphs were constructed such that each patch was connected to any other patch within a given spatial radius, which was set to allow connections to first-order lateral and diagonal neighbours. Connections were also made between patches showing the same tissue at different magnifications, with each low-magnification patch connected to four high-magnification patches (or fewer if not all high-magnification patches contained tissue).

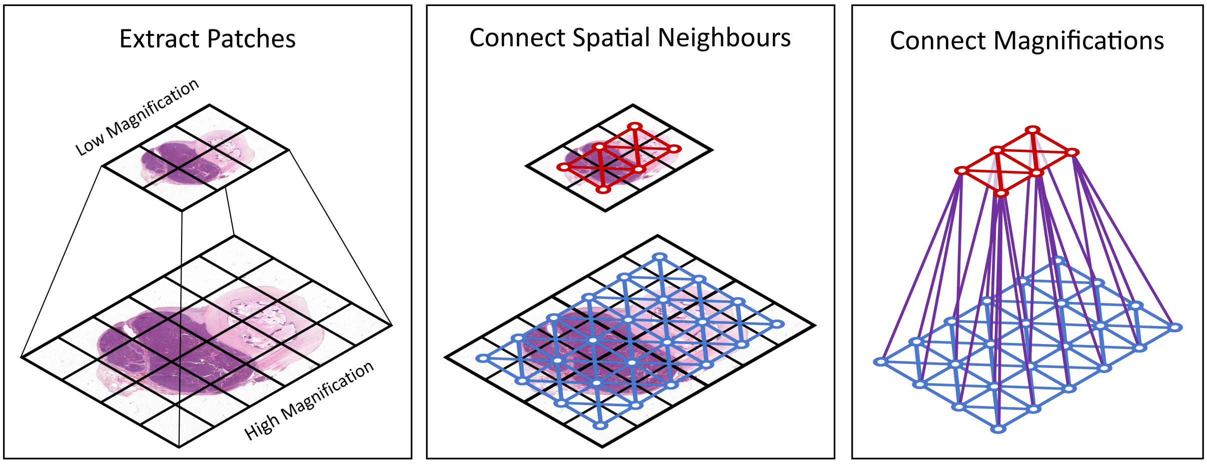


Figure 8.2 Multi-resolution graph construction, with each node representing a single patch. Actual graphs have many more patches/nodes, with an average of 1138 tissue patches per slide at 5x magnification and 4423 tissue patches per slide at 10x magnification.

Each graph model block contained at least one graph attention (GATv2) convolution layer [266] for message passing, followed by a ReLU activation and a self-attention graph pooling (SAGPool) [267] layer to reduce the number of nodes in the graph. The trainable attention mechanisms weighed the relative importance of the nodes, with GATv2 using node features to prioritise neighbours during message passing, and SAGPool using node features and the graph topology to prioritise important nodes during pooling. The GATv2 attention score $a_{i,j}$ for the edge between nodes i and j is calculated as:

$$a_{i,j} = \text{softmax}_j\{\mathbf{w}^\top \sigma(\mathbf{W} \cdot [\mathbf{x}_i || \mathbf{x}_j])\}, \tag{8.1}$$

where σ is the *Leaky ReLU* activation function, $||$ is the concatenation function, m' is the chosen output node feature dimension, and $\mathbf{w} \in \mathbb{R}^{2m'}$ and $\mathbf{W} \in \mathbb{R}^{m' \times m}$ are the trainable weight vector and matrix, respectively. The SAGPool attention score vector $\mathbf{a} \in \mathbb{R}^n$ is calculated as:

$$\mathbf{a} = \sigma(\tilde{\mathbf{D}}^{-0.5}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-0.5}\mathbf{X}\mathbf{p}), \tag{8.2}$$

where σ is the *tanh* activation function, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ is the graph adjacency matrix with self-connections and corresponding diagonal degree matrix $\tilde{\mathbf{D}} \in \mathbb{R}^{n \times n}$, and $\mathbf{p} \in \mathbb{R}^m$ is the trainable weight vector.

The outputs of each graph block were pooled using both mean and max pooling across all remaining nodes, and these pooled features were concatenated together to make a double-length feature set. All graph block outputs were summed to form a WSI-level feature set (to which dropout was applied during training), and finally, these were classified through a single fully connected network layer with five output neurons corresponding to the five ovarian cancer subtypes.

One complexity in extending GNNs to multiple resolutions is in handling the features. Different biological entities are represented by features at different magnifications; thus, it may be naive to share the same features across magnifications. Further, the vastly more common high-magnification patches may have an undue influence in a shared feature space. Previous studies have concatenated features from different resolutions [82, 262], though it is unclear whether this is beneficial. We compared the 'naive' approach in which the same features are shared across magnifications (making the model magnification-agnostic) to two approaches in which there was a separate set of features for each magnification. Each graph node initially represented a tissue patch at only a single magnification, and so the features for the other magnification were either initially set to zero ('concat_zero') or to the average of all patch features at the relevant magnification ('concat_avg'). The previous ovarian cancer GNN [82] instead directly extracted multi-magnification features by analysing only one high-resolution patch within each lower-resolution patch, though this discarded most of the high-resolution tissue, where our proposed approach used all available tissue at multiple magnifications.

8.2.2 Hyperparameter Tuning and Validation Procedures

GNN hyperparameters were tuned through the standard iterative grid search procedure (Section 4.3), starting from the optimal hyperparameters of the UNI-based ABMIL model in Chapter 7. At least 100 unique configurations were evaluated for each graph-based model, with the ABMIL results taken directly from Chapter 7. As shown in Table 8.1, a total of 13 hyperparameters were tuned for the GNNs, including nine previously described hyperparameters (Section 4.3) and four GNN architecture hyperparameters controlling the number of **message-passing layers** per graph block, the number of **graph blocks** (and hence the number of pooling layers), the graph **pooling factor** per graph block, and the **node feature dimension** per magnification. The node feature dimension hyperparameter was applied within the first message passing layer, where the dimension of the node features was reduced from the input dimension (from the patch feature extractor) to the selected size, and this size was maintained through the remaining graph network layers. In experiments where separate features were used for two different magnifications, the number of node features was doubled to retain the separate concatenated features for each magnification throughout the network.

	Tuning Iteration									
Hyperparameter	1	2	3	4	5	6	7	8	9	10
Learning Rate (LR)	✓							✓		
LR Decay Factor							✓			
LR Decay Patience							✓			
First Moment Decay (β_1)					✓					
Second Moment Decay (β_2)					✓					
Stability Parameter (ϵ)								✓		
Weight Decay						✓				
Dropout Rate						✓				
Max Patches	✓								✓	
Message-Passing Layers		✓								✓
Graph Blocks		✓	✓	✓						✓
Pooling Factor				✓						
Node Feature Dimension			✓						✓	

Table 8.1 Iterative hyperparameter tuning procedure, with check marks (✓) indicating the hyperparameters that were adjusted at each stage of tuning, with all others frozen. These are grouped into hyperparameters relating to the learning rate, Adam optimizer, regularisation, and model architecture.

Seven models were evaluated to compare feature extractors, magnifications, and MIL models. For the comparison of different architectures, we first created a baseline GNN which was a multi-resolution graph at 5x and 10x magnifications (chosen based on the analysis of magnifications in Chapter 6), using separate magnification-specific features with average initialisation (concat_avg). Comparisons were conducted using a multi-resolution GNN at higher magnifications (10x and 20x), a single-magnification GNN at 10x, and ABMIL at 10x. Another comparison swapped the UNI vision transformer feature extractor to an ImageNet-pretrained ResNet50. To compare different multi-resolution feature spaces, the baseline approach (concat_avg) was compared to the separate features with zero initialisation (concat_zero) and to magnification-agnostic features (naive). Each model was compared to the baseline using a paired *t*-test across the five cross-validation folds, with p-values adjusted for multiple testing using a false discovery rate correction [206].

The finalised LTHT dataset was used in this chapter, consisting of 1864 WSIs of adnexal tissue from 433 patients for training and 5-fold cross-validation, and a further 100 WSIs from 30 independent patients for hold-out testing. The Transcanadian Study dataset [203], consisting of 80 WSIs from 80 patients, and the OCEAN Challenge dataset [160], consisting of 513 WSIs from an unknown number of patients, were both used for external validation. These datasets are described further in Chapter 4.2. All validations were conducted using the HPC (Section 4.5), and the code was made available online at https://github.com/scjjb/MultiscalePathGraph.

8.3 Results

8.3.1 Hyperparameter Tuning Results

The best hyperparameters from tuning are shown in Table 8.2. The smallest tuned classifiers were the single-resolution (10x GNN, 0.5M parameters; 10x ABMIL, 0.8M) and magnification-agnostic models (naive features, 0.7M), followed closely by the zero-initialised model (1.2M) and the higher magnification model (1.2M), with the largest being the baseline model (7.9M) and the ResNet-based GNN (10.5M). In most cases, the classifier was much smaller than the respective feature extractor, with the UNI model having 303M parameters and ResNet50 having 9M. The multi-resolution GNNs were typically larger than the single-resolution ABMIL classifiers in Chapter 7, which had 0.1-1.6M parameters, with most under 1M.

	Model						
Hyperparameter	**ABMIL 10x only**	**GNN Baseline**	**GNN 10x only**	**GNN 10x+20x**	**Naive features**	**Concat_zero features**	**ImageNet-ResNet50**
Learning Rate (LR)	1e-5	1e-4	5e-5	1e-4	2e-4	1e-4	2e-3
LR Decay Factor	0.75	0.9	0.9	0.9	0.45	0.9	0.6
LR Decay Patience	10	10	10	20	15	15	20
First Moment Decay (β_1)	0.9	0.9	0.95	0.95	0.9	0.95	0.8
Second Moment Decay (β_2)	0.999	0.9999	0.999	0.999	0.99999	0.99	0.95
Stability Parameter (ϵ)	1e-5	1e-5	1e-7	1e-7	1e-7	1e-7	1e-2
Weight Decay	1e-3	1e-2	1e-1	1e-2	1e-3	1e-2	1e-3
Dropout Rate	0.0	0.2	0.0	0.1	0.2	0.4	0.2
Max Patches	1000	6000	4000	14000	4000	5000	5000
Message-Passing Layers	N/A	3	1	1	1	1	1
Graph Blocks	N/A	4	1	2	2	2	4
Pooling Factor	N/A	0.9	0.6	0.6	0.45	0.75	0.6
Node Feature Dimension	512	512	256	256	256	256	1024

Table 8.2 Optimal hyperparameters for each model found through an iterative grid search on the validation sets from five-fold cross-validation. These are grouped into hyperparameters relating to the learning rate, Adam optimizer, regularisation, and model architecture.

8.3.2 Graph Model Validation Results

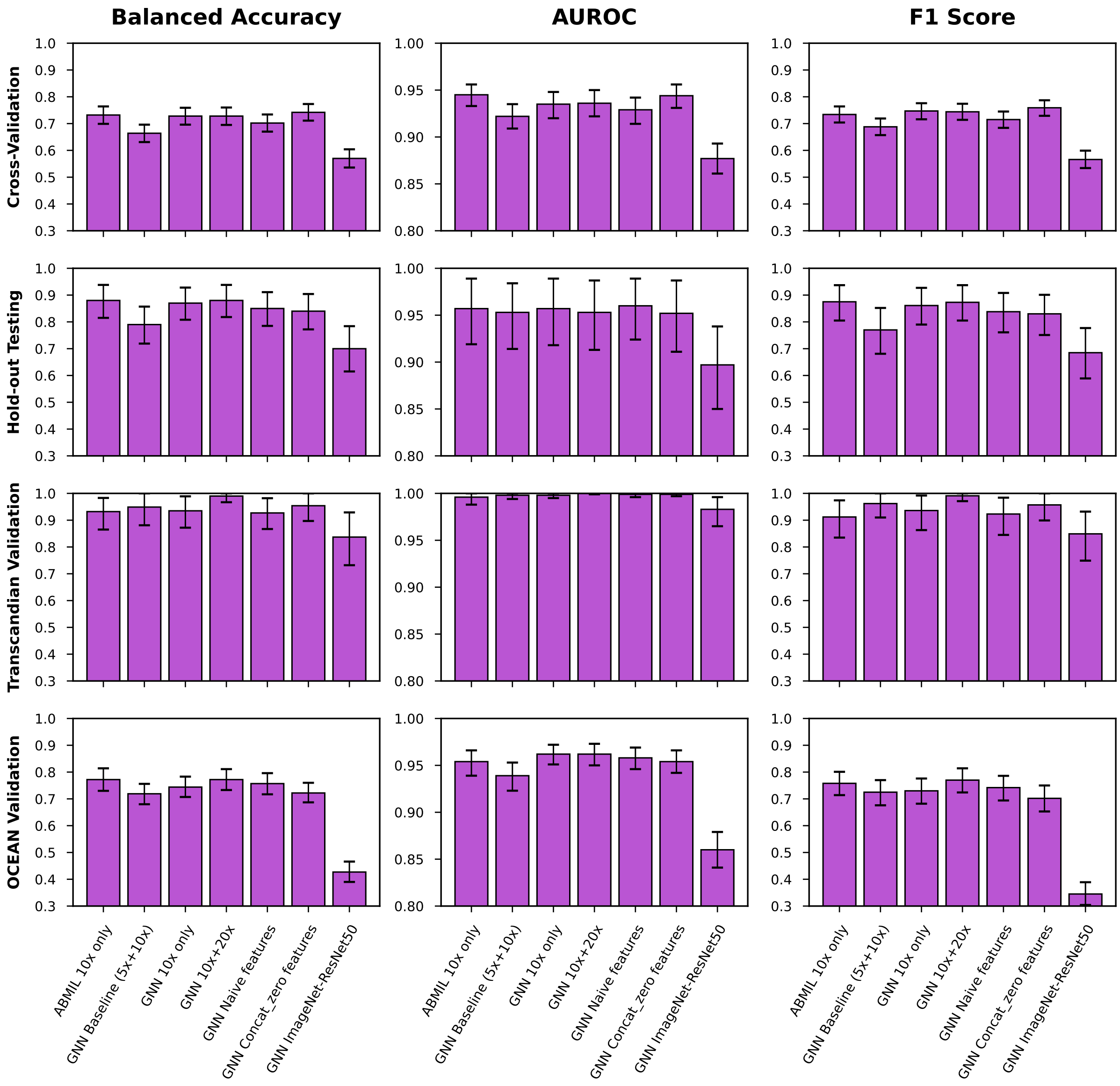


Figure 8.3 Ovarian cancer subtyping results (mean and 95% confidence interval from 10,000 iterations of bootstrapping) from cross-validation, internal hold-out testing, and external validations [160, 203]. In hold-out testing and external validations, predictions were ensembled from the five cross-validation models.

Results are shown in Figure 8.3. The best-performing model in cross-validation (Table 8.3) was the zero-initialised multi-resolution GNN, with a balanced accuracy of 74.2%, AUROC of 0.944 (second-best, behind the 0.945 of ABMIL), and F1 score of 0.759. ABMIL performed best in hold-out testing (Table 8.4), with a balanced accuracy of 88.0%, AUROC of 0.957 (second-best, behind the 0.960 of the naive GNN), and an F1 score of 0.875. In both external validations, the 10x+20x GNN performed best by all metrics, with a balanced accuracy of 99.0%, AUROC of 1.000, and an F1 score of 0.991 on the Transcandian Study dataset (Table 8.5), and 77.2%, 0.962, and 0.770 on the OCEAN Challenge dataset (Table 8.6).

Model	Balanced Accuracy	AUROC	F1 Score
ABMIL 10x only	73.2% (69.9-76.4%)	**0.945** (0.933-0.956)	0.734 (0.704-0.764)
GNN Baseline (5x+10x)	66.4% (63.1-69.6%)	0.922 (0.909-0.935)	0.688 (0.657-0.719)
GNN 10x only	72.8% (69.6-75.9%)	0.935 (0.920-0.948)	0.747 (0.716-0.776)
GNN 10x+20x	72.8% (69.5-76.0%)	0.936 (0.922-0.950)	0.744 (0.714-0.774)
GNN Naive features	70.2% (67.0-73.4%)	0.929 (0.914-0.942)	0.715 (0.684-0.745)
GNN Concat_zero features	**74.2%** (71.1-77.3%)	0.944 (0.931-0.956)	**0.759** (0.729-0.787)
GNN ImageNet-ResNet50	57.0% (53.6-60.4%)	0.877 (0.861-0.893)	0.566 (0.534-0.599)

Table 8.3 Cross-validation results shown as the mean and 95% confidence intervals generated by 10,000 iterations of bootstrapping. The best results are indicated in **bold**.

Model	Balanced Accuracy	AUROC	F1 Score
ABMIL 10x only	**88.0%** (81.5-93.8%)	0.957 (0.919-0.989)	**0.875** (0.805-0.937)
GNN Baseline (5x+10x)	79.0% (71.9-85.7%)	0.953 (0.914-0.984)	0.770 (0.681-0.852)
GNN 10x only	87.0% (80.8-92.8%)	0.957 (0.918-0.989)	0.861 (0.790-0.927)
GNN 10x+20x	**88.0%** (81.8-93.8%)	0.953 (0.913-0.987)	0.873 (0.805-0.937)
GNN Naive features	85.0% (78.5-91.1%)	**0.960** (0.924-0.989)	0.838 (0.761-0.908)
GNN Concat_zero features	84.0% (77.2-90.4%)	0.952 (0.911-0.987)	0.830 (0.751-0.901)
GNN ImageNet-ResNet50	70.0% (61.5-78.4%)	0.897 (0.850-0.938)	0.685 (0.589-0.777)

Table 8.4 Hold-out testing results (ensembled across the cross-validation folds) shown as the mean and 95% confidence intervals generated by 10,000 iterations of bootstrapping. The best results are indicated in **bold**.

No single model performed best in all evaluations, though the 10x+20x magnification model was the most consistent, never more than 0.014 behind the best for any given metric. In internal validations, it was slightly outperformed by ABMIL and the naive GNN, but it was best in external validations, with a clear margin in the Transcanadian Study validation (3.6% balanced accuracy, 0.001 AUROC, 0.029 F1 score).

Model	Balanced Accuracy	AUROC	F1 Score
ABMIL 10x only	93.2% (86.5-98.3%)	0.996 (0.988-1.000)	0.912 (0.835-0.974)
GNN Baseline (5x+10x)	94.9% (88.1-100.0%)	0.998 (0.994-1.000)	0.962 (0.910-1.000)
GNN 10x only	93.5% (87.2-98.9%)	0.998 (0.995-1.000)	0.936 (0.863-0.992)
GNN 10x+20x	**99.0%** (96.7-100.0%)	**1.000** (0.999-1.000)	**0.991** (0.971-1.000)
GNN Naive features	92.7% (86.7-98.2%)	0.999 (0.996-1.000)	0.923 (0.845-0.984)
GNN Concat_zero features	95.4% (89.7-100.0%)	0.999 (0.997-1.000)	0.957 (0.899-1.000)
GNN ImageNet-ResNet50	83.7% (73.2-92.9%)	0.983 (0.965-0.996)	0.849 (0.749-0.932)

Table 8.5 Transcanadian Study external validation results (ensembled across the cross-validation folds) shown as the mean and 95% confidence intervals generated by 10,000 iterations of bootstrapping. The best results are indicated in **bold**.

Model	Balanced Accuracy	AUROC	F1 Score
ABMIL 10x only	**77.2%** (73.0-81.4%)	0.954 (0.939-0.966)	0.758 (0.714-0.801)
GNN Baseline (5x+10x)	71.9% (68.0-75.6%)	0.939 (0.923-0.953)	0.725 (0.676-0.770)
GNN 10x only	74.4% (70.7-78.3%)	**0.962** (0.951-0.972)	0.730 (0.682-0.776)
GNN 10x+20x	**77.2%** (73.3-81.1%)	**0.962** (0.950-0.973)	**0.770** (0.724-0.814)
GNN Naive features	75.7% (71.7-79.6%)	0.958 (0.946-0.969)	0.742 (0.694-0.786)
GNN Concat_zero features	72.2% (68.7-76.0%)	0.954 (0.942-0.966)	0.702 (0.653-0.750)
GNN ImageNet-ResNet50	42.7% (39.0-46.6%)	0.860 (0.841-0.879)	0.345 (0.304-0.389)

Table 8.6 OCEAN Challenge external validation results (ensembled across the cross-validation folds) shown as the mean and 95% confidence intervals generated by 10,000 iterations of bootstrapping. The best results are indicated in **bold**.

The 10x+20x magnification GNN had the best average performance across the validations (Table 8.7), with an average balanced accuracy of 84.3%, an average AUROC of 0.962, and an average F1 score of 0.845. This was greater than the average performance of any ABMIL-based model in Chapter 7 (balanced accuracy 83.0%, AUROC 0.965, F1 score 0.822). The confusion matrices for the optimal 10x+20x GNN (Figure 8.4) showed improvements to be fairly evenly spread across the classes when compared to the optimal H-optimus-0 ABMIL model in Chapter 7. Performance remained particularly variable for the least common subtypes, with the F1 scores for LGSC being 0.483 and 0.603 in validations including IDS data, and 0.824 and 1.000 in those without IDS data. The optimal GNN was, however, slightly more consistent at classifying the most common subtype (HGSC), with an F1 score of at least 0.858 in all validations, compared to 0.804 for ABMIL. Overall, the improvements of the graph compared to ABMIL were relatively marginal.

The GNN using the ImageNet-pretrained ResNet50 feature extractor performed worst in every evaluation, with significantly lower performance compared to the baseline GNN in all validations except the Transcanadian Study external validation (Table 8.8). It thus had the lowest averaged performance of 63.3% balanced accuracy, 0.904 AUROC, and 0.611 F1 score. This performance was still greater than that of the ABMIL model using the same features in Chapter 7, which had an average balanced accuracy of 57.1%, AUROC of 0.893, and F1 score of 0.596.

Model	**Balanced Accuracy**	**AUROC**	**F1 Score**
ABMIL 10x only	82.9%	**0.963**	0.820
GNN Baseline (5x+10x)	78.1%	0.953	0.786
GNN 10x only	81.9%	**0.963**	0.819
GNN 10x+20x	**84.3%**	0.962	**0.845**
GNN Naive features	80.9%	0.962	0.805
GNN Concat_zero features	81.5%	0.962	0.812
GNN ImageNet-ResNet50	63.3%	0.904	0.611

Table 8.7 Averaged results across the four validations. The best results are indicated in **bold**.

The effects of modelling with different tissue magnifications varied across validations. The 10x-only GNN typically performed better than the 5x+10x baseline, often with a significant difference (Table 8.8), though it performed slightly worse in the external validation with the Transcanadian Study dataset. The 10x+20x model outperformed the 5x+10x model in all evaluations, though the differences were only statistically significant for the AUROC in hold-out testing, and both the AUROC and balanced accuracy in the OCEAN validation. The 10x+20x model gave a similar performance to the 10x-only model in internal validations but performed much better in external validations.

It was not clear which multi-resolution feature space was best overall. The baseline average-initialised feature space generally performed worst by a small margin, with the naive feature space best in hold-out testing and the OCEAN validation, and the zero-initialised features best in cross-validation and the Transcanadian validation. However, the improvements offered by the naive and zero-initialised features compared to the average-initialised features were not significant in most cases.

Cross-validation

		Predicted Subtype				
		HGSC	**LGSC**	**CCC**	**EC**	**MC**
Actual Subtype	**HGSC**	**1191**	38	18	16	3
	LGSC	44	**43**	4	1	0
	CCC	41	1	**151**	3	2
	EC	18	4	0	**176**	11
	MC	1	0	6	30	**62**

Hold-out Testing

		Predicted Subtype				
		HGSC	**LGSC**	**CCC**	**EC**	**MC**
Actual Subtype	**HGSC**	**19**	0	0	1	0
	LGSC	0	**14**	1	3	2
	CCC	5	0	**15**	0	0
	EC	0	0	0	**20**	0
	MC	0	0	0	0	**20**

External Validation – Transcanadian Study

		Predicted Subtype				
		HGSC	**LGSC**	**CCC**	**EC**	**MC**
Actual Subtype	**HGSC**	**30**	0	0	0	0
	LGSC	0	**9**	0	0	0
	CCC	1	0	**19**	0	0
	EC	0	0	0	**11**	0
	MC	0	0	0	0	**10**

External Validation – OCEAN Challenge

		Predicted Subtype				
		HGSC	**LGSC**	**CCC**	**EC**	**MC**
Actual Subtype	**HGSC**	**202**	1	8	3	3
	LGSC	19	**19**	2	1	1
	CCC	0	0	**93**	0	1
	EC	32	1	6	**67**	13
	MC	1	0	1	1	**38**

Figure 8.4 Confusion matrices for the optimal 10x+20x magnification GNN. Correct classifications are indicated in **bold**.

		p-values		
Validation Set	**Model**	**Balanced Accuracy**	**AUROC**	**F1 Score**
Cross-Validation	ABMIL 10x only	0.078	0.264	0.105
	GNN 10x only	**0.046**	0.756	**0.035**
	GNN 10x+20x	0.093	0.398	0.067
	GNN Naive features	0.226	0.360	0.158
	GNN Concat_zero features	**0.046**	0.264	**0.035**
	GNN ImageNet-ResNet50	**0.050**	**0.049**	**0.035**
Hold-out Testing	ABMIL 10x only	**0.015**	**0.023**	**0.020**
	GNN 10x only	**0.015**	**0.044**	**0.020**
	GNN 10x+20x	0.106	**0.023**	0.113
	GNN Naive features	0.106	**0.017**	0.136
	GNN Concat_zero features	0.106	0.119	0.107
	GNN ImageNet-ResNet50	**0.021**	**0.008**	**0.032**
Transcanadian Study	ABMIL 10x only	0.926	0.312	0.681
	GNN 10x only	0.418	0.464	0.166
	GNN 10x+20x	0.926	0.312	0.681
	GNN Naive features	0.418	0.364	0.166
	GNN Concat_zero features	0.926	0.312	0.681
	GNN ImageNet-ResNet50	0.103	0.272	0.114
OCEAN Challenge	ABMIL 10x only	**0.007**	**0.037**	0.379
	GNN 10x only	0.065	0.053	0.379
	GNN 10x+20x	**0.007**	**0.036**	0.379
	GNN Naive features	0.081	**0.036**	0.858
	GNN Concat_zero features	0.425	0.053	0.858
	GNN ImageNet-ResNet50	**0.007**	**0.036**	**0.005**

Table 8.8 Resulting p-values from paired two-tailed *t*-tests comparing each model to the baseline 5x+10x GNN with UNI features and concat_avg initialisation. These were calculated using the outputs of the five cross-validation models and were adjusted for multiple testing [206]. Those less than 0.050 (before rounding) are indicated in **bold**.

8.4 Discussion

The results indicate that multi-resolution GNN can offer modest improvements to ovarian carcinoma subtyping. In particular, the 10x+20x magnification model achieved near-perfect classification on the Transcanadian Study validation set, giving the greatest reported performance for this task to date [1]. However, multi-resolution GNNs offered only a small benefit over ABMIL overall, with the classification performance only improved in one of the four validations. Considering the relatively small size of the Transcanadian Study set, it was unclear how great a benefit the graph models offered overall. Given the particularly poor performance of the ResNet-based GNN and the relatively strong performance of the UNI-based ABMIL model, it was evident that the chosen feature encoder had a much greater effect on model performance than the subsequent MIL modelling approach.

The results may indicate that any spatial analysis required in ovarian cancer subtyping is sufficiently achieved by applying ABMIL with a transformer-based patch encoder at 10x magnification based on downsampled 1024 x 1024 pixel patches at 40x magnification. The transformer-based approach may capture spatial relationships on a cellular scale within patches, and the ABMIL aggregation may quantify relevant tissue types across the slide. However, the ABMIL approach cannot represent inter-patch (tissue-level) spatial relationships to the same extent as the graph model. Considering the similar performance of these approaches, it appears that tissue-level spatial relationships are not particularly important to ovarian cancer subtyping. Graphs may be more beneficial for other slide classification tasks [90, 99], particularly prognostic tasks in which spatial relationships between tumour, necrosis, and immune cells can be particularly important [85, 264, 268]. However, it remains to be seen whether graphs will still be relevant to these tasks when applied with the drastically improved patch features from histopathology foundation models, and whether the local spatial information encoded within the transformer-based patch encoders could be sufficient within an ABMIL model.

In the external validation on the Transcanadian Study dataset, all models achieved AUROC scores between 0.983 and 1.000, despite the balanced accuracy and F1 scores varying from 83.7%-99.0% and 0.849-0.991, respectively. In internal testing models often also had highly similar AUROC scores but clearly distinct scores by the other metrics. This highlights the limitations of the AUROC for imbalanced multi-class classification, with similarly high scores for all models despite obvious differences in the balanced accuracy and F1 scores, which better represent clinical utility.

As in Chapter 7, performance was greatest in the hold-out test set and the Transcanadian Study external validation set, which were the sets which specifically used primary staging samples without any IDS samples. This indicates that this finding was not unique to ABMIL-based methodologies, and is likely an inherent feature of this type of data. As such, further work should be conducted to understand the variable classification performance that can be achieved on different datasets.

One factor which was not accounted for in this study was the effects of constructing graphs in different configurations - all graphs were constructed such that each patch was connected to its direct neighbours (laterally and diagonally), though this may not be the optimal approach. However, the effect of this was likely mitigated by the tuning of the number of message-passing layers, with more layers having a similar effect to longer connections in the graph. Further, the attention-based methods increased the flexibility of the GNNs by allowing variable connection strengths between tissue patches.

The five-class balanced accuracies of 88%, 99%, and 77% in hold-out and external validations may be sufficient for clinical assistance tools, with these results comparing favourably to the 74-91% concordance of pathologists [32], and so future work should investigate whether pathologists can benefit from the assistance of such tools. However, some limitations remain. The hold-out and Transcanadian Study validations used data from only 30 and 80 patients, respectively, so cannot represent the vast array of variability seen in clinical diagnostic cases. The models are also currently incapable of indicating uncertainty, providing thorough explanations of classification decisions, or coping with tissue which does not contain one of the five most common subtypes of ovarian carcinoma (e.g. non-malignant tissue, carcinosarcomas and non-epithelial malignancies). The large vision transformer and multi-resolution graphs also carry a heavy computational burden, which is likely to lead to logistical difficulties in deploying such models in the clinical setting. None of these issues are insurmountable, and when they are overcome, these models could be invaluable as diagnostic assistive tools offering a rapid second opinion to pathologists.

8.5 Conclusion

Overall, we have shown that a multi-resolution GNN can slightly improve the accuracy of ovarian carcinoma subtyping at the whole-slide level above the previous state-of-the-art, though the benefit was not present in all validations. In an external validation of 80 WSIs, a GNN achieved a near-perfect 99% balanced accuracy, but in internal hold-out testing this was 88%, and in another external validation only 77%, no greater than ABMIL performance. The best GNN combined 10x and 20x magnification data, which was better than combining lower magnifications or using only 10x magnification data, though at an increased computational cost. While the highly accurate graph models may offer a useful second opinion to pathologists, more extensive validations are required to understand the reasons underlying performance variability across different datasets and to improve model consistency.

Chapter 9

Conclusions and Future Work

In this chapter, we provide an overview of the thesis, including our contributions and key findings, the limitations of the research, and potential ideas for future research.

9.1 Thesis Summary

This thesis has focused on the development and thorough validation of an AI pipeline for the classification of ovarian carcinoma subtypes at the whole slide level.

Chapter 1 introduced the thesis and set out the aims and objectives, which were to be achieved by systematically reviewing relevant literature, applying state-of-the-art approaches with a world-leading ovarian cancer dataset, building upon previous techniques with novel classification approaches, and rigorously validating the performance of the resulting classification pipelines. Finally, the overall structure of the thesis was laid out, with the subsequent six chapters addressing the thesis objectives.

Chapter 2 introduced ovarian cancer and the current clinical problems faced in the pathological diagnosis of ovarian carcinoma subtypes. It also gave context to the current state of digitisation and AI utilisation in histopathology, and described how an automated subtype classification pipeline could potentially improve the efficiency, accuracy, and objectivity of diagnosis. It also provided the technical background for the thesis, in particular describing the computer vision methods that are applied to digital pathology images.

Chapter 3 provided an in-depth analysis of previous AI research for the diagnosis and prognosis of ovarian cancer from histopathology slides. This was underpinned by a systematic literature review and brought up-to-date with recently repeated searches. The risks of bias in previous research were assessed, and recommendations were provided to reduce these risks and improve the clinical viability of future research.

Chapter 4 was a methodological chapter, describing the AI model development and validation methods used throughout the rest of the thesis. The concept of multiple instance learning was described in terms of the preprocessing, patching, embedding, and aggregation steps used, and the baseline ABMIL classification model was introduced. Approaches for ensuring rigorous validations were explored, including hyperparameter tuning, different classification metrics, and hypothesis testing. Breakdowns were provided for three ovarian carcinoma subtyping datasets, with a world-leading internal dataset and two external validation datasets used. Finally, the software and hardware used to create and test models were described.

Chapter 5 proposed an approach to improve the efficiency of slide-level classification by leveraging the patch attention scores of ABMIL to create an iterative active patch sampling approach for use during inference. This utilised the inherent spatial relationships within WSIs, with diagnostically relevant tissue patches often forming spatial clusters. Sampling drastically reduced the proportion of the total tissue area that was fully analysed, aiming to reduce the computational workload of classification.

Chapter 6 thoroughly analysed the performance of the standard ABMIL classifier with six different tissue magnifications from the clinical standard 40x down to 1.25x. This investigated the trade-off between the cellular-level detail at higher magnifications and the greater tissue-level context at lower magnifications. It also included analysing the efficiency of model training and slide inference at different magnifications.

Chapter 7 thoroughly analysed different feature extraction techniques in an ABMIL classifier, with a focus on comparing the newly available histopathology foundation models to traditional ImageNet-pretrained feature extractors. This included an exploration as to whether the ImageNet-pretrained ResNet50 model could be made competitive with the newer approaches through varied preprocessing techniques such as normalisation and augmentation. It also included an ablation study into the effects of hyperparameter tuning on downstream classification.

Chapter 8 proposed a novel multi-resolution graph MIL network, utilising the spatial relationships between patches in a pathology slide to improve classification performance. The graph model was compared in six configurations, using different tissue magnifications and multi-magnification feature modelling approaches. The effects of foundation models were further analysed in relation to this different classification approach.

9.2 Key Contributions and Findings

The key contributions of this thesis were the systematic review of previous literature in ovarian cancer histopathology, the development of novel WSI pipelines for the classification of ovarian carcinoma subtypes, and the rigorous analysis and validation of these subtyping models. The chapters, presented in chronological order, detailed the process of interpreting previous research, applying state-of-the-art histopathology models, analysing the classification performance and efficiency of these approaches in varied configurations, and finally, creating and thoroughly validating novel classifiers based on all of the previously learned lessons. The main findings of this research are as follows.

In Chapter 3, it was found that previous research had been conducted to investigate the utility of AI for a wide array of diagnostic and prognostic tasks in ovarian cancer histopathology, with subtyping being one of the most common. Key limitations were identified regarding the datasets, validations, and reporting in previous studies. The sparsity of available ovarian cancer datasets was a common issue, with few researchers able to assemble large enough datasets to thoroughly train and validate models. This was often compounded by methodological flaws, with studies conducted without cross-validation, external validation, bootstrapping, hyperparameter tuning, or statistical analyses. No study achieved an overall low risk of bias score, with the most promising papers only achieving an *unclear* risk of bias due to incomplete reporting.

In Chapter 5, it was found that the proposed active patch sampling method during inference gave a similar classification performance to the standard ABMIL approach, but with a drastically reduced computational burden given the reduced proportion of tissue being fully processed by the classifier. For binary classification of HGSC using the earliest version of the internal LTHT dataset, the baseline ABMIL classifier achieved an 80.1% balanced accuracy and 0.878 AUROC, while the sampling approach using only 5% of the available tissue patches achieved 79.1% and 0.868, respectively. This small reduction in classification performance allowed inference time to be reduced by up to 86%.

In Chapter 6, it was found that the 5x and 10x magnifications gave the best overall classification performance in the ABMIL classifier, with these also drastically reducing the computational requirements of model training and slide inference when compared to higher magnifications. In five-class hold-out testing, the optimal 10x model achieved 62.0% balanced accuracy and 0.850 AUROC, while reducing training time by 94% and inference time by 70% compared to the clinical standard 40x magnification.

In Chapter 7, it was found that histopathology foundation models drastically improved subtype classification performance compared to ImageNet-pretrained feature extractors, though at an increased computational cost. Where the baseline ImageNet-pretrained ResNet50 model gave balanced accuracies of 66.0%, 69.2%, and 52.4% in hold-out testing and two external validations, the optimal foundation model, H-optimus-0, achieved 89.0%, 96.7%, and 74.0%. Further, the UNI foundation model achieved similar performance to H-optimus-0 at a quarter of the computational cost. It was found that hyperparameter tuning was beneficial to classification performance even when employing the greatest feature extractors, with a median improvement of 1.9% balanced accuracy attained. The foundation model-based classifiers were the first models that were accurate enough to potentially compete with real pathologists, and thus, these may be able to aid pathologists in diagnostic decision-making.

In Chapter 8, it was found that graph networks gave modest classification improvements over ABMIL, specifically when combining 10x and 20x magnification data, though the benefit was variable and much smaller than the benefit given by using a foundation model rather than an ImageNet-pretrained encoder. Where the ABMIL model achieved balanced accuracies of 88.0%, 93.2%, and 77.2%, the optimal graph model achieved 88.0%, 99.0%, and 77.2%, only improving performance on one of the three validation sets. Given the graph networks had much greater computational costs than ABMIL, and the only improvement was found on the smallest test set, it was not clear that this benefit was worthwhile.

9.3 Limitations and Further Work

The eventual goal of this research is to create a clinically implementable assistive tool for pathologists, but there are several factors currently preventing this from being attained. These factors could all be addressed in future work, with the scope of potential work being that of several additional PhD theses.

While we conducted analyses with the largest ovarian cancer subtyping dataset to date, this was still not sufficient to ensure that classifiers were robust to all relevant sources of variation in histopathology data. The hold-out and Transcanadian Study external validation set were composed of only 80 and 100 WSIs, respectively, and the UBC-OCEAN dataset appeared to be of mixed diagnostic quality, without sufficient metadata provided to fully understand the resulting classification variability. Ideally, subtyping models would be made robust to lower quality data to reduce the burden on quality control in the lab, though if this goal proves unattainable, greater automated quality control measures will be required before subtyping models are applied.

The models reported in this thesis are limited to the slide-level classification of the five most common histological subtypes of ovarian carcinoma. Such models have no understanding of rare carcinoma subtypes, non-carcinoma ovarian cancers, mixed subtypes, or even non-ovarian cancer tissue, yet all of these would be classified as one of the five most common ovarian carcinoma subtypes. Future work may seek to collect data for the rarer subtypes, though given their rarity, it may never be possible to collect a sufficient quantity to attain high classification performance. As such, future work may instead seek to quantify the uncertainty in the classification predictions. It may be expected that any input data that does not match one of the common subtypes would be classified with a high level of uncertainty, and such cases could be prioritized for manual analysis by pathologists. Along with automated quality control methods, uncertainty quantification could be seen as a guardrail to aid the pathologists in safely using the AI models.

The analyses presented have been limited to resection specimens, with the best performance found on the diagnostically preferable primary resection specimens rather than IDS samples (Section 2.2). In many ovarian carcinoma cases, it is not possible to analyse a primary resection specimen, and as such it would improve the applicability of the models if they could be accurately applied to pre-treatment biopsies or IDS specimens. Achieving a high accuracy on these samples may not be possible due to the drastically reduced tissue quantity in biopsies and the degraded tissue quality in IDS specimens, though if a model could demonstrate comparable performance to an expert pathologist it may still be clinically beneficial in cases of imperfect information.

The methods used in this study are not easily interpretable, with the black-box models only interrogated in this study using attention heatmaps, which provide very limited insight into model decision-making. Truly explainable approaches would require the use of interpretable features, which could be achieved by the use of hand-crafted features or by extensive interrogations of the automated features extracted by histopathology foundation models. A better understanding of model decision-making would make it clearer whether models should be trusted, drastically improving their utility.

One of the key problems with translating these models into the clinic is that it is unclear how to best present the class predictions generated by these models to the pathologists. Usability studies could be key in uncovering key principles for presenting automated inferences to pathologists. Such studies could include investigating whether pathologists benefit from supplementary information, such as uncertainty scores and attention heatmaps, to understand whether this improves diagnostic accuracy or whether it simply distracts the pathologist. These investigations may also help to determine whether pathologists would require extra training to safely leverage AI-generated insights. Understanding the human elements of diagnostic AI assistance will help to maximise the benefits of these technologies.

The best-performing models in this thesis were also some of the most computationally expensive, which drastically limits the real-world utility of these models to pathology departments that can afford expensive computational infrastructure and those that are willing to export their data off-site. To broaden access to these models it will be essential to reduce their computational burden. This will likely require the combination of many efficiency gains, potentially including improved active sampling approaches and using the lowest viable tissue magnifications, but it will also require techniques not covered in this thesis, such as pruning, mixed-precision modelling, and knowledge distillation. It also may be possible to extend these methods to smaller tissue samples and lower-quality images, potentially even making it feasible that a pathologist could take a photograph through the microscope for computational analysis.

9.4 Closing Remarks

It is absolutely clear that significant progress has been made in AI for ovarian cancer subtyping (and for pathological diagnosis more widely) in recent years. This thesis has highlighted that modern classifiers (especially those built using histopathology foundation models) can accurately classify ovarian cancer subtypes to the extent that clinical trials may now be considered, and it has also shown that there are viable techniques to reduce the computational burden of such classifiers. It now seems inevitable that these technologies will achieve clinical utility in the coming years. While many questions remain unanswered, it is notable that some AI technologies are starting to receive regulatory approval, and many new companies are entering the field of digital pathology. To ensure the equitable deployment of these models, further work will be needed to broaden access to digital pathology services and to reduce the computational burden of the models. If successful, such tools may mitigate the worldwide shortage of pathologists and improve diagnostic accuracy to optimise the delivery of precision medicine.

References

[1] Breen J, Allen K, Zucker K, Adusumilli P, Scarsbrook A, Hall G, et al. Artificial intelligence in ovarian cancer histopathology: a systematic review. NPJ Precision Oncology. 2023;7(1):83.

[2] Breen J, Allen K, Zucker K, Godson L, Orsi NM, Ravikumar N. A Comprehensive Evaluation of Histopathology Foundation Models for Ovarian Cancer Subtype Classification. NPJ Precision Oncology. 2025;9:33.

[3] Aubreville M, Stathonikos N, Bertram CA, Klopfleisch R, Ter Hoeve N, Ciompi F, et al. Mitosis domain generalization in histopathology images—the MIDOG challenge. Medical Image Analysis. 2023;84:102699.

[4] Breen J, Zucker K, Orsi NM, Ravikumar N. Assessing domain adaptation techniques for mitosis detection in multi-scanner breast cancer histopathology images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 14-22.

[5] Breen J, Allen K, Zucker K, Hall G, Orsi NM, Ravikumar N. Efficient subtyping of ovarian cancer histopathology whole slide images using active sampling in multiple instance learning. In: Proceedings of SPIE 12471. vol. 12471. SPIE; 2023. p. 248-58.

[6] Breen J, Ravikumar N, Allen K, Zucker K, Orsi NM. Reducing Histopathology Slide Magnification Improves the Accuracy and Speed of Ovarian Cancer Subtyping. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI); 2024. p. 1-5.

[7] Breen J, Allen K, Zucker K, Orsi NM, Ravikumar N. Multi-Resolution Histopathology Patch Graphs for Ovarian Cancer Subtyping. arXiv preprint arXiv:240718105. 2024.

[8] Allen KE, Breen J, Hall G, Zucker K, Ravikumar N, Orsi NM. Comparative evaluation of ovarian carcinoma subtyping in primary versus interval debulking surgery specimen whole slide images using artificial intelligence. BMJ Specialist Journals; 2023.

[9] Breen J, Zucker K, Allen K, Ravikumar N, Orsi NM. Generative Adversarial Networks for Stain Normalisation in Histopathology. In: Applications of Generative AI. Springer; 2024. p. 227-47.

[10] Allen KE, Adusumilli P, Breen J, Hall G, Orsi NM. Artificial Intelligence in Ovarian Digital Pathology. In: Pathology of the Ovary, Fallopian Tube and Peritoneum. Springer; 2024. p. 731-49.

[11] Breen J, Allen K, Zucker K, Hall G, Ravikumar N, Orsi NM. Predicting Ovarian Cancer Treatment Response in Histopathology using Hierarchical Vision Transformers and Multiple Instance Learning. arXiv preprint arXiv:231012866. 2023.

[12] Eisenmann M, Reinke A, Weru V, Tizabi MD, Isensee F, Adler TJ, et al. Biomedical image analysis competitions: The state of current participation practice. arXiv preprint arXiv:221208568. 2022.

[13] Kyo S, Ishikawa N, Nakamura K, Nakayama K. The fallopian tube as origin of ovarian cancer: Change of diagnostic and preventive strategies. Cancer medicine. 2020;9(2):421-31.

[14] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2024;74(3):229-63.

[15] Ebell MH, Culp MB, Radke TJ. A systematic review of symptoms for the diagnosis of ovarian cancer. American journal of preventive medicine. 2016;50(3):384-94.

[16] Cancer Research UK. Ovarian Cancer Incidence Statistics (United Kingdom); 2023. Accessed: 2023-12-08. Available from: www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/ovarian-cancer/incidence.

[17] Menon U, Gentry-Maharaj A, Burnell M, Singh N, Ryan A, Karpinskyj C, et al. Ovarian cancer population screening and mortality after long-term follow-up

in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. The Lancet. 2021;397(10290):2182-93.

[18] Cancer Research UK. Ovarian cancer survival statistics; 2024. Accessed: 2024-07-22. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/ovarian-cancer/survival.

[19] Gaitskell K, Hermon C, Barnes I, Pirie K, Floud S, Green J, et al. Ovarian cancer survival by stage, histotype, and pre-diagnostic lifestyle factors, in the prospective UK Million Women Study. Cancer epidemiology. 2022;76:102074.

[20] Berek JS, Renz M, Kehoe S, Kumar L, Friedlander M. Cancer of the ovary, fallopian tube, and peritoneum: 2021 update. International Journal of Gynecology & Obstetrics. 2021;155:61-85.

[21] Falzone L, Scandurra G, Lombardo V, Gattuso G, Lavoro A, Distefano AB, et al. A multidisciplinary approach remains the best strategy to improve and strengthen the management of ovarian cancer. International Journal of Oncology. 2021;59(1):1-14.

[22] Kuroki L, Guntupalli SR. Treatment of epithelial ovarian cancer. Bmj. 2020;371.

[23] Ledermann J, Harter P, Gourley C, Friedlander M, Vergote I, Rustin G, et al. Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: a preplanned retrospective analysis of outcomes by BRCA status in a randomised phase 2 trial. The lancet oncology. 2014;15(8):852-61.

[24] Mirza MR, Monk BJ, Herrstedt J, Oza AM, Mahner S, Redondo A, et al. Niraparib maintenance therapy in platinum-sensitive, recurrent ovarian cancer. New England Journal of Medicine. 2016;375(22):2154-64.

[25] Ray-Coquard I, Pautier P, Pignata S, Pérol D, González-Martín A, Berger R, et al. Olaparib plus bevacizumab as first-line maintenance in ovarian cancer. New England Journal of Medicine. 2019;381(25):2416-28.

[26] Moch H. Female genital tumours: WHO Classification of Tumours, Volume 4. WHO Classification of Tumours. 2020;4.

[27] Prat J, on Gynecologic Oncology FC, et al. Staging classification for cancer of the ovary, fallopian tube, and peritoneum. International Journal of Gynecology & Obstetrics. 2014;124(1):1-5.

[28] Vroobel K. Overview of Ovarian Tumours: Pathogenesis and General Considerations. In: Pathology of the Ovary, Fallopian Tube and Peritoneum. Springer; 2024. p. 95-113.

[29] Köbel M, Kalloger SE, Boyd N, McKinney S, Mehl E, Palmer C, et al. Ovarian carcinoma subtypes are different diseases: implications for biomarker studies. PLoS medicine. 2008;5(12):e232.

[30] Kossaï M, Leary A, Scoazec JY, Genestie C. Ovarian cancer: a heterogeneous disease. Pathobiology. 2018;85(1-2):41-9.

[31] Matsuno RK, Sherman ME, Visvanathan K, Goodman MT, Hernandez BY, Lynch CF, et al. Agreement for tumor grade of ovarian carcinoma: analysis of archival tissues from the surveillance, epidemiology, and end results residual tissue repository. Cancer causes & control. 2013;24:749-57.

[32] Köbel M, Bak J, Bertelsen BI, Carpen O, Grove A, Hansen ES, et al. Ovarian carcinoma histotype determination is highly reproducible, and is improved through the use of immunohistochemistry. Histopathology. 2014;64(7):1004-13.

[33] Barnard ME, Pyden A, Rice MS, Linares M, Tworoger SS, Howitt BE, et al. Interpathologist and pathology report agreement for ovarian tumor characteristics in the Nurses' Health Studies. Gynecologic oncology. 2018;150(3):521-6.

[34] Goode EL, Block MS, Kalli KR, Vierkant RA, Chen W, Fogarty ZC, et al. Dose-response association of CD8+ tumor-infiltrating lymphocytes and survival time in high-grade serous ovarian cancer. JAMA oncology. 2017;3(12):e173290-0.

[35] Sieh W, Köbel M, Longacre TA, Bowtell DD, DeFazio A, Goodman MT, et al. Hormone-receptor expression and ovarian cancer survival: an Ovarian Tumor Tissue Analysis consortium study. The lancet oncology. 2013;14(9):853-62.

[36] Royal College of Pathologists. Meeting pathology demand: Histopathology workforce census; 2018. Accessed: 2024-09-25. Available

from: https://www.rcpath.org/static/952a934d-2ec3-48c9-a8e6e00fcdca700f/Meeting-Pathology-Demand-Histopathology-Workforce-Census-2018.pdf.

[37] Walsh E, Orsi NM. The current troubled state of the global pathology workforce: a concise review. Diagnostic Pathology. 2024;19(1):163.

[38] Wilson ML, Fleming KA, Kuti MA, Looi LM, Lago N, Ru K. Access to pathology and laboratory medicine services: a crucial gap. The Lancet. 2018;391(10133):1927-38.

[39] George J, Gkousis E, Feast A, Morris S, Pollard S, Vohra J. Estimating the cost of growing the NHS cancer workforce in England by 2029. Cancer Research UK. 2020. Accessed: 2024-09-25. Available from: https://www.cancerresearchuk.org/sites/default/files/estimating_the_cost_of_growing_the_nhs_cancer_workforce_in_england_by_2029_october_2020_-_full_report.pdf.

[40] Hanna TP, King WD, Thibodeau S, Jalink M, Paulin GA, Harvey-Jones E, et al. Mortality due to cancer treatment delay: systematic review and meta-analysis. bmj. 2020;371.

[41] Baidoshvili A, Bucur A, van Leeuwen J, van der Laak J, Kluin P, van Diest PJ. Evaluating the benefits of digital pathology implementation: time savings in laboratory logistics. Histopathology. 2018;73(5):784-94.

[42] Hanna MG, Reuter VE, Samboy J, England C, Corsale L, Fine SW, et al. Implementation of digital pathology offers clinical and operational increase in efficiency and cost savings. Archives of pathology & laboratory medicine. 2019;143(12):1545-55.

[43] Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). The American journal of surgical pathology. 2018;42(1):39.

[44] Azam AS, Tsang YW, Thirlwall J, Kimani PK, Sah S, Gopalakrishnan K, et al. Digital pathology for reporting histopathology samples, including cancer screening samples–definitive evidence from a multisite study. Histopathology. 2024;84(5):847-62.

[45] Pinto DG, Bychkov A, Tsuyama N, Fukuoka J, Eloy C. Exploring the adoption of digital pathology in clinical settings-Insights from a cross-continent study. medRxiv. 2023:2023-04.

[46] Ahmed AA, Abouzid M, Kaczmarek E. Deep Learning Approaches in Histopathology. Cancers. 2022;14(21):5264.

[47] Zimmermann E, Vorontsov E, Viret J, Casson A, Zelechowski M, Shaikovski G, et al. Virchow 2: Scaling Self-Supervised Mixed Magnification Models in Pathology. arXiv preprint arXiv:240800738. 2024.

[48] Lyon HO, De Leenheer A, Horobin R, Lambert W, Schulte E, Van Liedekerke B, et al. Standardization of reagents and methods used in cytological and histological practice with emphasis on dyes, stains and chromogenic reagents. The Histochemical Journal. 1994;26:533-44.

[49] Rajaganesan S, Kumar R, Rao V, Pai T, Mittal N, Sahay A, et al. Comparative assessment of digital pathology systems for primary diagnosis. Journal of Pathology Informatics. 2021;12(1):25.

[50] Aubreville M, Stathonikos N, Donovan TA, Klopfleisch R, Ammeling J, Ganz J, et al. Domain generalization across tumor types, laboratories, and species—Insights from the 2022 edition of the Mitosis Domain Generalization Challenge. Medical Image Analysis. 2024;94:103155.

[51] da Silva LM, Pereira EM, Salles PG, Godrich R, Ceballos R, Kunz JD, et al. Independent real-world application of a clinical-grade automated prostate cancer detection system. The Journal of pathology. 2021;254(2):147-58.

[52] Matthews GA, McGenity C, Bansal D, Treanor D. Public evidence on AI products for digital pathology. medRxiv. 2024:2024-02.

[53] Raciti P, Sue J, Retamero JA, Ceballos R, Godrich R, Kunz JD, et al. Clinical validation of artificial intelligence–augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. Archives of Pathology & Laboratory Medicine. 2023;147(10):1178-85.

[54] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-6.

[55] O'Mahony N, Campbell S, Carvalho A, Harapanahalli S, Hernandez GV, Krpalkova L, et al. Deep learning vs. traditional computer vision. In: Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC). vol. 1. Springer; 2020. p. 128-44.

[56] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;86(11):2278-324.

[57] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.

[58] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015;115:211-52.

[59] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.

[60] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing. 2021;452:48-62.

[61] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[62] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.

[63] Jelassi S, Sander M, Li Y. Vision transformers provably learn spatial structure. Advances in Neural Information Processing Systems. 2022;35:37822-36.

[64] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: A survey. Medical Image Analysis. 2023;88:102802.

[65] Dehghani M, Djolonga J, Mustafa B, Padlewski P, Heek J, Gilmer J, et al. Scaling vision transformers to 22 billion parameters. In: International Conference on Machine Learning. PMLR; 2023. p. 7480-512.

[66] Awais M, Naseer M, Khan S, Anwer RM, Cholakkal H, Shah M, et al. Foundational models defining a new era in vision: A survey and outlook. arXiv preprint arXiv:230713721. 2023.

[67] Zhou C, Li Q, Li C, Yu J, Liu Y, Wang G, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. International Journal of Machine Learning and Cybernetics. 2024:1-65.

[68] Song AH, Jaume G, Williamson DF, Lu MY, Vaidya A, Miller TR, et al. Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering. 2023;1(12):930-49.

[69] Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence. 1997;89(1-2):31-71.

[70] Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2424-33.

[71] Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems. 2021;34:2136-47.

[72] Carbonneau MA, Cheplygina V, Granger E, Gagnon G. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition. 2018;77:329-53.

[73] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning. Advances in neural information processing systems. 2002;15.

[74] Ramon J, De Raedt L. Multi instance neural networks. In: Proceedings of the ICML-2000 workshop on attribute-value and relational learning; 2000. p. 53-60.

[75] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine. 2019;25(8):1301-9.

[76] Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. Scientific reports. 2020;10(1):1504.

[77] Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: International conference on machine learning. PMLR; 2018. p. 2127-36.

[78] Lu MY, Williamson DF, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering. 2021;5(6):555-70.

[79] Guo Z, Zhao W, Wang S, Yu L. HIGT: Hierarchical Interaction Graph-Transformer for Whole Slide Image Analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2023. p. 755-64.

[80] Hou W, He Y, Yao B, Yu L, Yu R, Gao F, et al. Multi-scope Analysis Driven Hierarchical Graph Transformer for Whole Slide Image Based Cancer Survival Prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2023. p. 745-54.

[81] Godson L, Alemi N, Nsengimana J, Cook GP, Clarke EL, Treanor D, et al. Immune subtyping of melanoma whole slide images using multiple instance learning. Medical Image Analysis. 2024;93:103097.

[82] Mirabadi AK, Archibald G, Darbandsari A, Contreras-Sanz A, Nakhli RE, Asadi M, et al. GRASP: GRAph-Structured Pyramidal Whole Slide Image Representation. arXiv preprint arXiv:240203592. 2024.

[83] Chen RJ, Chen C, Li Y, Chen TY, Trister AD, Krishnan RG, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 16144-55.

[84] Gao Z, Shi J, Wang J. GQ-GCN: Group quadratic graph convolutional network for classification of histopathological images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. Springer; 2021. p. 121-31.

[85] Lu C, Koyuncu C, Corredor G, Prasanna P, Leo P, Wang X, et al. Feature-driven local cell graph (FLocK): new computational pathology-based descriptors for prognosis of lung cancer and HPV status of oropharyngeal cancers. Medical image analysis. 2021;68:101903.

[86] Wang Z, Li J, Pan Z, Li W, Sisk A, Ye H, et al. Hierarchical graph pathomic network for progression free survival prediction. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. Springer; 2021. p. 227-37.

[87] Ahmedt-Aristizabal D, Armin MA, Denman S, Fookes C, Petersson L. A survey on graph-based deep learning for computational histopathology. Computerized Medical Imaging and Graphics. 2022;95:102027.

[88] Bai Y, Mi Y, Su Y, Zhang B, Zhang Z, Wu J, et al. A scalable graph-based framework for multi-organ histology image classification. IEEE Journal of Biomedical and Health Informatics. 2022;26(11):5506-17.

[89] Hou W, Huang H, Peng Q, Yu R, Yu L, Wang L. Spatial-hierarchical graph neural network with dynamic structure learning for histological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2022. p. 181-91.

[90] Pati P, Jaume G, Foncubierta-Rodriguez A, Feroce F, Anniciello AM, Scognamiglio G, et al. Hierarchical graph representations in digital pathology. Medical image analysis. 2022;75:102264.

[91] Sims J, Grabsch HI, Magee D. Using Hierarchically Connected Nodes and Multiple GNN Message Passing Steps to Increase the Contextual Information in Cell-Graph Classification. In: MICCAI Workshop on Imaging Systems for GI Endoscopy. Springer; 2022. p. 99-107.

[92] Abbas SF, Le Vuong TT, Kim K, Song B, Kwak JT. Multi-cell type and multi-level graph aggregation network for cancer grading in pathology images. Medical Image Analysis. 2023;90:102936.

[93] Krishna A, Gupta RK, Kurian NC, Jeevan P, Sethi A. Heterogeneous Graphs Model Spatial Relationship Between Biological Entities for Breast Cancer Diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2023. p. 97-106.

[94] Wang H, Huang G, Zhao Z, Cheng L, Juncker-Jensen A, Nagy ML, et al. Ccf-gnn: A unified model aggregating appearance, microenvironment, and topology for pathology image classification. IEEE Transactions on Medical Imaging. 2023;42(11):3179-93.

[95] Vanea C, Campbell J, Dodi O, Salumäe L, Meir K, Hochner D, et al. A New Graph Node Classification Benchmark: Learning Structure from Histology Cell Graphs. In: NeurIPS 2022 Workshop: New Frontiers in Graph Learning; 2022. p. 1-17.

[96] Vanea C, Džigurski J, Rukins V, Dodi O, Siigur S, Salumäe L, et al. Mapping cell-to-tissue graphs across human placenta histology whole slide images using deep learning with HAPPY. Nature Communications. 2024;15(1):2710.

[97] Alzoubi I, Zhang L, Zheng Y, Loh C, Wang X, Graeber MB. PathoGraph: An Attention-Based Graph Neural Network Capable of Prognostication Based on CD276 Labelling of Malignant Glioma Cells. Cancers. 2024;16(4):750.

[98] Yang Z, Qiu Z, Lin T, Chao H, Chang W, Yang Y, et al. From Histopathology Images to Cell Clouds: Learning Slide Representations with Hierarchical Cell Transformer. arXiv preprint arXiv:241216715. 2024.

[99] Lu W, Toss M, Dawood M, Rakha E, Rajpoot N, Minhas F. SlideGraph+: Whole slide image level graphs to predict HER2 status in breast cancer. Medical Image Analysis. 2022;80:102486.

[100] Pati P, Jaume G, Ayadi Z, Thandiackal K, Bozorgtabar B, Gabrani M, et al. Weakly supervised joint whole-slide segmentation and classification in prostate cancer. Medical Image Analysis. 2023;89:102915.

[101] Brussee S, Buzzanca G, Schrader AM, Kers J. Graph neural networks in histopathology: Emerging trends and future directions. Medical Image Analysis. 2025:103444.

[102] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE transactions on neural networks. 2008;20(1):61-80.

[103] Zhou ZH, Sun YY, Li YF. Multi-instance learning by treating instances as non-iid samples. In: Proceedings of the 26th annual international conference on machine learning; 2009. p. 1249-56.

[104] Akazawa M, Hashimoto K. Artificial intelligence in gynecologic cancers: Current status and future challenges–A systematic review. Artificial Intelligence in Medicine. 2021;120:102164.

[105] Fiste O, Liontos M, Zagouri F, Stamatakos G, Dimopoulos MA. Machine Learning applications in gynecological cancer: a critical review. Critical Reviews in Oncology/Hematology. 2022:103808.

[106] Xu HL, Gong TT, Liu FH, Chen HY, Xiao Q, Hou Y, et al. Artificial intelligence performance in image-based ovarian cancer identification: A systematic review and meta-analysis. EClinicalMedicine. 2022;53.

[107] Zhou J, Cao W, Wang L, Pan Z, Fu Y. Application of artificial intelligence in the diagnosis and prognostic prediction of ovarian cancer. Computers in Biology and Medicine. 2022;146:105608.

[108] Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC medical research methodology. 2022;22(1):1-16.

[109] Dong J, Li J, Lu J, Fu A. Automatic segmentation for ovarian cancer immunohistochemical image based on chroma criterion. In: 2010 2nd International Conference on Advanced Computer Control. vol. 2. IEEE; 2010. p. 147-50.

[110] Dong J, Li J, Fu A, Lv H. Automatic segmentation for ovarian cancer immunohistochemical image based on yuv color space. In: 2010 International Conference on Biomedical Engineering and Computer Science. IEEE; 2010. p. 1-4.

[111] Signolle N, Revenu M, Plancoulaine B, Herlin P. Wavelet-based multiscale texture segmentation: Application to stromal compartment characterization on virtual slides. Signal Processing. 2010;90(8):2412-22.

[112] Janowczyk A, Chandran S, Feldman M, Madabhushi A. Local morphologic scale: application to segmenting tumor infiltrating lymphocytes in ovarian cancer TMAs. In: Medical Imaging 2011: Image Processing. vol. 7962. SPIE; 2011. p. 827-40.

[113] Janowczyk A, Chandran S, Singh R, Sasaroli D, Coukos G, Feldman MD, et al. High-throughput biomarker segmentation on ovarian cancer tissue microarrays via hierarchical normalized cuts. IEEE transactions on biomedical engineering. 2012;59(5):1240-52.

[114] Kothari S, Phan JH, Osunkoya AO, Wang MD. Biological interpretation of morphological patterns in histopathological whole-slide images. In: Proceedings of the ACM conference on bioinformatics, computational biology and biomedicine; 2012. p. 218-25.

[115] Poruthoor A, Phan JH, Kothari S, Wang MD. Exploration of genomic, proteomic, and histopathological image data integration methods for clinical prediction. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing. IEEE; 2013. p. 259-63.

[116] BenTaieb A, Li-Chang H, Huntsman D, Hamarneh G. Automatic diagnosis of ovarian carcinomas via sparse multiresolution tissue representation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18. Springer; 2015. p. 629-36.

[117] BenTaieb A, Nosrati MS, Li-Chang H, Huntsman D, Hamarneh G. Clinically-inspired automatic classification of ovarian carcinoma subtypes. Journal of pathology informatics. 2016;7(1):28.

[118] BenTaieb A, Li-Chang H, Huntsman D, Hamarneh G. A structured latent model for ovarian carcinoma subtyping from histopathology slides. Medical image analysis. 2017;39:194-205.

[119] Lorsakul A, Andersson E, Harring SV, Sade H, Grimm O, Bredno J. Automated wholeslide analysis of multiplex-brightfield IHC images for cancer cells and carcinoma-associated fibroblasts. In: Medical Imaging 2017: Digital Pathology. vol. 10140. SPIE; 2017. p. 41-6.

[120] Du Y, Zhang R, Zargari A, Thai TC, Gunderson CC, Moxley KM, et al. Classification of tumor epithelium and stroma by exploiting image features learned by deep convolutional neural networks. Annals of biomedical engineering. 2018;46:1988-99.

[121] Heindl A, Khan AM, Rodrigues DN, Eason K, Sadanandam A, Orbegoso C, et al. Microenvironmental niche divergence shapes BRCA1-dysregulated ovarian cancer morphological plasticity. Nature communications. 2018;9(1):3917.

[122] Kalra S, Tizhoosh HR, Shah S, Choi C, Damaskinos S, Safarpoor A, et al. Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. NPJ digital medicine. 2020;3(1):31.

[123] Levine AB, Peng J, Farnell D, Nursey M, Wang Y, Naso JR, et al. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. The Journal of pathology. 2020;252(2):178-88.

[124] Yaar A, Asif A, Raza SEA, Rajpoot N, Minhas F. Cross-domain knowledge transfer for prediction of chemosensitivity in ovarian cancer patients. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020. p. 928-9.

[125] Yu KH, Hu V, Wang F, Matulonis UA, Mutter GL, Golden JA, et al. Deciphering serous ovarian carcinoma histopathology and platinum response by convolutional neural networks. BMC medicine. 2020;18(1):1-14.

[126] Gentles L, Howarth R, Lee WJ, Sharma-Saha S, Ralte A, Curtin N, et al. Integration of computer-aided automated analysis algorithms in the development and validation of immunohistochemistry biomarkers in ovarian cancer. Journal of Clinical Pathology. 2021;74(7):469-74.

[127] Ghoniem RM, Algarni AD, Refky B, Ewees AA. Multi-modal evolutionary deep learning model for ovarian cancer diagnosis. Symmetry. 2021;13(4):643.

[128] Jiang J, Tekin B, Guo R, Liu H, Huang Y, Wang C. Digital pathology-based study of cell-and tissue-level morphologic features in serous borderline ovarian tumor and high-grade serous ovarian cancer. Journal of Pathology Informatics. 2021;12(1):24.

[129] Laury AR, Blom S, Ropponen T, Virtanen A, Carpén OM. Artificial intelligence-based image analysis can predict outcome in high-grade serous carcinoma via histology alone. Scientific Reports. 2021;11(1):19165.

[130] Paijens S, Vledder A, Loiero D, Duiker E, Bart J, Hendriks A, et al. Prognostic image-based quantification of CD8CD103 T cell subsets in high-grade serous ovarian cancer patients. Oncoimmunology. 2021;10(1):1935104.

[131] Shin SJ, You SC, Jeon H, Jung JW, An MH, Park RW, et al. Style transfer strategy for developing a generalizable deep learning application in digital pathology. Computer Methods and Programs in Biomedicine. 2021;198:105815.

[132] Zeng H, Chen L, Zhang M, Luo Y, Ma X. Integration of histopathological images and multi-dimensional omics analyses predicts molecular features and prognosis in high-grade serous ovarian cancer. Gynecologic oncology. 2021;163(1):171-80.

[133] Boehm KM, Aherne EA, Ellenson L, Nikolovski I, Alghamdi M, Vázquez-García I, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature cancer. 2022;3(6):723-33.

[134] Boschman J, Farahani H, Darbandsari A, Ahmadvand P, Van Spankeren A, Farnell D, et al. The utility of color normalization for ai-based diagnosis of hematoxylin and eosin-stained pathology images. The Journal of Pathology. 2022;256(1):15-24.

[135] Elie N, Giffard F, Blanc-Fournier C, Morice PM, Brachet PE, Dutoit S, et al. Impact of automated methods for quantitative evaluation of immunostaining: Towards digital pathology. Frontiers in oncology. 2022;12:931035.

[136] Farahani H, Boschman J, Farnell D, Darbandsari A, Zhang A, Ahmadvand P, et al. Deep learning-based histotype diagnosis of ovarian carcinoma whole-slide pathology images. Modern Pathology. 2022;35(12):1983-90.

[137] Hu Y, Sirinukunwattana K, Gaitskell K, Wood R, Verrill C, Rittscher J. Predicting molecular traits from tissue morphology through self-interactive multi-instance

learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2022. p. 130-9.

[138] Jiang J, Tekin B, Yuan L, Armasu S, Winham SJ, Goode EL, et al. Computational tumor stroma reaction evaluation led to novel prognosis-associated fibrosis and molecular signature discoveries in high-grade serous ovarian carcinoma. Frontiers in Medicine. 2022;9:994467.

[139] Kasture KR, Choudhari D, Matte PN. Prediction and Classification of Ovarian Cancer using Enhanced Deep Convolutional Neural Network. International Journal of Engineering Trends and Technology. 2022;70:310-8.

[140] Kowalski PA, Błoniarz J, Chmura Ł. Convolutional neural networks in the ovarian cancer detection. In: Computational Intelligence and Mathematics for Tackling Complex Problems 2. Springer; 2022. p. 55-64.

[141] Lazard T, Bataillon G, Naylor P, Popova T, Bidard FC, Stoppa-Lyonnet D, et al. Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images. Cell Reports Medicine. 2022;3(12).

[142] Liu T, Su R, Sun C, Li X, Wei L. EOCSA: Predicting prognosis of Epithelial ovarian cancer with whole slide histopathological images. Expert Systems with Applications. 2022;206:117643.

[143] Mayer RS, Gretser S, Heckmann LE, Ziegler PK, Walter B, Reis H, et al. How to learn with intentional mistakes: NoisyEnsembles to overcome poor tissue quality for deep learning in computational pathology. Frontiers in Medicine. 2022;9:959068.

[144] Nero C, Boldrini L, Lenkowicz J, Giudice MT, Piermattei A, Inzani F, et al. Deep-learning to predict brca mutation and survival from digital h&e slides of epithelial ovarian cancer. International Journal of Molecular Sciences. 2022;23(19):11326.

[145] Salguero J, Prasanna P, Corredor G, Cruz-Roa A, Becerra D, Romero E. Selecting training samples for ovarian cancer classification via a semi-supervised clustering approach. In: Medical Imaging 2022: Digital and Computational Pathology. vol. 12039. SPIE; 2022. p. 20-4.

[146] Wang CW, Lee YC, Chang CC, Lin YJ, Liou YA, Hsu PC, et al. A weakly supervised deep learning method for guiding ovarian cancer treatment and identifying an effective biomarker. Cancers. 2022;14(7):1651.

[147] Wang CW, Chang CC, Lee YC, Lin YJ, Lo SC, Hsu PC, et al. Weakly supervised deep learning for prediction of treatment effectiveness on ovarian cancer from histopathology images. Computerized Medical Imaging and Graphics. 2022;99:102093.

[148] Yokomizo R, Lopes TJ, Takashima N, Hirose S, Kawabata A, Takenaka M, et al. O3C Glass-Class: A Machine-Learning Framework for Prognostic Prediction of Ovarian Clear-Cell Carcinoma. Bioinformatics and Biology Insights. 2022;16:11779322221134312.

[149] Ho DJ, Chui MH, Vanderbilt CM, Jung J, Robson ME, Park CS, et al. Deep Interactive Learning-based ovarian cancer segmentation of H&E-stained whole slide images to study morphological patterns of BRCA mutation. Journal of Pathology Informatics. 2023;14:100160.

[150] Meng Z, Wang G, Su F, Liu Y, Wang Y, Yang J, et al. A Deep Learning–Based System Trained for Gastrointestinal Stromal Tumor Screening Can Identify Multiple Types of Soft Tissue Tumors. The American Journal of Pathology. 2023;193(7):899-912.

[151] Ramasamy S, Kaliyaperumal V. A hybridized channel selection approach with deep convolutional neural network for effective ovarian cancer prediction in periodic acid-Schiff-stained images. Concurrency and Computation: Practice and Experience. 2023;35(5):e7568.

[152] Wang CW, Lee YC, Lin YJ, Chang CC, Wang CH, Chao TK, et al. Interpretable attention-based deep learning ensemble for personalized ovarian cancer treatment without manual annotations. Computerized Medical Imaging and Graphics. 2023;107:102233.

[153] Wu M, Zhu C, Yang J, Cheng S, Yang X, Gu S, et al. Exploring prognostic indicators in the pathological images of ovarian cancer based on a deep survival network. Frontiers in Genetics. 2023;13:1069673.

[154] Holback C, Jarosz R, Prior F, Mutch DG, Bhosale P, Garcia K, et al. The cancer genome atlas ovarian cancer collection (tcga-ov)(version 4)[data set]. The Cancer Imaging Archive. 2016.

[155] He K, Gan C, Li Z, Rekik I, Yin Z, Ji W, et al. Transformers in medical image analysis: A review. Intelligent Medicine. 2022.

[156] Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of internal medicine. 2019;170(1):51-8.

[157] Shrestha P, Poudyal B, Yadollahi S, Wright DE, Gregory AV, Warner JD, et al. A systematic review on the use of artificial intelligence in gynecologic imaging–Background, state of the art, and future directions. Gynecologic Oncology. 2022.

[158] Bueno LM, Caun DL, Comelis MT, Beguelini MR, Taboga SR, Morielle-Versute E. Ovarian morphology and folliculogenesis and ovulation process in the flat-faced fruit-eating bat Artibeus planirostris and the Argentine brown bat Eptesicus furinalis: A comparative analysis. Acta Zoologica. 2019;100(3):245-56.

[159] Siar M, Teshnehlab M. Brain tumor detection using deep neural network and machine learning algorithm. In: 2019 9th international conference on computer and knowledge engineering (ICCKE). IEEE; 2019. p. 363-8.

[160] Asadi-Aghbolaghi M, Farahani H, Zhang A, Akbari A, Kim S, Chow A, et al. Machine Learning-driven Histotype Diagnosis of Ovarian Carcinoma: Insights from the OCEAN AI Challenge. medRxiv. 2024:2024-04.

[161] Ahmed A, Xiaoyang Z, Tunio MH, Butt MH, Shah SA, Chengxiao Y, et al. OCCNET: Improving Imbalanced Multi-Centred Ovarian Cancer Subtype Classification in Whole Slide Images. In: 2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE; 2023. p. 1-8.

[162] Paayas P, Annamalai R. OCEAN-Ovarian Cancer subtypE clAssification and outlier detectioN using DenseNet121. In: 2023 Seventh International Conference on Image Information Processing (ICIIP). IEEE; 2023. p. 827-31.

[163] Aelgani V, Vadlakonda D. A novel interpretable regularized cnn with a modified xlnet transformer for segmenting and classifying the ovarian cancer. Multimedia Tools and Applications. 2024:1-28.

[164] Ziyambe B, Yahya A, Mushiri T, Tariq MU, Abbas Q, Babar M, et al. A deep learning framework for the prediction and diagnosis of ovarian cancer in pre-and post-menopausal women. Diagnostics. 2023;13(10):1703.

[165] Kasture KR, Patil WV, Shankar A. Comparative Analysis of Deep Learning Models for Early Prediction and Subtype Classification of Ovarian Cancer: A Comprehensive Study. International Journal of Intelligent Systems and Applications in Engineering. 2024;12(7s):507-15.

[166] Sundari MJ, Brintha N. TLOD: Innovative ovarian tumor detection for accurate multiclass classification and clinical application. Network Modeling Analysis in Health Informatics and Bioinformatics. 2024;13(1):18.

[167] Radhakrishnan M, Sampathila N, Muralikrishna H, Swathi K. Advancing Ovarian Cancer Diagnosis through Deep Learning and eXplainable AI: A Multiclassification Approach. IEEE Access. 2024.

[168] Nakhli R, Rich K, Zhang A, Darbandsari A, Shenasa E, Hadjifaradji A, et al. Volta: an environment-aware contrastive cell representation learning for histopathology. Nature Communications. 2024;15(1):3942.

[169] Behera SK, Das A, Sethy PK. Deep fine-KNN classification of ovarian cancer subtypes using efficientNet-B0 extracted features: a comprehensive analysis. Journal of cancer research and clinical oncology. 2024;150(7):361.

[170] Alahmadi A. Towards Ovarian Cancer Diagnostics: A Vision Transformer-based Computer-Aided Diagnosis Framework with Enhanced Interpretability. Results in Engineering. 2024:102651.

[171] Asadi-Aghbolaghi M, Darbandsari A, Zhang A, Contreras-Sanz A, Boschman J, Ahmadvand P, et al. Learning generalizable AI models for multi-center histopathology image classification. NPJ Precision Oncology. 2024;8(1):151.

[172] Ma J, Guo Z, Zhou F, Wang Y, Xu Y, Cai Y, et al. Towards A Generalizable Pathology Foundation Model via Unified Knowledge Distillation. arXiv preprint arXiv:240718449. 2024.

[173] Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, et al. A whole-slide foundation model for digital pathology from real-world data. Nature. 2024:1-8.

[174] Rodriguez EM, Lopez JS, Becerra D, Castro ER. Estimation of the Survival Time of Ovarian Serous Carcinoma by approximating a Tumor Microenvironment. In: 2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM). IEEE; 2023. p. 1-4.

[175] Machuca-Aguado J, Conde-Martín AF, Alvarez-Muñoz A, Rodríguez-Zarco E, Polo-Velasco A, Rueda-Ramos A, et al. Machine Learning Quantification of Intraepithelial Tumor-Infiltrating Lymphocytes as a Significant Prognostic Factor in High-Grade Serous Ovarian Carcinomas. International Journal of Molecular Sciences. 2023;24(22):16060.

[176] van Wagensveld L, Walker C, Hahn K, Sanders J, Kruitwagen R, van der Aa M, et al. The prognostic value of tumor-stroma ratio and a newly developed computer-aided quantitative analysis of routine H&E slides in high-grade serous ovarian cancer. Research Square. 2023:rs-3.

[177] Aggarwal A, Khalighi S, Babu D, Li H, Azarianpour-Esfahani S, Corredor G, et al. Computational pathology identifies immune-mediated collagen disruption to predict clinical outcomes in gynecologic malignancies. Communications Medicine. 2024;4(1):2.

[178] Hamada K, Murakami R, Ueda A, Kashima Y, Miyagawa C, Taki M, et al. A Deep Learning–Based Assessment Pipeline for Intraepithelial and Stromal Tumor-Infiltrating Lymphocytes in High-Grade Serous Ovarian Carcinoma. The American Journal of Pathology. 2024.

[179] Yang Z, Zhang Y, Zhuo L, Sun K, Meng F, Zhou M, et al. Prediction of prognosis and treatment response in ovarian cancer patients from histopathology images using graph deep learning: a multicenter retrospective study. European Journal of Cancer. 2024;199:113532.

[180] Zhou R, Zhao B, Ding H, Fu Y, Li H, Wei Y, et al. Survival prediction of ovarian serous carcinoma based on machine learning combined with pathological images and clinical information. AIP Advances. 2024;14(4).

[181] Bontempo G, Bartolini N, Lovino M, Bolelli F, Virtanen A, Ficarra E. Enhancing PFI Prediction with GDS-MIL: A Graph-based Dual Stream MIL Approach. In: International Conference on Image Analysis and Processing. Springer; 2023. p. 550-62.

[182] Wang CW, Lee YC, Lin YJ, Firdi NP, Muzakky H, Liu TC, et al. Deep Learning Can Predict Bevacizumab Therapeutic Effect and Microsatellite Instability Directly from Histology in Epithelial Ovarian Cancer. Laboratory Investigation. 2023;103(11):100247.

[183] Ahn B, Moon D, Kim HS, Lee C, Cho NH, Choi HK, et al. Histopathologic image–based deep learning classifier for predicting platinum-based treatment responses in high-grade serous ovarian cancer. Nature Communications. 2024;15(1):4253.

[184] Kilim O, Olar A, Biricz A, Madaras L, Pollner P, Szallasi Z, et al. Histopathology and proteomics are synergistic for High-Grade Serous Ovarian Cancer platinum response prediction. medRxiv. 2024:2024-06.

[185] Wang Q, Bi Q, Qu L, Deng Y, Wang X, Zheng Y, et al. MAMILNet: advancing precision oncology with multi-scale attentional multi-instance learning for whole slide image analysis. Frontiers in Oncology. 2024;14.

[186] Zhu Q, Dai H, Qiu F, Lou W, Wang X, Deng L, et al. Heterogeneity of computational pathomic signature predicts drug resistance and intra-tumor heterogeneity of ovarian cancer. Translational Oncology. 2024;40:101855.

[187] Liao X, Li K, Gan Z, Pu Y, Qian G, Zheng X. Prognostic prediction of ovarian cancer based on hierarchical sampling & fine-grained recognition convolution neural network. Alexandria Engineering Journal. 2024;102:264-78.

[188] Mao Y, Hu Z, Zhang X, Tong T. TransPBMIL: Transformer-Based Weakly Supervised Prognostic Prediction in Ovarian Cancer with Pseudo-Bag Strategy. In: International Conference on Intelligent Computing. Springer; 2024. p. 171-80.

[189] Wang CW, Firdi NP, Chu TC, Faiz MFI, Iqbal MZ, Li Y, et al. ATEC23 Challenge: Automated prediction of treatment effectiveness in ovarian cancer using histopathological images. Medical Image Analysis. 2024:103342.

[190] Das A, Chilakarao M, Biswas P, Sethy PK, Dalai MK, Behera SK. DeepOvaNet: A Comprehensive Deep Learning Framework for Predicting and Diagnosing Ovarian Cancer in Women Across Menopausal Transitions. In: 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). IEEE; 2024. p. 1-7.

[191] Falana WO, Serener A, Serte S. Deep Learning for Comparative Study of Ovarian Cancer Detection on Histopathological Images. In: 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE; 2023. p. 1-6.

[192] Salguero J, Prasanna P, Corredor G, Cruz-Roa A, Becerra D, Romero E. Data distillation in computational pathology by choosing few representants of the original variance: A use case in ovarian cancer. Expert Systems with Applications. 2024;245:123028.

[193] Singh S, Maurya MK, Singh NP. STRAMPN: Histopathological image dataset for ovarian cancer detection incorporating AI-based methods. Multimedia Tools and Applications. 2024;83(9):28175-96.

[194] Mishra R, Kumar M, Brindha S, Sharma AK, Raja C, Moharekar TT. An Efficient Deep Learning Model for Intraoperative Tissue Classification in Gynecological Cancer. In: 2023 9th International Conference on Smart Structures and Systems (ICSSS). IEEE; 2023. p. 1-6.

[195] Kwatra CV, Kaur H. Enhancing Ovarian Cancer Detection: A Deep Learning Approach with MobileNetV3 and ResNet50. In: 2023 Seventh International Conference on Image Information Processing (ICIIP). IEEE; 2023. p. 188-93.

[196] Ariotta V, Lehtonen O, Salloum S, Micoli G, Lavikka K, Rantanen V, et al. H&E image analysis pipeline for quantifying morphological features. Journal of pathology informatics. 2023;14:100339.

[197] Bourgade R, Rabilloud N, Perennec T, Pécot T, Garrec C, Guédon AF, et al. Deep Learning for Detecting BRCA Mutations in High-Grade Ovarian Cancer Based on an Innovative Tumor Segmentation Method From Whole Slide Images. Modern Pathology. 2023;36(11):100304.

[198] Bogaerts JM, Bokhorst JM, Simons M, van Bommel MH, Steenbeek MP, de Hullu JA, et al. Deep learning detects premalignant lesions in the Fallopian tube. npj Women's Health. 2024;2(1):11.

[199] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. Circulation. 2015;131(2):211-9.

[200] Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. bmj. 2024;385.

[201] Dehkharghanian T, Bidgoli AA, Riasatian A, Mazaheri P, Campbell CJ, Pantanowitz L, et al. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. Diagnostic pathology. 2023;18(1):1-12.

[202] Perivolaris A, Adams-McGavin C, Madan Y, Kishibe T, Antoniou T, Mamdani M, et al. Quality of Interaction Between Clinicians and Artificial Intelligence Systems. A Systematic Review. Future Healthcare Journal. 2024:100172.

[203] Köbel M, Kalloger SE, Baker PM, Ewanowich CA, Arseneau J, Zherebitskiy V, et al. Diagnosis of ovarian carcinoma cell type is highly reproducible: a transcanadian study. The American journal of surgical pathology. 2010;34(7):984-93.

[204] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

[205] Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nature Machine Intelligence. 2020;2(11):665-73.

[206] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289-300.

[207] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019;32.

[208] Chen RJ, Ding T, Lu MY, Williamson DF, Jaume G, Song AH, et al. Towards a general-purpose foundation model for computational pathology. Nature Medicine. 2024:1-13.

[209] The National Pathology Imaging Co-operative (NPIC). The Leeds Guide to Digital Pathology Volume 2; 2023. Accessed: 2024-07-24. Available from: https://npic.ac.uk/wp-content/uploads/sites/71/2022/12/Horizontal-Leeds-guide-volume-2.pdf.

[210] Lerousseau M, Deutsch E, Paragios N. Multimodal brain tumor classification. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6. Springer; 2021. p. 475-86.

[211] Zhu X, Yao J, Zhu F, Huang J. Wsisa: Making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 7234-42.

[212] Couture HD, Marron JS, Perou CM, Troester MA, Niethammer M. Multiple instance learning for heterogeneous images: Training a cnn for histopathology. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. Springer; 2018. p. 254-62.

[213] Katharopoulos A, Fleuret F. Processing megapixel images with deep attention-sampling models. In: International Conference on Machine Learning. PMLR; 2019. p. 3282-91.

[214] Koriakina N, Sladoje N, Lindblad J. The effect of within-bag sampling on end-to-end multiple instance learning. In: 2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA). IEEE; 2021. p. 183-8.

[215] Koriakina N, Sladoje N, Bašić V, Lindblad J. Deep multiple instance learning versus conventional deep single instance learning for interpretable oral cancer detection. Plos one. 2024;19(4):e0302169.

[216] Combalia M, Vilaplana V. Monte-Carlo sampling applied to multiple instance learning for histological image classification. In: International Workshop on Deep Learning in Medical Image Analysis. Springer; 2018. p. 274-81.

[217] Broad A, Wright AI, de Kamps M, Treanor D. Attention-guided sampling for colorectal cancer analysis with digital pathology. Journal of Pathology Informatics. 2022;13:100110.

[218] Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih N, et al. High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection. PloS one. 2018;13(5):e0196828.

[219] Qu L, Yang Z, Duan M, Ma Y, Wang S, Wang M, et al. Boosting whole slide image classification from the perspectives of distribution, correlation and magnification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 21463-73.

[220] Hashimoto N, Fukushima D, Koga R, Takagi Y, Ko K, Kohno K, et al. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 3852-61.

[221] D'Amato M, Szostak P, Torben-Nielsen B. A comparison between single-and multi-scale approaches for classification of histopathology images. Frontiers in Public Health. 2022;10:892658.

[222] Rasoolijaberi M, Babaei M, Riasatian A, Hemati S, Ashrafi P, Gonzalez R, et al. Multi-magnification image search in digital pathology. IEEE Journal of Biomedical and Health Informatics. 2022;26(9):4611-22.

[223] Zaffar I, Jaume G, Rajpoot N, Mahmood F. Embedding space augmentation for weakly supervised learning in whole-slide images. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). IEEE; 2023. p. 1-4.

[224] Ciga O, Xu T, Martel AL. Self supervised contrastive learning for digital histopathology. Machine Learning with Applications. 2022;7:100198.

[225] Wang X, Yang S, Zhang J, Wang M, Zhang J, Yang W, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis. 2022;81:102559.

[226] Kang M, Song H, Park S, Yoo D, Pereira S. Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 3344-54.

[227] Filiot A, Ghermi R, Olivier A, Jacob P, Fidon L, Mac Kain A, et al. Scaling self-supervised learning for histopathology with masked image modeling. medRxiv. 2023:2023-07.

[228] Wang W, Ma S, Xu H, Usuyama N, Ding J, Poon H, et al. When an image is worth 1,024 x 1,024 words: A case study in computational pathology. arXiv preprint arXiv:231203558. 2023.

[229] Nechaev D, Pchelnikov A, Ivanova E. Hibou: A Family of Foundational Vision Transformers for Pathology. arXiv preprint arXiv:240605074. 2024.

[230] Vorontsov E, Bozkurt A, Casson A, Shaikovski G, Zelechowski M, Severson K, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Medicine. 2024:1-12.

[231] Saillard C, Jenatton R, Llinares-López F, Mariet Z, Cahané D, Durand E, et al.. H-optimus-0; 2024. Accessed: 2024-09-25. Available from: https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

[232] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.

[233] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.

[234] Campanella G, Chen S, Verma R, Zeng J, Stock A, Croken M, et al. A Clinical Benchmark of Public Self-Supervised Pathology Foundation Models. arXiv preprint arXiv:240706508. 2024.

[235] Neidlinger P, Nahhas OSME, Muti HS, Lenz T, Hoffmeister M, Brenner H, et al. Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. arXiv preprint arXiv:240815823. 2024.

[236] kaiko ai, Aben N, de Jong ED, Gatopoulos I, Känzig N, Karasikov M, et al. Towards Large-Scale Training of Pathology Foundation Models. arXiv preprint arXiv:240415217. 2024.

[237] PyTorch. ResNet50 Documentation; 2022. Accessed: 2024-09-26. Available from: https://pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html.

[238] PyTorch. ResNet18 Documentation; 2024. Accessed: 2024-09-26. Available from: https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html.

[239] Google Research. ViT-L Documentation; 2024. Accessed: 2024-09-26. Available from: https://github.com/google-research/vision_transformer?tab=readme-ov-file#available-vit-models.

[240] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR; 2020. p. 1597-607.

[241] Zbontar J, Jing L, Misra I, LeCun Y, Deny S. Barlow twins: Self-supervised learning via redundancy reduction. In: International conference on machine learning. PMLR; 2021. p. 12310-20.

[242] Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 9650-60.

[243] Zhou J, Wei C, Wang H, Shen W, Xie C, Yuille A, et al. ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:211107832. 2021.

[244] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:230407193. 2023.

[245] Kanwal N, Pérez-Bueno F, Schmidt A, Engan K, Molina R. The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review. Ieee Access. 2022;10:58821-44.

[246] Otsu N. A threshold selection method from gray-level histograms. Automatica. 1975;11(285-296):23-7.

[247] Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: an open-source quality control tool for digital pathology slides. JCO clinical cancer informatics. 2019;3:1-7.

[248] Shakhawat H, Hossain S, Kabir A, Mahmud SH, Islam MM, Tariq F. Review of artifact detection methods for automated analysis and diagnosis in digital pathology. In: Artificial Intelligence For Disease Diagnosis And Prognosis In Smart Healthcare. CRC Press; 2023. p. 177-202.

[249] Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. IEEE Computer graphics and applications. 2001;21(5):34-41.

[250] Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE international symposium on biomedical imaging: from nano to macro. IEEE; 2009. p. 1107-10.

[251] Dooper S, Pinckaers H, Aswolinskiy W, Hebeda K, Jarkman S, van der Laak J, et al. Gigapixel end-to-end training using streaming and attention. Medical Image Analysis. 2023;88:102881.

[252] Shao Z, Dai L, Wang Y, Wang H, Zhang Y, et al. AugDiff: Diffusion based feature augmentation for multiple instance learning in whole slide image. IEEE Transactions on Artificial Intelligence. 2024.

[253] Wang Y, Farnell D, Farahani H, Nursey M, Tessier-Cloutier B, Jones SJ, et al. Classification of epithelial ovarian carcinoma whole-slide pathology images using deep transfer learning. arXiv preprint arXiv:200510957. 2020.

[254] Wölflein G, Ferber D, Meneghetti AR, Nahhas OSME, Truhn D, Carrero ZI, et al. Benchmarking Pathology Feature Extractors for Whole Slide Image Classification. arXiv preprint arxiv:231111772. 2024.

[255] Gilks C, Davidson B, Köbel M, Ledermann J, Lim D, Malpica A, et al. Ovary, Fallopian Tube and Primary Peritoneal Carcinoma Histopathology Reporting Guide. International Collaboration on Cancer Reporting. 2021; 2nd edition.

[256] Bibal A, Cardon R, Alfter D, Wilkens R, Wang X, François T, et al. Is attention explanation? an introduction to the debate. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022. p. 3889-900.

[257] Altman AD, Nelson GS, Ghatage P, McIntyre JB, Capper D, Chu P, et al. The diagnostic utility of TP53 and CDKN2A to distinguish ovarian high-grade serous carcinoma from low-grade serous ovarian tumors. Modern Pathology. 2013;26(9):1255-63.

[258] Adnan M, Kalra S, Tizhoosh HR. Representation learning of histopathology images using graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020. p. 988-9.

[259] Chen RJ, Lu MY, Shaban M, Chen C, Chen TY, Williamson DF, et al. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. Springer; 2021. p. 339-49.

[260] Guan Y, Zhang J, Tian K, Yang S, Dong P, Xiang J, et al. Node-aligned graph convolutional network for whole-slide image representation and classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 18813-23.

[261] Liang M, Chen Q, Li B, Wang L, Wang Y, Zhang Y, et al. Interpretable classification of pathology whole-slide images using attention based context-

aware graph convolutional neural network. Computer methods and programs in biomedicine. 2023;229:107268.

[262] Bazargani R, Fazli L, Gleave M, Goldenberg L, Bashashati A, Salcudean S. Multi-scale relational graph convolutional network for multiple instance learning in histopathology images. Medical Image Analysis. 2024;96:103197.

[263] Xing X, Ma Y, Jin L, Sun T, Xue Z, Shi F, et al. A Multi-scale Graph Network with Multi-head Attention for Histopathology Image Diagnosisn. In: COMPAY@ MICCAI; 2021. p. 227-35.

[264] Godson L, Alemi N, Nsengimana J, Cook GP, Clarke EL, Treanor D, et al. Multi-level Graph Representations of Melanoma Whole Slide Images for Identifying Immune Subgroups. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2023. p. 85-96.

[265] Shi J, Tang L, Li Y, Zhang X, Gao Z, Zheng Y, et al. A structure-aware hierarchical graph-based multiple instance learning framework for pt staging in histopathological image. IEEE Transactions on Medical Imaging. 2023;42(10):3000-11.

[266] Brody S, Alon U, Yahav E. How attentive are graph attention networks? arXiv preprint arXiv:210514491. 2021.

[267] Lee J, Lee I, Kang J. Self-attention graph pooling. In: International conference on machine learning. PMLR; 2019. p. 3734-43.

[268] Lee Y, Park JH, Oh S, Shin K, Sun J, Jung M, et al. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. Nature Biomedical Engineering. 2022:1-15.

[269] Miller RE, Elyashiv O, El-Shakankery KH, Ledermann JA. Ovarian cancer therapy: homologous recombination deficiency as a predictive biomarker of response to PARP inhibitors. OncoTargets and Therapy. 2022:1105-17.

[270] Li R, Yao J, Zhu X, Li Y, Huang J. Graph CNN for survival analysis on whole slide pathological images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2018. p. 174-82.

[271] Wang CW, Chang CC, Khalil MA, Lin YJ, Liou YA, Hsu PC, et al. Histopathological whole slide image dataset for classification of treatment effectiveness to ovarian cancer. Scientific Data. 2022;9(1):25.

Appendices

A Systematic Review Search Strategy

The full searches used in the systematic literature review (Chapter 3) are shown here, with any text which was not directly input to the search bar in **bold**. These searches were each a combination of three aspects - artificial intelligence, ovarian cancer, and histopathology. No filters were applied, and all options were left on their default settings. The wildcard character, *, was used to search for multiple versions of the same word, for example, “patholog*” searched for all of “pathology”, “pathologist”, “pathologists”, and “pathological”.

A.1 PubMed

(“Machine Learning”[Mesh] OR “Artificial Intelligence”[Mesh] OR “Neural Networks, Computer”[Mesh] OR “support vector machine”[MeSH] OR “Deep Learning”[Mesh] OR “diagnosis, computer-assisted”[Mesh] OR “Machine learn*” OR “Artificial Intelligen*” OR (ML[Title/Abstract] NOT (μgml[Title/Abstract] OR μ/ml[Title/Abstract] OR mgml[Title/Abstract] OR pgml[Title/Abstract] OR ngml[Title/Abstract] OR uiml[Title/Abstract] OR iuml[Title/Abstract] OR miuml[Title/Abstract] OR muiml[Title/Abstract] OR uml[Title/Abstract] OR gml[Title/Abstract] OR mlkg[Title/Abstract] OR milliliter*[Title/Abstract])) OR AI[Title/Abstract] OR “Computer Vision” OR “Neural network*” OR “Deep Network*” OR “Computer-aided Diagnosis” OR “Computer aided Diagnosis” OR Perceptron* OR “Convolutional Network*” OR “Recurrent Network*” OR “Graph Network*” OR “Deep Learn*” OR “Deep-Learn*” OR Backprop* OR “support vector*” OR ensemble* OR “random forest*” OR “nearest neighbor*” OR “nearest neighbour*” OR “k-nearest neighbor*” OR “k-nearest neighbour*” OR “Gradient boost*” OR “XGBoost*” OR “segmentation” OR “instance learning” OR “multi-instance learning” OR “Active Learning”)

AND (((ovar* OR fallopian) AND (cancer* OR mass* OR carcinoma* OR tumour* OR tumor* OR neoplasm* OR malignan* OR “carcinoma”[Mesh] OR “neoplasms”[Mesh]))

OR “Ovarian Neoplasms”[Mesh] OR “peritoneal cancer” OR “peritoneal carcinoma” OR “peritoneal tumo∗”)

AND ((digit∗ AND patholog∗) OR “computational patholog∗” OR “tissue microarray∗” OR histopath∗ OR histolog∗ OR “Whole Slide Imag∗” OR “Tissue slide∗” OR “pathology slide∗” OR “pathology image∗” OR Immunohistochem∗ OR ((Haematoxylin OR Hematoxylin) AND Eosin) OR Histology[Mesh])

A.2 Scopus

TITLE-ABS-KEY(“Machine learn∗” OR “Artificial Intelligen∗” OR (“ML” AND NOT “∗ μ ml” AND NOT “∗g ml” AND NOT “∗ui ml” AND NOT “∗UI ml” AND NOT “∗iu ml” AND NOT “∗u ml” AND NOT “∗g ml” AND NOT “∗ml kg” AND NOT milliliter∗) OR AI OR “Computer Vision” OR “Neural network∗” OR “Deep Network∗” OR “Computer-aided Diagnosis” OR “Computer aided Diagnosis” OR Perceptron∗ OR “Convolutional Network∗” OR “Recurrent Network∗” OR “Graph Network∗” OR “Deep Learn∗” OR “Deep-Learn∗” OR Backprop∗ OR “support vector∗” OR ensemble∗ OR “random forest∗” OR “nearest neighbor∗” OR “nearest neighbour∗” OR “k-nearest neighbor∗” OR “k-nearest neighbour∗” OR “Gradient boost∗” OR “XGBoost∗” OR “segmentation” OR “instance learning” OR “multi-instance learning” OR “Active Learning”)

AND TITLE-ABS-KEY(((ovar∗ OR fallopian) AND (cancer∗ OR mass∗ OR carcinoma∗ OR tumour∗ OR tumor∗ OR neoplasm∗ OR malignan∗)) OR “peritoneal cancer” OR “peritoneal carcinoma” OR “peritoneal tumo∗”)

AND TITLE-ABS-KEY((digit∗ AND patholog∗) OR “computational patholog∗” OR “tissue microarray∗” OR histopath∗ OR histolog∗ OR “Whole Slide Imag∗” OR “Tissue slide∗” OR “pathology slide∗” OR “pathology image∗” OR Immunohistochem∗ OR ((Haematoxylin OR Hematoxylin) AND Eosin))

A.3 Web of Science

(ALL=(“Machine learn∗” OR “Artificial Intelligen∗” OR “Computer Vision” OR “Neural network∗” OR “Deep Network∗” OR “Computer-aided Diagnosis” OR “Computer aided Diagnosis” OR Perceptron∗ OR “Convolutional Network∗” OR “Recurrent Network∗”

OR “Graph Network∗” OR “Deep Learn∗” OR “Deep-Learn∗” OR Backprop∗ OR “support vector∗” OR ensemble∗ OR “random forest∗” OR “nearest neighbor∗” OR “nearest neighbour∗” OR “k-nearest neighbor∗” OR “k-nearest neighbour∗” OR “Gradient boost∗” OR “XGBoost∗” OR “segmentation” OR “instance learning” OR “multi-instance learning” OR “Active Learning”) OR TS=(AI OR (“ML” NOT (“∗ μ ml” OR “∗g ml” OR “∗ui ml” OR “∗UI ml” OR “∗iu ml” OR “∗u ml” OR “∗g ml” OR “∗ml kg” OR milliliter∗))))

AND ALL=(((ovar∗ OR fallopian) AND (cancer∗ OR mass∗ OR carcinoma∗ OR tumour∗ OR tumor∗ OR neoplasm∗ OR malignan∗)) OR “peritoneal cancer” OR “peritoneal carcinoma” OR “peritoneal tumo∗”)

AND ALL=((digit∗ AND patholog∗) OR “computational patholog∗” OR “tissue microarray∗” OR histopath∗ OR histolog∗ OR “Whole Slide Imag∗” OR “Tissue slide∗” OR “pathology slide∗” OR “pathology image∗” OR Immunohistochem∗ OR ((Haematoxylin OR Hematoxylin) AND Eosin))

A.4 Cochrane Central Register of Controlled Trials

Search #1:

All text: (“Machine learn∗” OR “Artificial Intelligen∗” OR “Computer Vision” OR “Neural network∗” OR “Deep Network∗” OR “Computer-aided Diagnosis” OR “Computer aided Diagnosis” OR Perceptron∗ OR “Convolutional Network∗” OR “Recurrent Network∗” OR “Graph Network∗” OR “Deep Learn∗” OR “Deep-Learn∗” OR Backprop∗ OR “support vector∗” OR ensemble∗ OR “random forest∗” OR “nearest neighbor∗” OR “nearest neighbour∗” OR “k-nearest neighbor∗” OR “k-nearest neighbour∗” OR “Gradient boost∗” OR “XGBoost∗” OR “segmentation” OR “instance learning” OR “multi-instance learning” OR “Active Learning”)

Search #2:

Title-Abstract-Keyword: (“AI” OR (“ML” NOT (“∗ μ ml” OR “∗g ml” OR “∗ui ml” OR “∗UI ml” OR “∗iu ml” OR “∗u ml” OR “∗g ml” OR “∗ml kg” OR milliliter∗))) in Title Abstract Keyword

Search #3:

All text: (((ovar∗ OR fallopian) AND (cancer∗ OR mass∗ OR carcinoma∗ OR tumour∗ OR tumor∗ OR neoplasm∗ OR malignan∗)) OR “peritoneal cancer” OR “peritoneal carcinoma” OR “peritoneal tumo∗”)

AND ((digit∗ AND patholog∗) OR “computational patholog∗” OR “tissue microarray∗” OR histopath∗ OR histolog∗ OR “Whole Slide Imag∗” OR “Tissue slide∗” OR “pathology slide∗” OR “pathology image∗” OR Immunohistochem∗ OR ((Haematoxylin OR Hematoxylin) AND Eosin))

Final search:

(#1 OR #2) AND #3

A.5 WHO-ICTRP

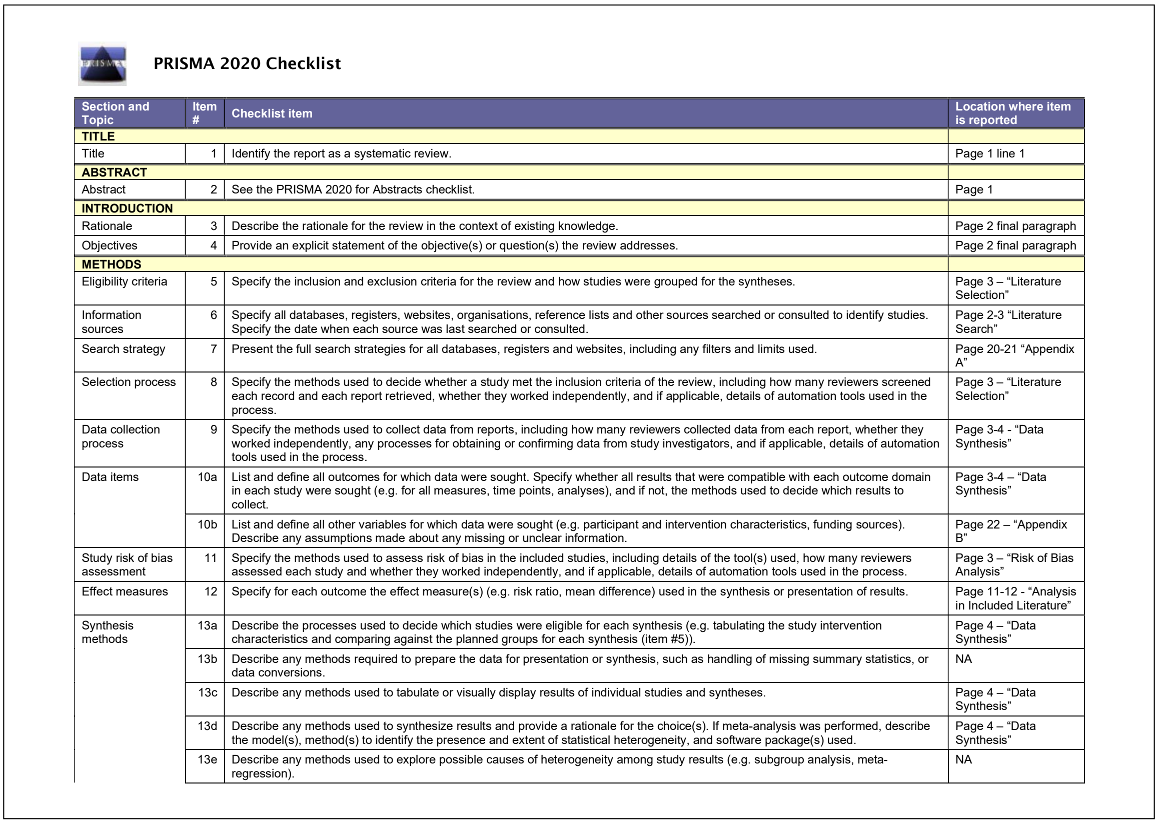
((“Machine learn∗” OR “Artificial Intelligen∗” OR “Computer Vision” OR “Neural network∗” OR “Deep Network∗” OR “Computer-aided Diagnosis” OR “Computer aided Diagnosis” OR Perceptron∗ OR “Convolutional Network∗” OR “Recurrent Network∗” OR “Graph Network∗” OR “Deep Learn∗” OR “Deep-Learn∗” OR Backprop∗ OR “support vector∗” OR ensemble∗ OR “random forest∗” OR “nearest neighbor∗” OR “nearest neighbour∗” OR “k-nearest neighbor∗” OR “k-nearest neighbour∗” OR “Gradient boost∗” OR “XGBoost∗” OR “segmentation” OR “instance learning” OR “multi-instance learning” OR “Active Learning”) OR (“AI” OR (“ML” NOT (“μ/ml” OR “g/ml” OR “ui/ml” OR “UI/ml” OR “iu/ml” OR “u/ml” OR “g/ml” OR “ml/kg” OR milliliter∗))))

AND (((ovar∗ OR fallopian) AND (cancer∗ OR mass∗ OR carcinoma∗ OR tumour∗ OR tumor∗ OR neoplasm∗ OR malignan∗)) OR “peritoneal cancer” OR “peritoneal carcinoma” OR “peritoneal tumo∗”)

AND ((digit∗ AND patholog∗) OR “computational patholog∗” OR “tissue microarray∗” OR histopath∗ OR histolog∗ OR “Whole Slide Imag∗” OR “Tissue slide∗” OR “pathology slide∗” OR “pathology image∗” OR Immunohistochem∗ OR ((Haematoxylin OR Hematoxylin) AND Eosin))

B PRISMA 2020 Reporting Checklist

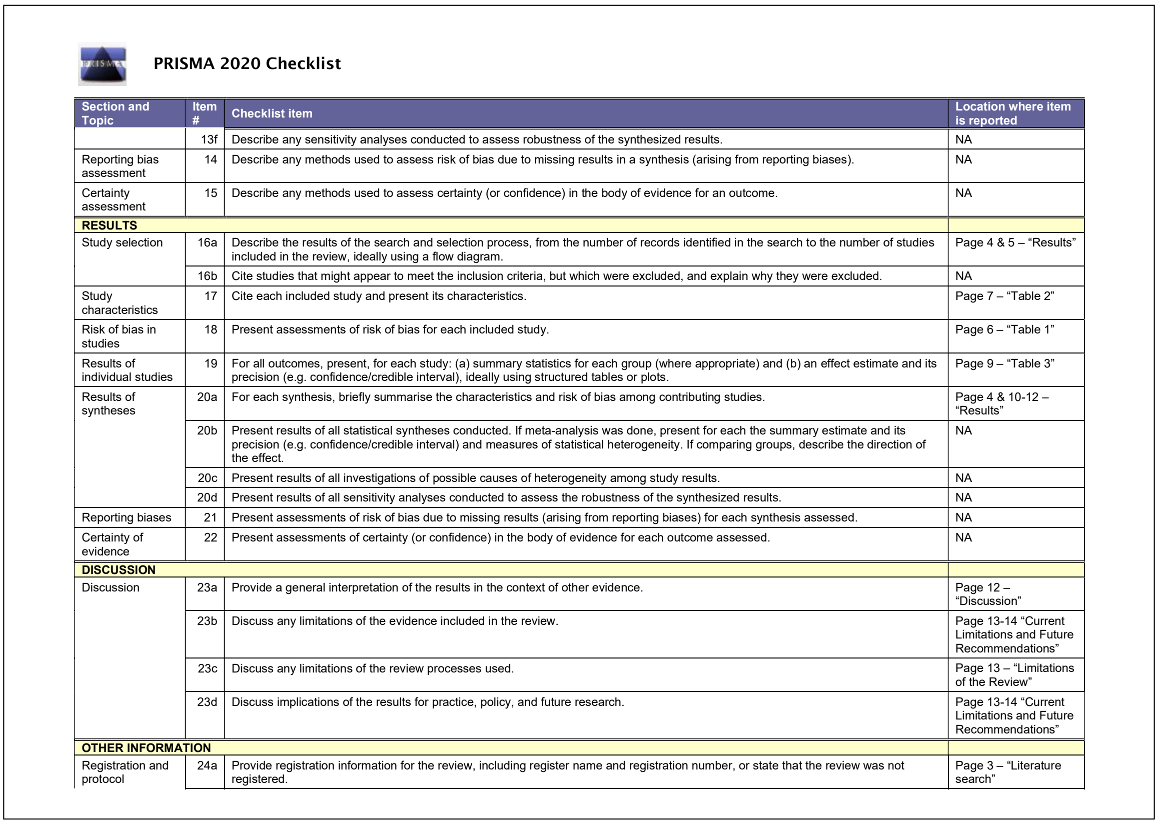
The following PRISMA 2020 reporting checklist is from the published version of the systematic literature review presented in Chapter 3.



PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Page 1 line 1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Page 1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Page 2 final paragraph
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Page 2 final paragraph
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Page 3 – “Literature Selection”
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Page 2-3 “Literature Search”
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Page 20-21 “Appendix A”
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Page 3 – “Literature Selection”
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Page 3-4 - “Data Synthesis”
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Page 3-4 – “Data Synthesis”
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Page 22 – “Appendix B”
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Page 3 – “Risk of Bias Analysis”
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Page 11-12 - “Analysis in Included Literature”
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Page 4 – “Data Synthesis”
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	NA
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Page 4 – “Data Synthesis”
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Page 4 – “Data Synthesis”
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	NA

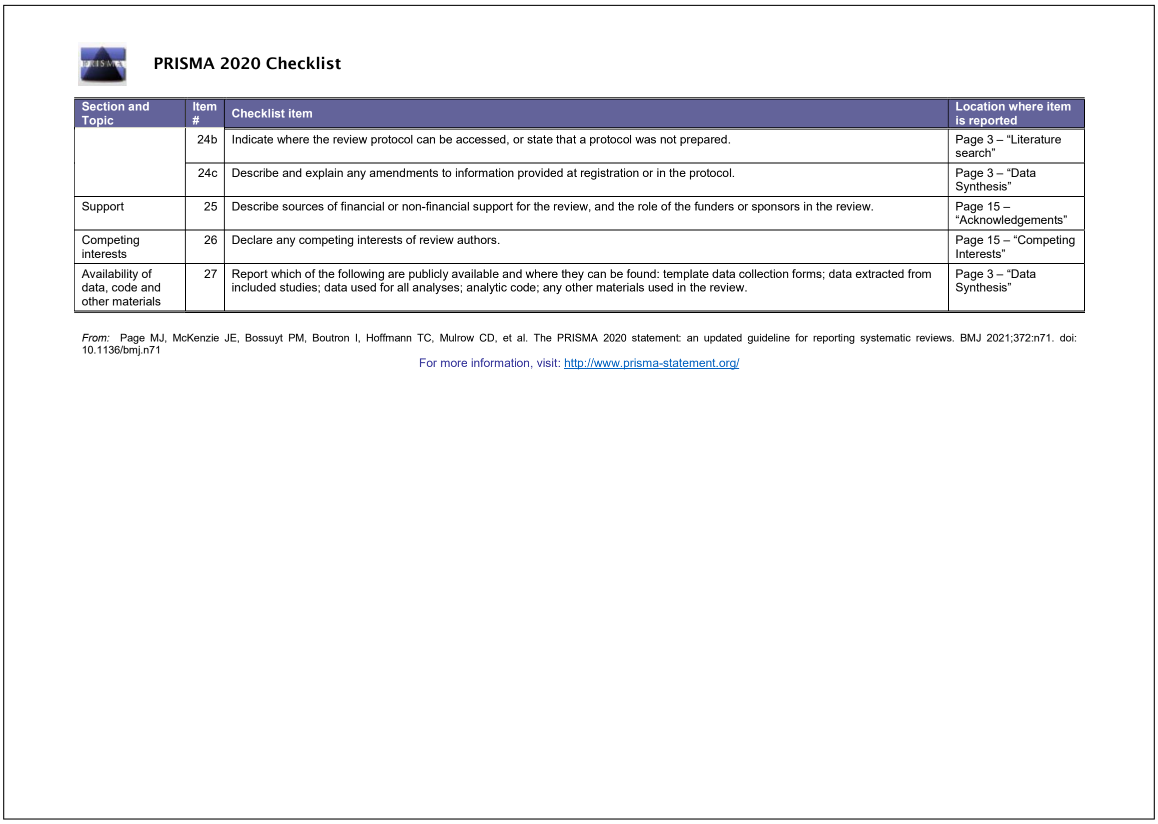
Figure B.1 PRISMA 2020 Checklist Page 1.



PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	NA
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	NA
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	NA
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Page 4 & 5 – "Results"
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	NA
Study characteristics	17	Cite each included study and present its characteristics.	Page 7 – "Table 2"
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Page 6 – "Table 1"
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Page 9 – "Table 3"
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Page 4 & 10-12 – "Results"
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	NA
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	NA
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	NA
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	NA
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	NA
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Page 12 – "Discussion"
	23b	Discuss any limitations of the evidence included in the review.	Page 13-14 "Current Limitations and Future Recommendations"
	23c	Discuss any limitations of the review processes used.	Page 13 – "Limitations of the Review"
	23d	Discuss implications of the results for practice, policy, and future research.	Page 13-14 "Current Limitations and Future Recommendations"
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Page 3 – "Literature search"

Figure B.2 PRISMA 2020 Checklist Page 2.



PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Page 3 – "Literature search"
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Page 3 – "Data Synthesis"
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Page 15 – "Acknowledgements"
Competing interests	26	Declare any competing interests of review authors.	Page 15 – "Competing Interests"
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Page 3 – "Data Synthesis"

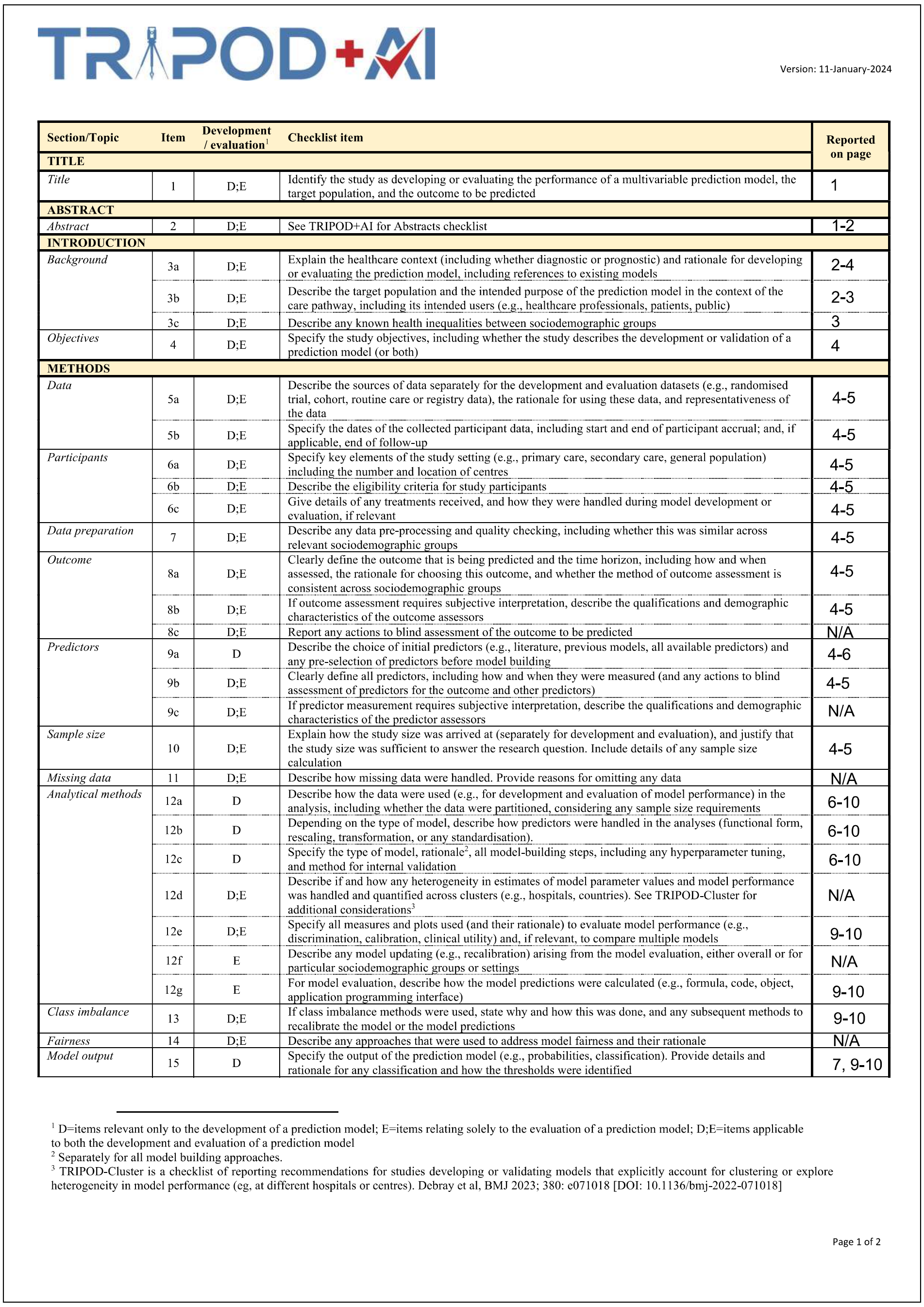
From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

For more information, visit: http://www.prisma-statement.org/

Figure B.3 PRISMA 2020 Checklist Page 3.

C TRIPOD+AI Reporting Checklist

The following TRIPOD+AI reporting checklist is from the preprint paper created from the work presented in Chapter 7.



TRIPOD+AI

Version: 11-January-2024

Section/Topic	Item	Development / evaluation[1]	Checklist item	Reported on page
TITLE				
Title	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	1
ABSTRACT				
Abstract	2	D;E	See TRIPOD+AI for Abstracts checklist	1-2
INTRODUCTION				
Background	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	2-4
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	2-3
	3c	D;E	Describe any known health inequalities between sociodemographic groups	3
Objectives	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	4
METHODS				
Data	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	4-5
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	4-5
Participants	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	4-5
	6b	D;E	Describe the eligibility criteria for study participants	4-5
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	4-5
Data preparation	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	4-5
Outcome	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	4-5
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	4-5
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	N/A
Predictors	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	4-6
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	4-5
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	N/A
Sample size	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	4-5
Missing data	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	N/A
Analytical methods	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	6-10
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	6-10
	12c	D	Specify the type of model, rationale[2], all model-building steps, including any hyperparameter tuning, and method for internal validation	6-10
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations[3]	N/A
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	9-10
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	N/A
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	9-10
Class imbalance	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	9-10
Fairness	14	D;E	Describe any approaches that were used to address model fairness and their rationale	N/A
Model output	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	7, 9-10

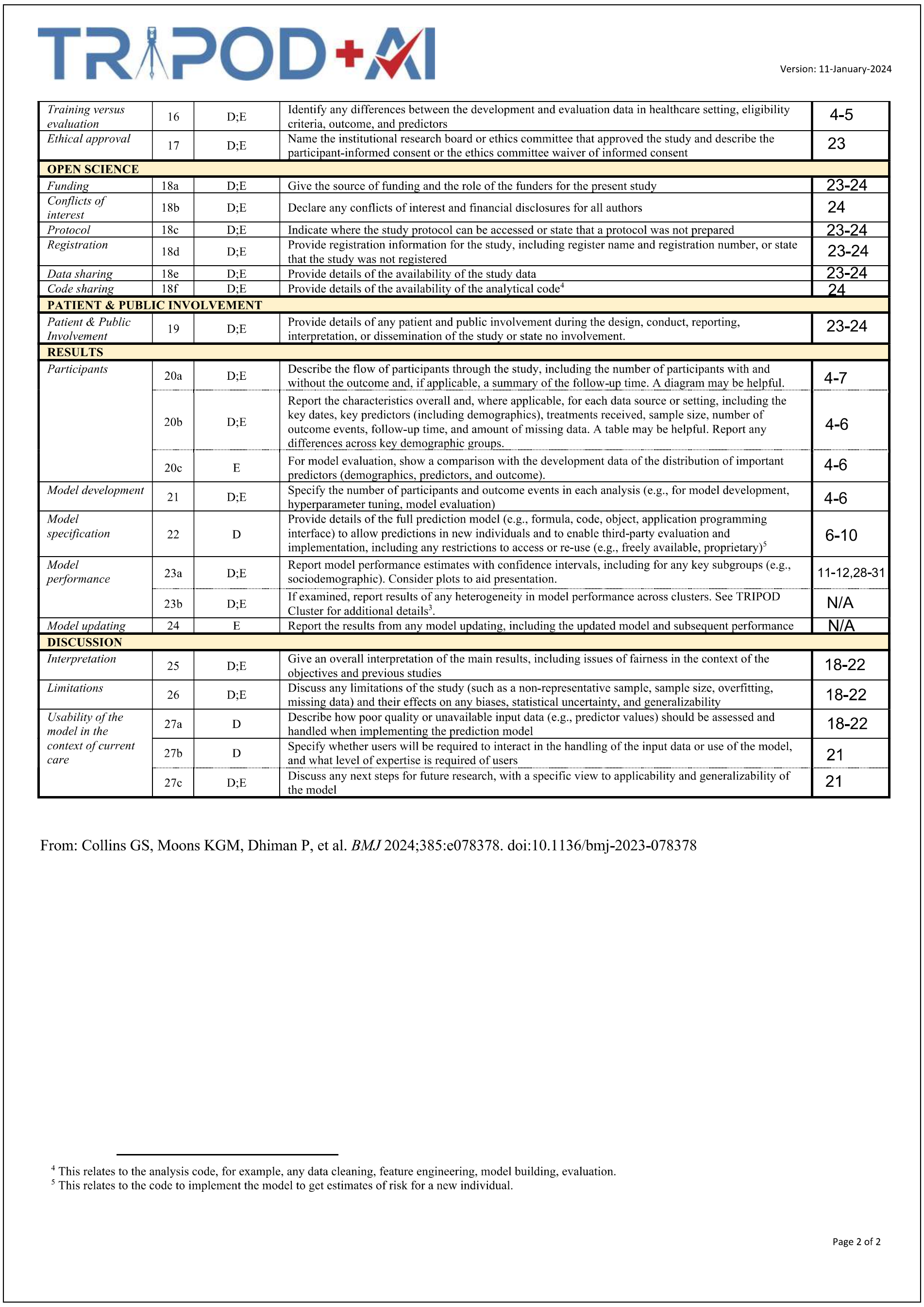
[1] D=items relevant only to the development of a prediction model; E=items relating solely to the evaluation of a prediction model; D;E=items applicable to both the development and evaluation of a prediction model

[2] Separately for all model building approaches.

[3] TRIPOD-Cluster is a checklist of reporting recommendations for studies developing or validating models that explicitly account for clustering or explore heterogeneity in model performance (eg, at different hospitals or centres). Debray et al, BMJ 2023; 380: e071018 [DOI: 10.1136/bmj-2022-071018]

Page 1 of 2

Figure C.1 TRIPOD+AI Checklist Page 1.



TRIPOD+AI

Version: 11-January-2024

Training versus evaluation	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	4-5
Ethical approval	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	23
OPEN SCIENCE				
Funding	18a	D;E	Give the source of funding and the role of the funders for the present study	23-24
Conflicts of interest	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	24
Protocol	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	23-24
Registration	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	23-24
Data sharing	18e	D;E	Provide details of the availability of the study data	23-24
Code sharing	18f	D;E	Provide details of the availability of the analytical code[4]	24
PATIENT & PUBLIC INVOLVEMENT				
Patient & Public Involvement	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	23-24
RESULTS				
Participants	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	4-7
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	4-6
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	4-6
Model development	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	4-6
Model specification	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary)[5]	6-10
Model performance	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	11-12,28-31
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details[3].	N/A
Model updating	24	E	Report the results from any model updating, including the updated model and subsequent performance	N/A
DISCUSSION				
Interpretation	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	18-22
Limitations	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	18-22
Usability of the model in the context of current care	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	18-22
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	21
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	21

From: Collins GS, Moons KGM, Dhiman P, et al. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378

[4] This relates to the analysis code, for example, any data cleaning, feature engineering, model building, evaluation.
[5] This relates to the code to implement the model to get estimates of risk for a new individual.

Page 2 of 2

Figure C.2 TRIPOD+AI Checklist Page 2.

D Results of Hypothesis Testing

The following results are supplementary to the hyperparameter tuning ablation in Section 7.3.3.

	Cross-Validation p-values			Hold-out p-values		
Model	**Balanced Accuracy**	**AUROC**	**F1 Score**	**Balanced Accuracy**	**AUROC**	**F1 Score**
RN50	0.617	0.264	0.133	0.171	0.252	0.133
RN18	0.967	0.259	0.170	0.326	0.252	0.170
ViT-L	**0.012**	**0.005**	**0.010**	**0.006**	0.095	**0.010**
RN18-Histo	**0.002**	0.095	0.145	0.086	0.671	0.145
Lunit	0.555	**0.011**	0.168	0.054	0.124	0.074
RN50-Histo	0.864	0.630	0.902	0.912	0.100	0.895
CTransPath	0.144	0.987	**0.042**	**0.030**	0.099	**0.042**
Hibou-B	0.159	0.069	**0.009**	**0.008**	0.207	**0.009**
Phikon	0.709	0.280	0.741	0.619	0.114	0.741
Kaiko-B8	**0.039**	0.089	0.124	0.099	0.063	0.124
GPFM	0.500	**0.029**	0.236	0.262	0.055	0.236
UNI	**0.003**	**0.015**	**0.021**	**0.033**	0.614	**0.021**
Hibou-L	0.104	0.050	0.070	**0.019**	0.193	**0.016**
Virchow	**0.039**	0.059	0.104	0.069	0.076	0.104
Virchow2-CLS	0.194	**0.035**	0.095	0.083	0.108	0.095
H-Optimus-0	0.111	0.069	0.133	0.119	0.089	0.133
Prov-GigaPath	0.412	0.297	0.215	0.194	**0.035**	0.215

	Transcanadian Study p-values			OCEAN Challenge p-values		
Model	**Balanced Accuracy**	**AUROC**	**F1 Score**	**Balanced Accuracy**	**AUROC**	**F1 Score**
RN50	0.190	0.178	0.219	0.303	0.098	0.716
RN18	0.240	0.217	0.106	0.339	0.056	0.279
ViT-L	**0.014**	0.109	**0.014**	**0.001**	**0.006**	**0.021**
RN18-Histo	0.578	0.774	0.973	0.212	0.182	0.620
Lunit	0.099	0.774	0.099	0.104	**0.049**	0.056
RN50-Histo	0.601	0.135	0.479	0.818	**0.023**	0.605
CTransPath	0.853	0.341	0.998	0.790	0.630	0.562
Hibou-B	0.300	0.076	0.286	0.700	0.172	0.590
Phikon	0.740	0.085	0.306	0.119	0.467	0.189
Kaiko-B8	0.213	0.125	0.342	0.102	**0.014**	**0.028**
GPFM	0.386	0.120	0.405	0.861	0.080	0.176
UNI	0.370	0.085	0.959	**0.002**	**0.008**	**0.014**
Hibou-L	0.087	**0.040**	**0.003**	0.142	**0.004**	0.196
Virchow	0.478	0.379	0.460	**0.049**	**0.012**	**0.057**
Virchow2-CLS	0.057	0.871	0.057	**0.017**	0.066	**0.035**
H-Optimus-0	0.167	0.751	0.192	0.866	0.054	0.168
Prov-GigaPath	0.274	0.060	0.209	0.258	0.124	0.416

Table D.1 Resulting p-values from paired *t*-tests comparing the subtype classification results for each feature extractor with and without hyperparameter tuning applied to the ABMIL classifier. Values below 0.05 are indicated in **bold**.

E Supplementary Attention Heatmap Analysis

This section contains supplementary information about the analysis of heatmaps performed in Chapter 7. Two pathologists (KA and NMO) qualitatively compared the UNI and ImageNet-pretrained ResNet50 attention heatmaps for ten class-balanced example WSIs from the internal hold-out test set. These WSIs (shown in Figures 5.1, E.1, and E.2) were selected from those in which a different classification had been determined by each model (specifically using the first-fold model of the five-model ensemble). Out of 39 total disagreements, the UNI-based model gave the correct classification in 26 cases, the ResNet50-based model in 3 cases, and neither was correct in 10 cases. The pathologists were only provided the heatmaps, and were blinded to the models used and the predictions made.

The heatmaps were determined to be similar between models, with both giving high attention to tumour regions and low attention to most stroma regions. Where differences occurred, the ResNet50-based model typically gave high attention to a larger tissue area, often including relevant stromal features (e.g. necrosis and psammoma bodies), but sometimes also including irrelevant stroma. When considering whether heatmaps had focused on diagnostically relevant regions, the pathologists expressed a preference for the UNI-based heatmap in four cases and the ResNet50-based heatmap in three cases, with no preference expressed for the remaining three cases due to their overwhelming similarity. In eight of the selected cases, the UNI model had correctly determined the classification, including all three cases in which the pathologists had preferred the ResNet50-based heatmap. In these cases, the UNI model did not appear to give sufficient attention to all relevant tissue, though it still determined the correct classification. Thus, there was some level of divergence between the pathologists' interpretations and the model heatmaps.

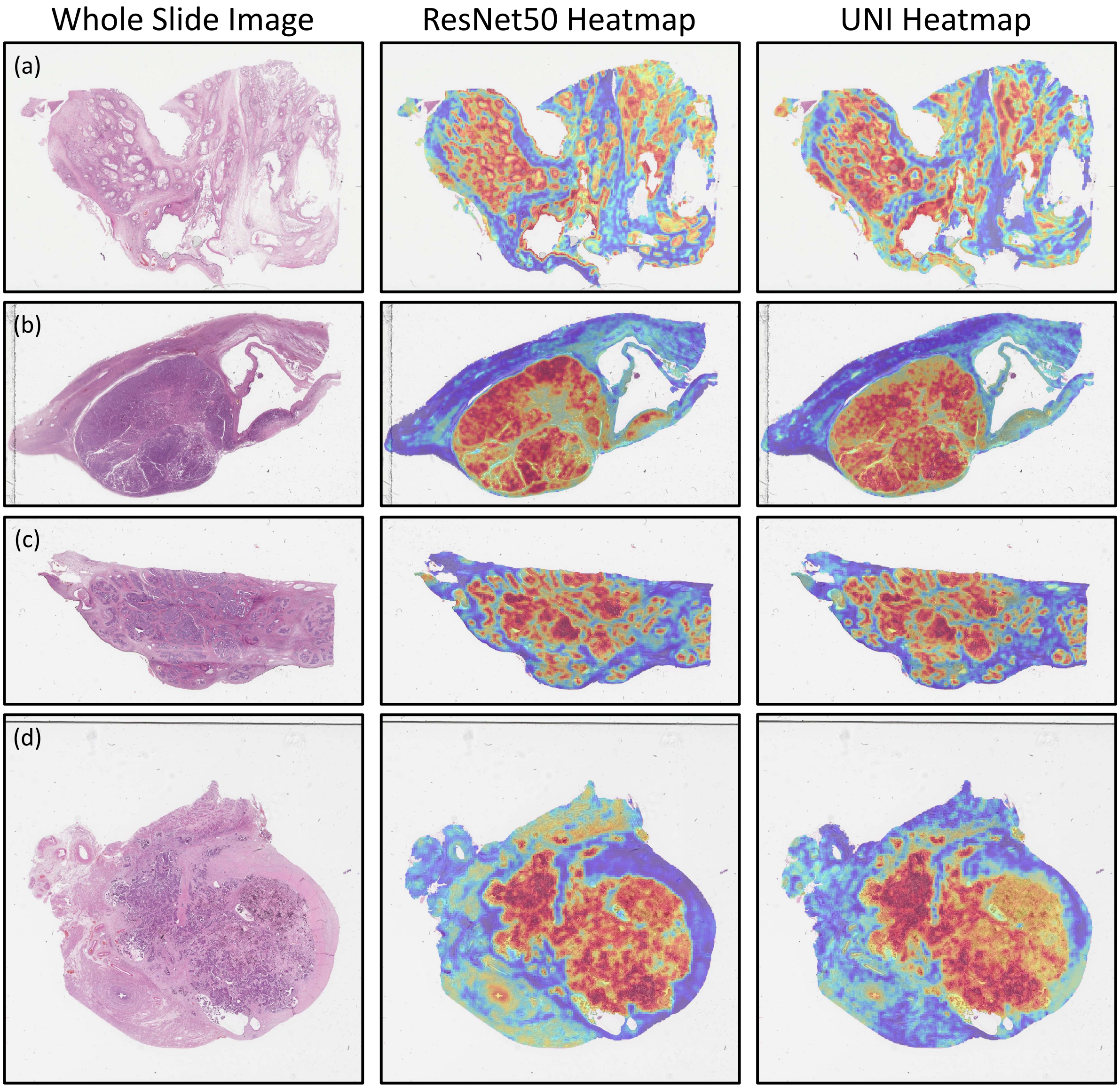


Figure E.1 Attention heatmaps from the ABMIL classifier using ImageNet-pretrained ResNet50 and UNI foundation model features, where the classification differed between the two models. (a) Ground truth: MC, ResNet50: CCC, UNI: MC. (b) Ground truth: CCC, ResNet50: HGSC, UNI: CCC. (c) Ground truth: EC, ResNet50: HGSC, UNI: EC. (d) Ground truth: LGSC, ResNet50: HGSC, UNI: LGSC.

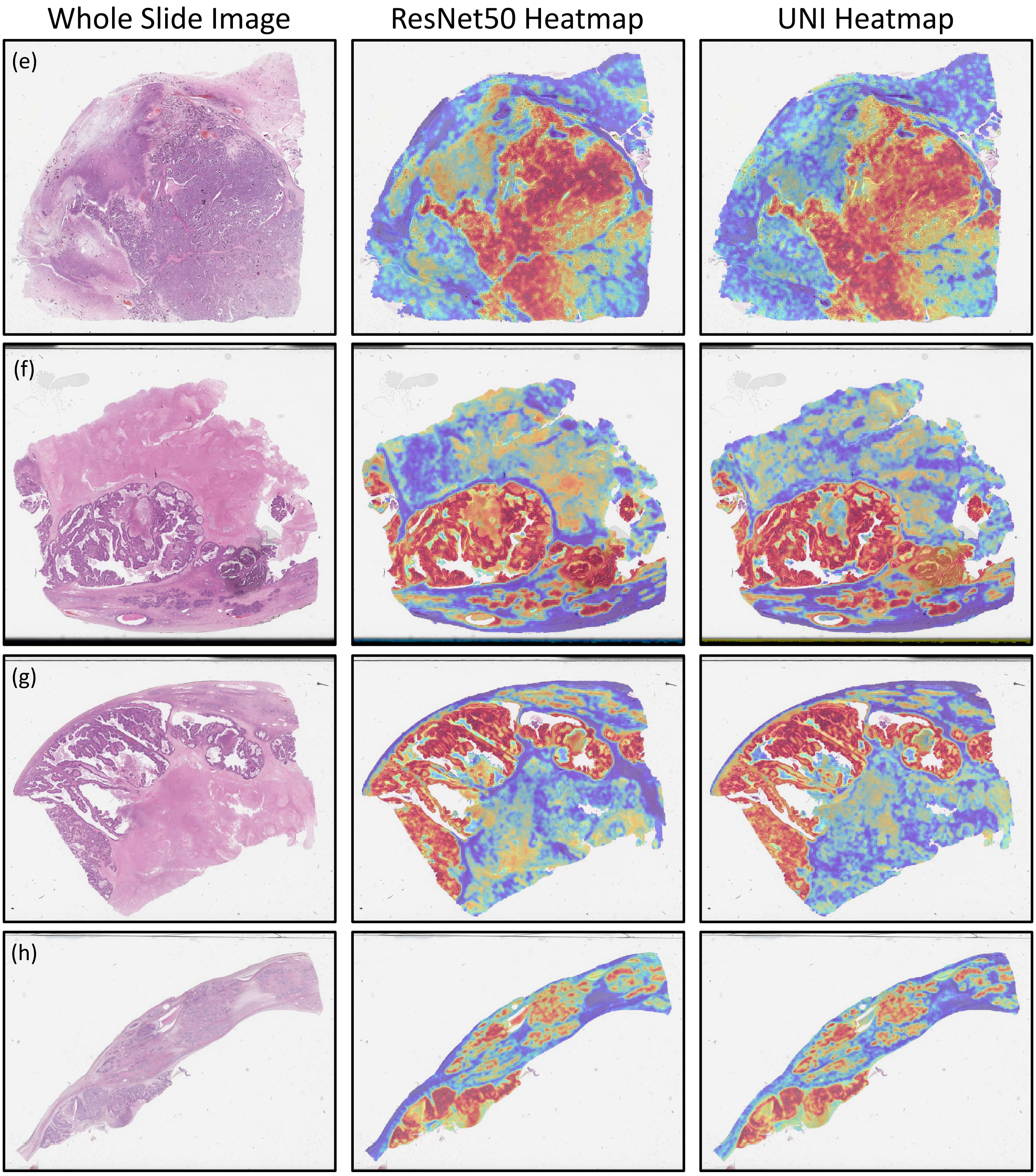


Figure E.2 Attention heatmaps from the ABMIL classifier using ImageNet-pretrained ResNet50 and UNI foundation model features, where the classification differed between the two models. (e) Ground truth: LGSC, ResNet50: HGSC, UNI: LGSC. (f) Ground truth: HGSC, ResNet50: HGSC, UNI: EC. (g) Ground truth: HGSC, ResNet50: HGSC, UNI: EC. (h) Ground truth: EC, ResNet50: HGSC, UNI: EC.

F Predicting Treatment Response

This section is based on work conducted to predict treatment response from ovarian cancer WSIs as part of the ATEC23 challenge [189] at MICCAI 2023. This strand of our research has not yet been continued due to a lack of high-quality data for prognostic tasks in ovarian cancer digital pathology. We instead continued to focus on histological subtyping, where there was sufficient data to conduct rigorous validations.

F.1 Introduction

Treatment options are guided by the stage, grade, and morphological subtype of ovarian cancer, and can often involve surgery, chemotherapy, and increasingly, immunotherapy. However, response to therapy can vary significantly, and the underlying causes are not well understood despite significant progress in defined subgroups, such as homologous recombination deficient tumours [269]. Due to this knowledge gap, some patients may be exposed to the adverse effects of a given therapy without deriving any clinical benefit. The ATEC23 challenge aimed to identify non-responders using pre-treatment histopathology WSIs alone.

Studies reporting higher accuracy in this particular area have used IHC panels [152], with performance being poorer in studies using H&E-stained tissue [124, 147]. A prediction model using H&E WSIs alone would offer greater clinical benefit given that this staining method is routine in all histopathological diagnostic interpretation of ovarian cancer specimens. Instead, dependence on IHC staining would add financial and time burdens to the diagnostic pathway.

An H&E baseline model was developed by the ATEC23 challenge organisers [147], in which a hierarchical attention approach was used to segment the most relevant tissue. ABMIL was then applied to this segmented tissue to classify WSIs. The reported results from 5-fold cross-validation presented an accuracy of 88.2% and an F1 score of 0.917, although the reported accuracy on an independent test set was no greater than random guessing.

None of the previous ovarian cancer treatment response studies have employed methods that capture spatial relationships within WSIs, such as vision transformers

[62] or graph networks [270]. Such methods are likely to be beneficial as there are established correlations between patient prognosis and the spatial arrangement of cellular structures visible in WSIs, with tumour-infiltrating lymphocytes being associated with survival in some ovarian cancer subtypes [34]. For this challenge, we combined vision transformers with ABMIL to classify whether patients would respond to a specific course of bevacizumab-based therapy from histopathology WSIs alone, as defined by measurable recurrence/progression within 6 months of treatment.

F.2 Methods

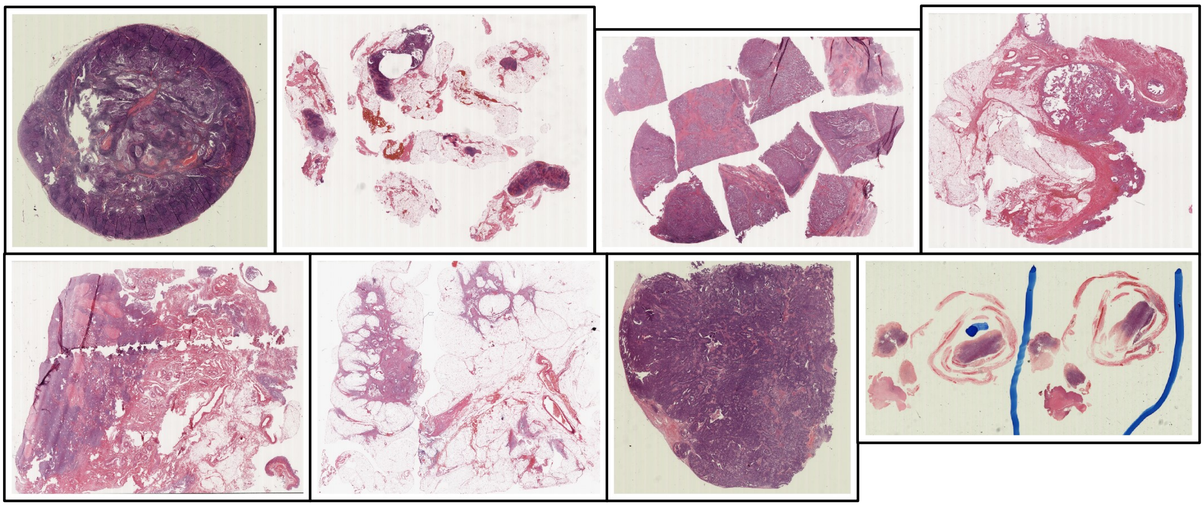


Figure F.1 Eight whole slide images from the ATEC23 challenge training set

The challenge training data [271] comprised 288 H&E-stained tissue section WSIs from 78 tubo-ovarian and primary peritoneal cancer patients, of which 53 were determined to have an *effective* response to treatment, and 25 were determined to have an *invalid* response to treatment. We used 282 WSIs from 78 patients due to two WSIs being inaccessible, two being duplicated, and two being erroneously excluded. All patients received debulking surgery, chemotherapy, and bevacizumab therapy, with treatment classified as *effective* if CA-125 levels fell and there was no tumour progression/recurrence found in CT/PET images within 6 months of treatment. All samples were originally collected from a single data centre and scanned using a single Leica AT Turbo scanner at 20x magnification. Patients had a range of morphological subtype diagnoses, including HGSC (n=58), CCC (n=7), unclassified carcinoma (n=7), EC (n=4), and MC (n=2). The slides in the dataset were highly heterogeneous (Figure F.1). Samples appeared to include a combination of adnexal, omental, and lymph node tissue, with some slides having differing colour profiles and artefacts, such as pen markings. An independent challenge test set was collected at the same data centre, consisting of 180 H&E-stained TMA single core images from patients diagnosed with HGSC.

Our HIPT-ABMIL classification approach, shown in Figure F.2, used ABMIL to classify WSIs based on region-level (4096 x 4096 pixel) features encoded through a two-stage vision transformer [83]. Before modelling, we used Otsu thresholding to segment tissue, then extracted 4096 x 4096 non-overlapping tissue regions for modelling. On average, the tissue patching procedure generated 91 regions per slide (range of 13 to 166).

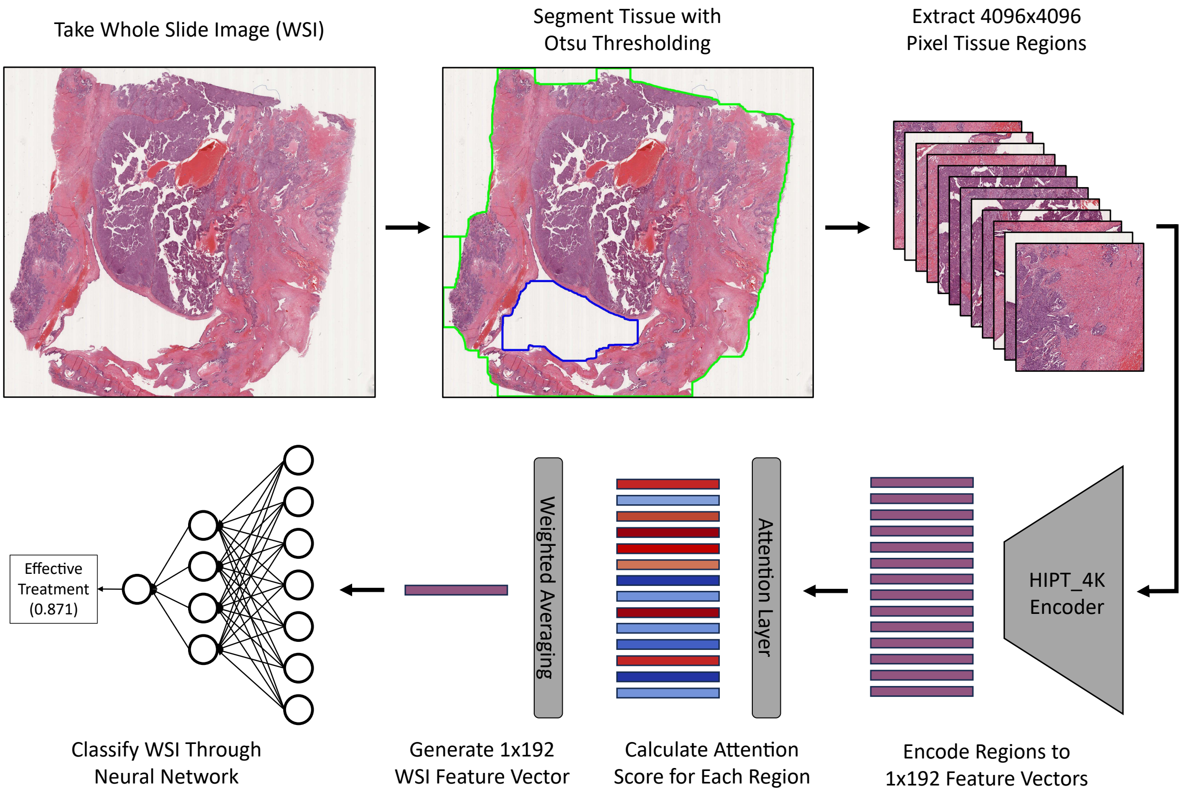

Take Whole Slide Image (WSI)
Segment Tissue with Otsu Thresholding
Extract 4096x4096 Pixel Tissue Regions
HIPT_4K Encoder
Encode Regions to 1x192 Feature Vectors
Attention Layer
Calculate Attention Score for Each Region
Weighted Averaging
Generate 1x192 WSI Feature Vector
Effective Treatment (0.871)
Classify WSI Through Neural Network

Figure F.2 HIPT-ABMIL whole slide image classification pipeline.

We extracted features from each tissue region using the two-stage Hierarchical Image Pyramid Transformer (HIPT_4K) [83]. This approach first uses a vision transformer [62] to aggregate cell-level information (16 x 16 pixels) to patch-level (256 x 256), then uses a second vision transformer to aggregate patch-level information to region-level (4096 x 4096). This feature extractor was pretrained using over ten thousand total histopathology slides from 33 cancer types using the self-supervised method *DINO*

[242]. We trained the ABMIL network using these HIPT region embeddings to classify WSIs. We also compared three other approaches - HIPT-CLAM, ResNet-ABMIL, and HistoResNet-ABMIL. HIPT-CLAM replaced the ABMIL classifier with CLAM. ResNet-ABMIL was the baseline model described in Section 4.1. HistoResNet-ABMIL was the same model but with features extracted through a ResNet18 encoder which was pretrained on a collection of 57 histopathology datasets [224] using the self-supervised technique *SimCLR* [240]. The smaller patch size in the ResNet approaches gave more patches per slide, with an average of 20214 (range of 2043 to 38828).

We trained our models using a cross-entropy loss and an Adam optimiser. As shown in Table F.1, we tuned hyperparameters across 5-fold cross-validation experiments, using a grid search strategy for five hyperparameters. The parameters were the learning rate, dropout rate, weight decay, model size, and number of patches per slide for training. The model size hyperparameter controlled the dimension of the attention layer, and the subsequent hidden layer in the classification network had a dimension half this size. One extra hyperparameter, B, was tuned for the CLAM model, which controlled the number of regions which were clustered in feature space during training. Each tuning configuration was repeated three times and the average loss was taken to account for random variations. Multiple stages of hyperparameter grid tuning were used, with earlier runs covering a wider range of parameters and influencing the hyperparameter options available in later stages. Each model was evaluated with over 500 total hyperparameter configurations.

We selected the hyperparameters which minimised the average validation loss across the 5-fold cross-validation to train the final model. Internal performance was measured on the cross-validation test sets, and the same hyperparameters were used to train a 4-fold ensemble model with 75%-25% train-val splits, with the mean predictions for the external TMA images submitted to the ATEC23 challenge. Due to the relatively small size of test set images, each one was represented as a single 4096 x 4096 region. All experiments were run on the HPC and code was made available at https://github.com/scjjb/HIPT_ABMIL_ATEC23, alongside further details of the hyperparameter tuning.

Hyperparameter	Function	Initial Tuning Options	Second Tuning Options	Third Tuning Options	Final Selection
Learning Rate	Sets the rate of change of model parameters trained using the Adam optimiser	1e-3, 1e-4, 1e-5	1e-3, 5e-4, 1e-4	1e-3, 5e-4	1e-3
Dropout Rate	Sets the proportion of model weights to drop in each training iteration	0.25, 0.5, 0.75	0.6, 0.75, 0.9	0.8, 0.85, 0.9, 0.95	0.85
Regularisation	Sets the level of weight decay in the Adam optimiser	1e-2, 1e-3, 1e-4	1e-1, 1e-2, 1e-3	1e-0, 5e-1, 1e-1, 5e-2	5e-1
Attention Layer Size	Sets the size of the attention layer, with the following hidden layer size set to half of this	64, 32, 16	32, 16, 8	32, 16	16
Patches per Slide	Sets the number of patches randomly selected from each slide per training epoch	25, 50, 75	25, 50, 75	50, 75, 100	75

Table F.1 HIPT-ABMIL hyperparameters tuned using a three-stage grid search. The same hyperparameters were tuned for ResNet-ABMIL and HistoResNet-ABMIL. An additional hyperparameter was tuned for HIPT-CLAM to set the number of regions used for clustering [78].

F.3 Results and Discussion

Method	Balanced Accuracy	AUROC	Accuracy	F1 Score
Challenge Baseline* [147]	NA	NA	88.2% ± 6%*	0.917 ± 0.07*
HIPT-ABMIL	**60.2% ± 2.9%**	0.646 ± 0.033	**61.0% ± 2.9%**	0.656 ± 0.031
HIPT-CLAM	57.6% ± 2.9%	0.624 ± 0.033	58.9% ± 2.9%	0.650 ± 0.031
ResNet-ABMIL	52.7% ± 2.9%	0.569 ± 0.034	54.3% ± 3.0%	0.617 ± 0.031
HistoResNet-ABMIL	58.1% ± 2.9%	**0.655 ± 0.032**	59.6% ± 2.9%	**0.660 ± 0.030**

Table F.2 5-fold cross-validation classification performance on the internal 282 WSIs (mean ± standard deviation from 100,000 iteration bootstrapping). *The best results are highlighted in **bold** except for the baseline results, which were generated by different researchers using different validation methods [147], as described in the Discussion.

The HIPT-ABMIL model had the greatest performance for two evaluated metrics and HistoResNet-ABMIL had the greatest performance for the other two. On the internal 5-fold test set, the HIPT-ABMIL model achieved a balanced accuracy of 60.2%±2.9%, AUROC of 0.646±0.033, and F1 score of 0.656±0.031 (mean ± one standard deviation from 100,000 iteration bootstrapping). The results were highly varied, with the AUROC per cross-validation fold being 0.381-0.825. There were also large differences between validation and test set performance in most folds, including a fold where validation AUROC was 0.400 higher (0.781 vs. 0.381) and another where test AUROC was 0.389 higher (0.436 vs. 0.825). The performance of the histopathology-pretrained models was much greater than the ImageNet-pretrained ResNet-ABMIL, which achieved just 52.7% balanced accuracy, barely greater than random guessing. No clear classification benefit was found from using hierarchical transformers compared to a ResNet, or from using CLAM rather than standard ABMIL. The optimal HIPT-ABMIL and HistoResNet-ABMIL models were each applied to the external ATEC23 TMA test set, though neither generalised well to this data (accuracies of 35% and 55% respectively).

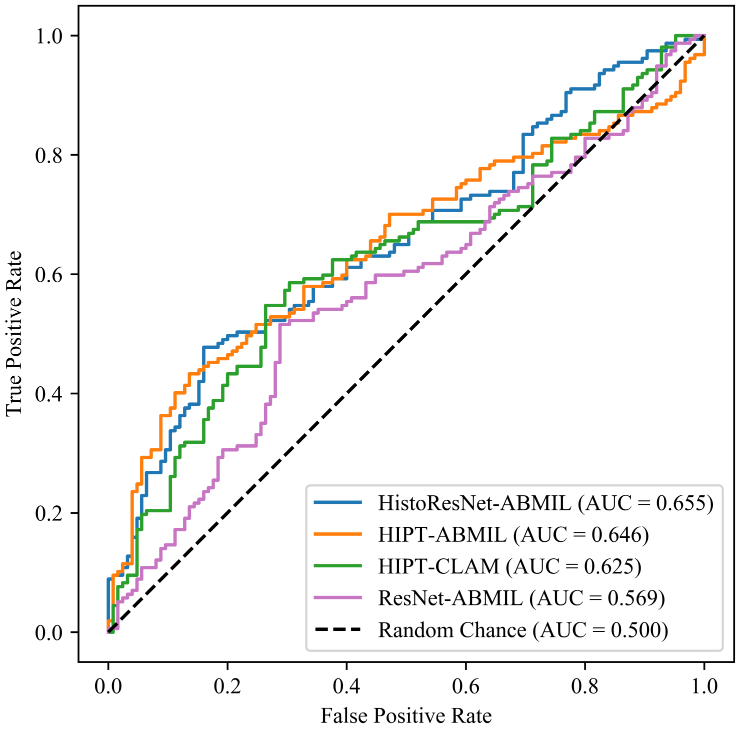
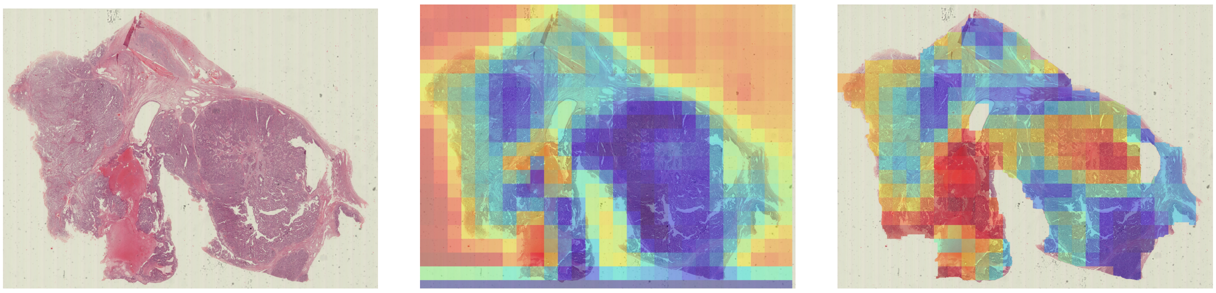


Figure F.3 Receiver operating characteristic (ROC) curves and the AUROC for each model from 5-fold cross-validation.

Our internal performance scores were much lower than the reported performance of the baseline approach (optimal F1 of 0.660 compared to 0.917, accuracy of 60.2% compared to 88.2% [147]). However, this is unlikely to be a fair comparison due to differences in the pre-processing, validation, and data used. Further validation would be beneficial in evaluating both approaches as there is a high risk that results were artificially inflated by confounding and bias caused by the high levels of heterogeneity in the relatively small dataset. We partially mitigated this by splitting data into train-val-test splits per patient, reducing the unduly high level of correlation between training and testing sets. However, there were other likely confounders which were not adequately controlled, with the dataset containing small quantities of WSIs with significant differences to the majority, such as different histological subtypes (HGSC, CCC, EC, MC, and unclassified carcinomas), tissue types/background histology (omentum, peritoneum, lymph node) and artefacts (pen markings, image stitching, out of focus regions). Such confounding could be moderated by using a larger, more clinically representative dataset. The large standard deviations in the results were also likely attributable to the relatively small dataset size, with a 95% confidence interval for the optimal balanced accuracy being 54.5% to 66.0%. No challenge participant achieved an accuracy greater than random chance, which may indicate that TMAs do not contain sufficient prognostic information.

The clinical utility of these models would benefit from a more precise and clinically relevant definition of outcome, as the ATEC23 binary classification grouped patients who relapsed after just over 6 months together with patients who never relapsed. Significant consideration should be given to the impact of carcinoma stage, grade and morphological subtype on outcome beyond the Cox models presented in previous research, which found strong but not statistically significant correlations between the subtype and outcome, and between the stage and outcome [147]. Further details about the cohort's patients in terms of their differing responses to platinum-based chemotherapy would also be informative as the model may be predicting response to a mixture of combination and single-agent therapies.



(a) Raw histopathology slide from the ATEC23 challenge training set.

(b) Corresponding ABMIL heatmap from the initial HIPT-ABMIL model.

(c) Corresponding ABMIL heatmap from the final HIPT-ABMIL model.

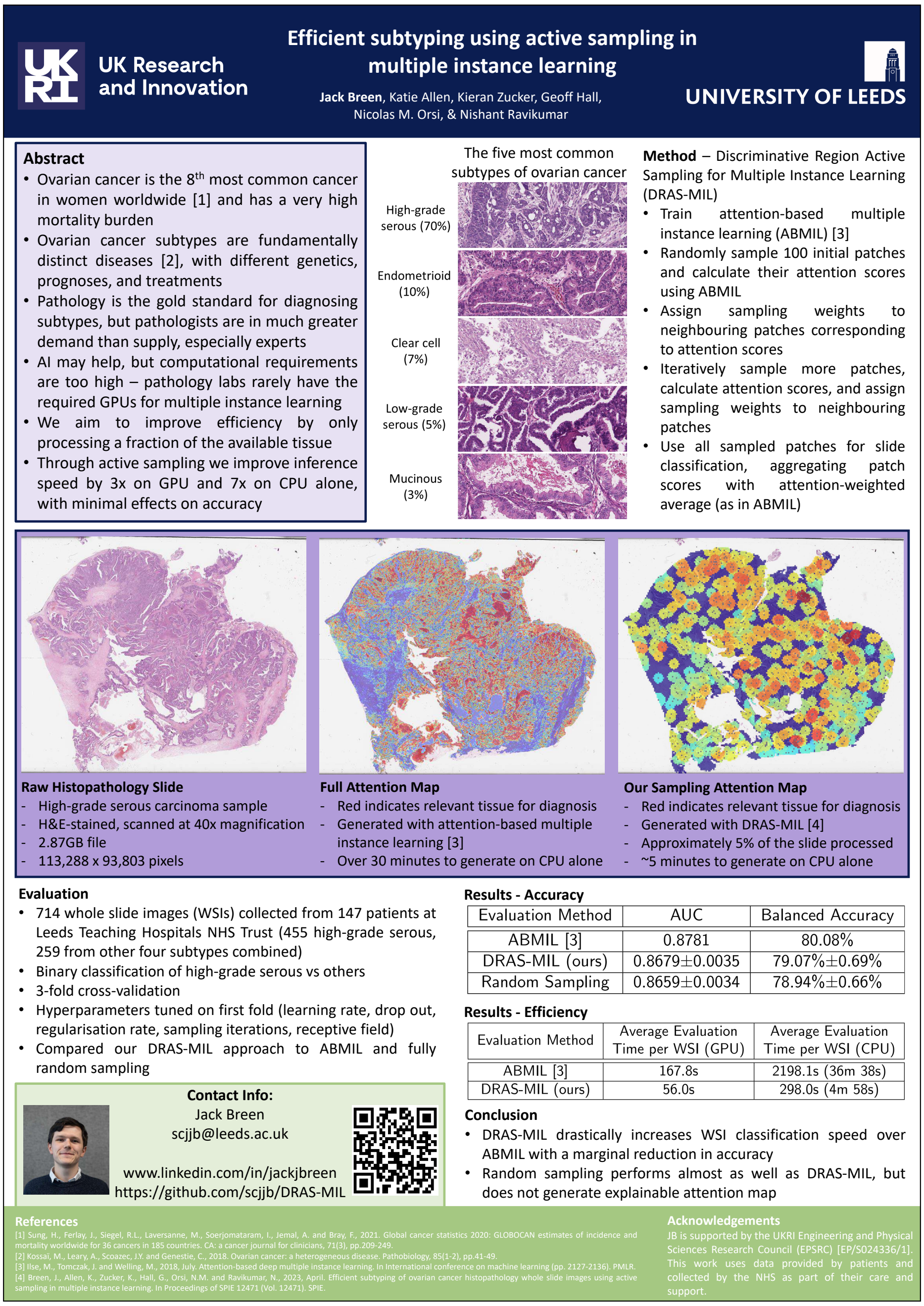
Figure F.4 Example of a WSI in which the initial model heatmap exhibited a high level of confounding, as evidenced by the background receiving greater attention than the tissue. In the final model, the tissue segmentation preprocessing step was improved to avoid including background, leading to heatmaps that do not exhibit clear confounding.

In our initial experiments, attention heatmaps showed that, in some WSIs, background regions were given much higher attention scores than tissue regions, indicating that these slides were being classified according to irrelevant information. The chromatic variability of WSIs was leading to inconsistent tissue segmentations, with some including a large amount of non-tissue areas. As a result, we adjusted our tissue segmentation parameters to achieve a more consistent performance, with the changes to a resulting heatmap shown in Figure F.4. All results presented in this paper were generated using these updated segmentations. Before this update, our 5-fold cross-validation accuracy was 89.7% and the F1 score was 0.915, which was very similar to the reported baseline model performance [147]. Improving the initial background segmentation significantly reduced internal classification performance, indicating that the slide backgrounds contained confounding information that could artificially inflate internal performance. This highlights the need for explainability in digital pathology AI to understand any model's decision-making process.

F.4 Conclusion

Overall, it is unclear whether treatment response can be accurately predicted from ovarian cancer histopathology slides alone, with our results indicating that WSIs may contain some prognostic signal that can be leveraged using hierarchical transformers and ABMIL. We found that it was beneficial to use feature extractors that were pre-trained using large sets of histopathology data, though did not find transformer-based models to outperform ResNet-based models. Given that the internal experiments were conducted on a set of only 282 histopathology WSIs from 78 patients and that external validations were conducted only on TMAs, more robust validations are required before the scale of the clinical utility of these algorithms can be adequately evaluated.

G Presented Posters



UK Research and Innovation

Efficient subtyping using active sampling in multiple instance learning

Jack Breen, Katie Allen, Kieran Zucker, Geoff Hall, Nicolas M. Orsi, & Nishant Ravikumar

UNIVERSITY OF LEEDS

Abstract

- Ovarian cancer is the 8th most common cancer in women worldwide [1] and has a very high mortality burden
- Ovarian cancer subtypes are fundamentally distinct diseases [2], with different genetics, prognoses, and treatments
- Pathology is the gold standard for diagnosing subtypes, but pathologists are in much greater demand than supply, especially experts
- AI may help, but computational requirements are too high – pathology labs rarely have the required GPUs for multiple instance learning
- We aim to improve efficiency by only processing a fraction of the available tissue
- Through active sampling we improve inference speed by 3x on GPU and 7x on CPU alone, with minimal effects on accuracy

The five most common subtypes of ovarian cancer

High-grade serous (70%)

Endometrioid (10%)

Clear cell (7%)

Low-grade serous (5%)

Mucinous (3%)

Method – Discriminative Region Active Sampling for Multiple Instance Learning (DRAS-MIL)

- Train attention-based multiple instance learning (ABMIL) [3]
- Randomly sample 100 initial patches and calculate their attention scores using ABMIL
- Assign sampling weights to neighbouring patches corresponding to attention scores
- Iteratively sample more patches, calculate attention scores, and assign sampling weights to neighbouring patches
- Use all sampled patches for slide classification, aggregating patch scores with attention-weighted average (as in ABMIL)

Raw Histopathology Slide

- High-grade serous carcinoma sample
- H&E-stained, scanned at 40x magnification
- 2.87GB file
- 113,288 x 93,803 pixels

Full Attention Map

- Red indicates relevant tissue for diagnosis
- Generated with attention-based multiple instance learning [3]
- Over 30 minutes to generate on CPU alone

Our Sampling Attention Map

- Red indicates relevant tissue for diagnosis
- Generated with DRAS-MIL [4]
- Approximately 5% of the slide processed
- ~5 minutes to generate on CPU alone

Evaluation

- 714 whole slide images (WSIs) collected from 147 patients at Leeds Teaching Hospitals NHS Trust (455 high-grade serous, 259 from other four subtypes combined)
- Binary classification of high-grade serous vs others
- 3-fold cross-validation
- Hyperparameters tuned on first fold (learning rate, drop out, regularisation rate, sampling iterations, receptive field)
- Compared our DRAS-MIL approach to ABMIL and fully random sampling

Results - Accuracy

Evaluation Method	AUC	Balanced Accuracy
ABMIL [3]	0.8781	80.08%
DRAS-MIL (ours)	0.8679±0.0035	79.07%±0.69%
Random Sampling	0.8659±0.0034	78.94%±0.66%

Results - Efficiency

Evaluation Method	Average Evaluation Time per WSI (GPU)	Average Evaluation Time per WSI (CPU)
ABMIL [3]	167.8s	2198.1s (36m 38s)
DRAS-MIL (ours)	56.0s	298.0s (4m 58s)

Contact Info:
Jack Breen
scjjb@leeds.ac.uk

www.linkedin.com/in/jackjbreen
https://github.com/scjjb/DRAS-MIL

Conclusion

- DRAS-MIL drastically increases WSI classification speed over ABMIL with a marginal reduction in accuracy
- Random sampling performs almost as well as DRAS-MIL, but does not generate explainable attention map

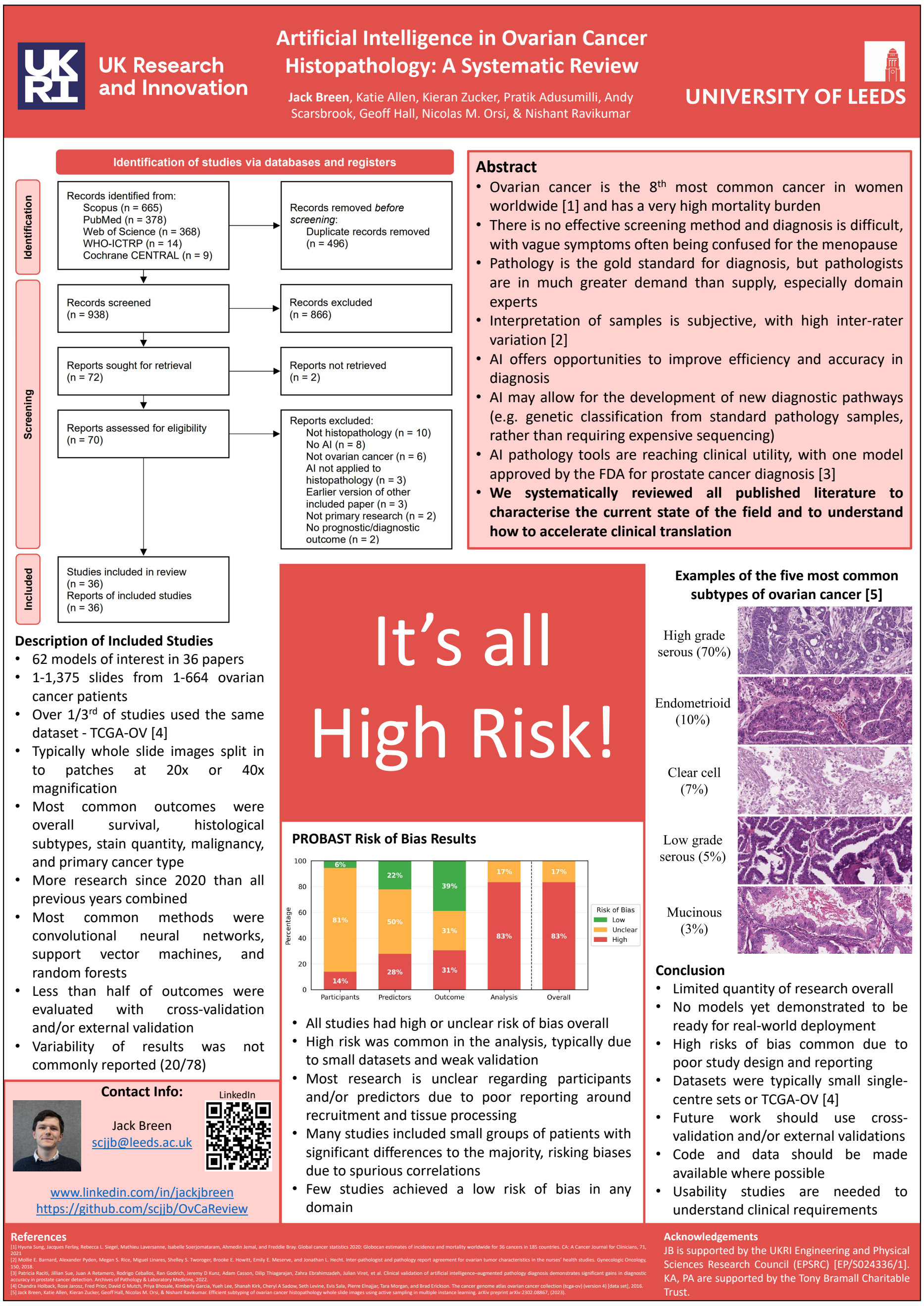
References

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F., 2021. 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), pp.209-249.
[2] Kossaï, M., Leary, A., Scoazec, J.Y. and Genestie, C., 2018. Ovarian cancer: a heterogeneous disease. Pathobiology, 85(1-2), pp.41-49.
[3] Ilse, M., Tomczak, J. and Welling, M., 2018, July. Attention-based deep multiple instance learning. In International conference on machine learning (pp. 2127-2136). PMLR.
[4] Breen, J., Allen, K., Zucker, K., Hall, G., Orsi, N.M. and Ravikumar, N., 2023, April. Efficient subtyping of ovarian cancer histopathology whole slide images using active sampling in multiple instance learning. In Proceedings of SPIE 12471 (Vol. 12471). SPIE.

Acknowledgements

JB is supported by the UKRI Engineering and Physical Sciences Research Council (EPSRC) [EP/S024336/1]. This work uses data provided by patients and collected by the NHS as part of their care and support.

Figure G.1 Poster from the 2023 AI4Health Summer School based on the work in Chapter 5. Winner of the Best Poster Prize.



UK Research and Innovation

Artificial Intelligence in Ovarian Cancer Histopathology: A Systematic Review

Jack Breen, Katie Allen, Kieran Zucker, Pratik Adusumilli, Andy Scarsbrook, Geoff Hall, Nicolas M. Orsi, & Nishant Ravikumar

UNIVERSITY OF LEEDS

Identification of studies via databases and registers

Identification

Records identified from:
Scopus (n = 665)
PubMed (n = 378)
Web of Science (n = 368)
WHO-ICTRP (n = 14)
Cochrane CENTRAL (n = 9)

Records removed *before screening*:
Duplicate records removed
(n = 496)

Screening

Records screened
(n = 938)

Records excluded
(n = 866)

Reports sought for retrieval
(n = 72)

Reports not retrieved
(n = 2)

Reports assessed for eligibility
(n = 70)

Reports excluded:
Not histopathology (n = 10)
No AI (n = 8)
Not ovarian cancer (n = 6)
AI not applied to histopathology (n = 3)
Earlier version of other included paper (n = 3)
Not primary research (n = 2)
No prognostic/diagnostic outcome (n = 2)

Included

Studies included in review
(n = 36)
Reports of included studies
(n = 36)

Abstract

- Ovarian cancer is the 8th most common cancer in women worldwide [1] and has a very high mortality burden
- There is no effective screening method and diagnosis is difficult, with vague symptoms often being confused for the menopause
- Pathology is the gold standard for diagnosis, but pathologists are in much greater demand than supply, especially domain experts
- Interpretation of samples is subjective, with high inter-rater variation [2]
- AI offers opportunities to improve efficiency and accuracy in diagnosis
- AI may allow for the development of new diagnostic pathways (e.g. genetic classification from standard pathology samples, rather than requiring expensive sequencing)
- AI pathology tools are reaching clinical utility, with one model approved by the FDA for prostate cancer diagnosis [3]
- **We systematically reviewed all published literature to characterise the current state of the field and to understand how to accelerate clinical translation**

Description of Included Studies

- 62 models of interest in 36 papers
- 1-1,375 slides from 1-664 ovarian cancer patients
- Over 1/3rd of studies used the same dataset - TCGA-OV [4]
- Typically whole slide images split in to patches at 20x or 40x magnification
- Most common outcomes were overall survival, histological subtypes, stain quantity, malignancy, and primary cancer type
- More research since 2020 than all previous years combined
- Most common methods were convolutional neural networks, support vector machines, and random forests
- Less than half of outcomes were evaluated with cross-validation and/or external validation
- Variability of results was not commonly reported (20/78)

Contact Info:

LinkedIn

Jack Breen
scjjb@leeds.ac.uk

www.linkedin.com/in/jackjbreen
https://github.com/scjjb/OvCaReview

It's all High Risk!

PROBAST Risk of Bias Results

- All studies had high or unclear risk of bias overall
- High risk was common in the analysis, typically due to small datasets and weak validation
- Most research is unclear regarding participants and/or predictors due to poor reporting around recruitment and tissue processing
- Many studies included small groups of patients with significant differences to the majority, risking biases due to spurious correlations
- Few studies achieved a low risk of bias in any domain

Examples of the five most common subtypes of ovarian cancer [5]

High grade serous (70%)

Endometrioid (10%)

Clear cell (7%)

Low grade serous (5%)

Mucinous (3%)

Conclusion

- Limited quantity of research overall
- No models yet demonstrated to be ready for real-world deployment
- High risks of bias common due to poor study design and reporting
- Datasets were typically small single-centre sets or TCGA-OV [4]
- Future work should use cross-validation and/or external validations
- Code and data should be made available where possible
- Usability studies are needed to understand clinical requirements

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, 2021

[2] Mollie E. Barnard, Alexander Pyden, Megan S. Rice, Miguel Linares, Shelley S. Tworoger, Brooke E. Howitt, Emily E. Meserve, and Jonathan L. Hecht. Inter-pathologist and pathology report agreement for ovarian tumor characteristics in the nurses' health studies. Gynecologic Oncology, 150, 2018.

[3] Patricia Raciti, Jillian Sue, Juan A Retamero, Rodrigo Ceballos, Ran Godrich, Jeremy D Kunz, Adam Casson, Dilip Thiagarajan, Zahra Ebrahimzadeh, Julian Viret, et al. Clinical validation of artificial intelligence–augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. Archives of Pathology & Laboratory Medicine, 2022.

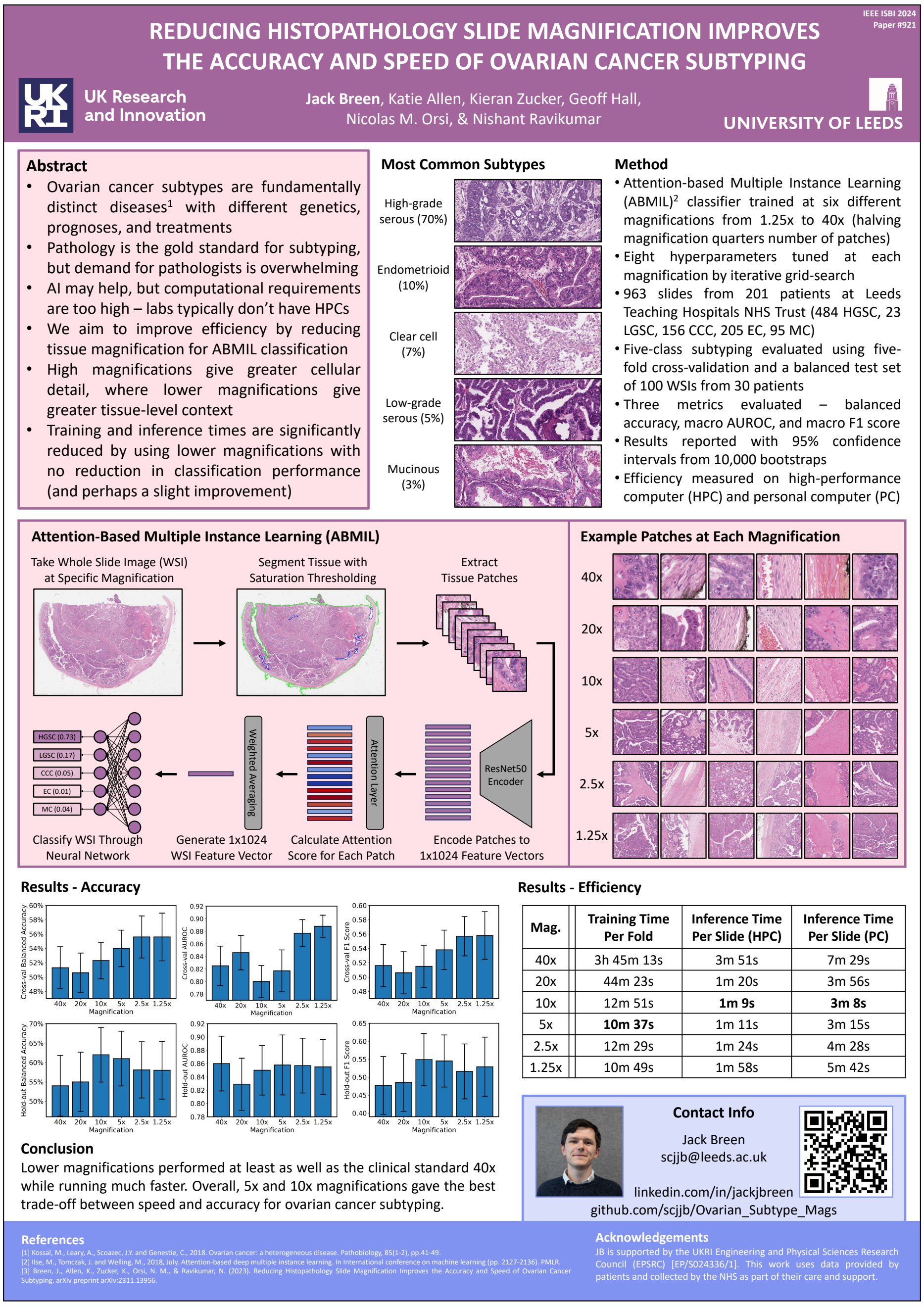
[4] Chandra Holback, Rose Jarosz, Fred Prior, David G Mutch, Priya Bhosale, Kimberly Garcia, Yueh Lee, Shanah Kirk, Cheryl A Sadow, Seth Levine, Evis Sala, Pierre Elnajjar, Tara Morgan, and Brad Erickson. The cancer genome atlas ovarian cancer collection (tcga-ov) (version 4) [data set], 2016.

[5] Jack Breen, Katie Allen, Kieran Zucker, Geoff Hall, Nicolas M. Orsi, & Nishant Ravikumar. Efficient subtyping of ovarian cancer histopathology whole slide images using active sampling in multiple instance learning. arXiv preprint arXiv:2302.08867, (2023).

Acknowledgements

JB is supported by the UKRI Engineering and Physical Sciences Research Council (EPSRC) [EP/S024336/1]. KA, PA are supported by the Tony Bramall Charitable Trust.

Figure G.2 Poster from the 2023 UKRI AI CDTs in Healthcare Conference based on the work in Chapter 3 (before peer-review feedback).



IEEE ISBI 2024
Paper #921

REDUCING HISTOPATHOLOGY SLIDE MAGNIFICATION IMPROVES THE ACCURACY AND SPEED OF OVARIAN CANCER SUBTYPING

UK Research and Innovation

Jack Breen, Katie Allen, Kieran Zucker, Geoff Hall, Nicolas M. Orsi, & Nishant Ravikumar

UNIVERSITY OF LEEDS

Abstract

- Ovarian cancer subtypes are fundamentally distinct diseases[1] with different genetics, prognoses, and treatments
- Pathology is the gold standard for subtyping, but demand for pathologists is overwhelming
- AI may help, but computational requirements are too high – labs typically don't have HPCs
- We aim to improve efficiency by reducing tissue magnification for ABMIL classification
- High magnifications give greater cellular detail, where lower magnifications give greater tissue-level context
- Training and inference times are significantly reduced by using lower magnifications with no reduction in classification performance (and perhaps a slight improvement)

Most Common Subtypes

Method

- Attention-based Multiple Instance Learning (ABMIL)[2] classifier trained at six different magnifications from 1.25x to 40x (halving magnification quarters number of patches)
- Eight hyperparameters tuned at each magnification by iterative grid-search
- 963 slides from 201 patients at Leeds Teaching Hospitals NHS Trust (484 HGSC, 23 LGSC, 156 CCC, 205 EC, 95 MC)
- Five-class subtyping evaluated using five-fold cross-validation and a balanced test set of 100 WSIs from 30 patients
- Three metrics evaluated – balanced accuracy, macro AUROC, and macro F1 score
- Results reported with 95% confidence intervals from 10,000 bootstraps
- Efficiency measured on high-performance computer (HPC) and personal computer (PC)

Attention-Based Multiple Instance Learning (ABMIL)

Take Whole Slide Image (WSI) at Specific Magnification
Segment Tissue with Saturation Thresholding
Extract Tissue Patches
ResNet50 Encoder
Encode Patches to 1x1024 Feature Vectors
Attention Layer
Calculate Attention Score for Each Patch
Weighted Averaging
Generate 1x1024 WSI Feature Vector
HGSC (0.73)
LGSC (0.17)
CCC (0.05)
EC (0.01)
MC (0.04)
Classify WSI Through Neural Network

Example Patches at Each Magnification

Results - Accuracy

Results - Efficiency

Mag.	Training Time Per Fold	Inference Time Per Slide (HPC)	Inference Time Per Slide (PC)
40x	3h 45m 13s	3m 51s	7m 29s
20x	44m 23s	1m 20s	3m 56s
10x	12m 51s	**1m 9s**	**3m 8s**
5x	**10m 37s**	1m 11s	3m 15s
2.5x	12m 29s	1m 24s	4m 28s
1.25x	10m 49s	1m 58s	5m 42s

Conclusion

Lower magnifications performed at least as well as the clinical standard 40x while running much faster. Overall, 5x and 10x magnifications gave the best trade-off between speed and accuracy for ovarian cancer subtyping.

Contact Info

Jack Breen
scjjb@leeds.ac.uk

linkedin.com/in/jackjbreen
github.com/scjjb/Ovarian_Subtype_Mags

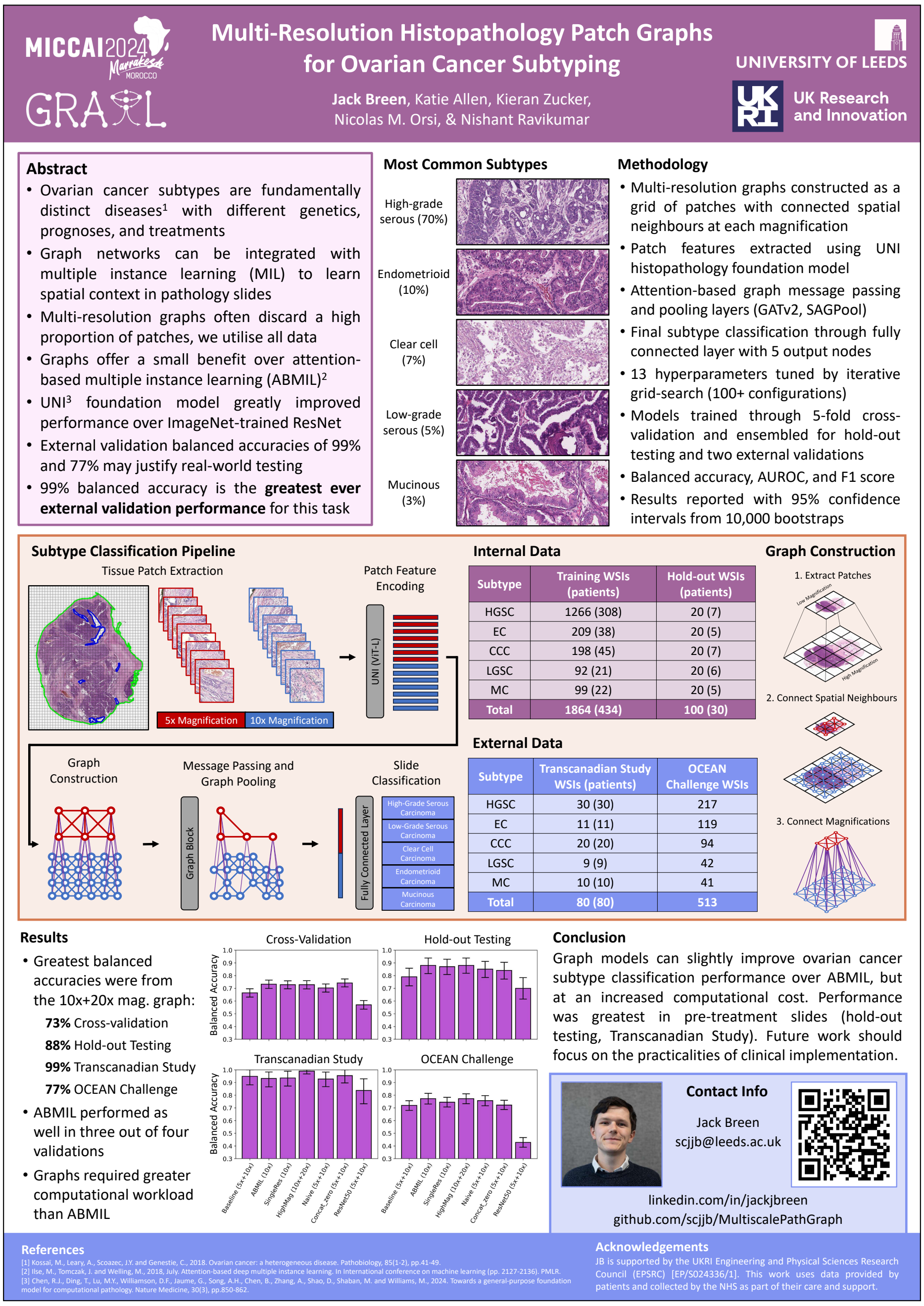
References

[1] Kossaï, M., Leary, A., Scoazec, J.Y. and Genestie, C., 2018. Ovarian cancer: a heterogeneous disease. Pathobiology, 85(1-2), pp.41-49.
[2] Ilse, M., Tomczak, J. and Welling, M., 2018, July. Attention-based deep multiple instance learning. In International conference on machine learning (pp. 2127-2136). PMLR.
[3] Breen, J., Allen, K., Zucker, K., Orsi, N. M., & Ravikumar, N. (2023). Reducing Histopathology Slide Magnification Improves the Accuracy and Speed of Ovarian Cancer Subtyping. arXiv preprint arXiv:2311.13956.

Acknowledgements

JB is supported by the UKRI Engineering and Physical Sciences Research Council (EPSRC) [EP/S024336/1]. This work uses data provided by patients and collected by the NHS as part of their care and support.

Figure G.3 Poster from ISBI 2024 based on the work in Chapter 6. This poster was not actually presented until the 2024 UKRI AI CDTs in Healthcare Conference due to being changed to an oral presentation.


MICCAI2024 Marrakesh MOROCCO
GRAIL
UNIVERSITY OF LEEDS
UKRI UK Research and Innovation

Multi-Resolution Histopathology Patch Graphs for Ovarian Cancer Subtyping

Jack Breen, Katie Allen, Kieran Zucker,
Nicolas M. Orsi, & Nishant Ravikumar

Abstract

- Ovarian cancer subtypes are fundamentally distinct diseases[1] with different genetics, prognoses, and treatments
- Graph networks can be integrated with multiple instance learning (MIL) to learn spatial context in pathology slides
- Multi-resolution graphs often discard a high proportion of patches, we utilise all data
- Graphs offer a small benefit over attention-based multiple instance learning (ABMIL)[2]
- UNI[3] foundation model greatly improved performance over ImageNet-trained ResNet
- External validation balanced accuracies of 99% and 77% may justify real-world testing
- 99% balanced accuracy is the **greatest ever external validation performance** for this task

Most Common Subtypes

Methodology

- Multi-resolution graphs constructed as a grid of patches with connected spatial neighbours at each magnification
- Patch features extracted using UNI histopathology foundation model
- Attention-based graph message passing and pooling layers (GATv2, SAGPool)
- Final subtype classification through fully connected layer with 5 output nodes
- 13 hyperparameters tuned by iterative grid-search (100+ configurations)
- Models trained through 5-fold cross-validation and ensembled for hold-out testing and two external validations
- Balanced accuracy, AUROC, and F1 score
- Results reported with 95% confidence intervals from 10,000 bootstraps

Subtype Classification Pipeline

Internal Data

Subtype	Training WSIs (patients)	Hold-out WSIs (patients)
HGSC	1266 (308)	20 (7)
EC	209 (38)	20 (5)
CCC	198 (45)	20 (7)
LGSC	92 (21)	20 (6)
MC	99 (22)	20 (5)
Total	**1864 (434)**	**100 (30)**

External Data

Subtype	Transcanadian Study WSIs (patients)	OCEAN Challenge WSIs
HGSC	30 (30)	217
EC	11 (11)	119
CCC	20 (20)	94
LGSC	9 (9)	42
MC	10 (10)	41
Total	**80 (80)**	**513**

Graph Construction

Results

- Greatest balanced accuracies were from the 10x+20x mag. graph:
  - **73%** Cross-validation
  - **88%** Hold-out Testing
  - **99%** Transcanadian Study
  - **77%** OCEAN Challenge
- ABMIL performed as well in three out of four validations
- Graphs required greater computational workload than ABMIL

Conclusion

Graph models can slightly improve ovarian cancer subtype classification performance over ABMIL, but at an increased computational cost. Performance was greatest in pre-treatment slides (hold-out testing, Transcanadian Study). Future work should focus on the practicalities of clinical implementation.

Contact Info

Jack Breen
scjjb@leeds.ac.uk

linkedin.com/in/jackjbreen
github.com/scjjb/MultiscalePathGraph

References

[1] Kossaï, M., Leary, A., Scoazec, J.Y. and Genestie, C., 2018. Ovarian cancer: a heterogeneous disease. Pathobiology, 85(1-2), pp.41-49.
[2] Ilse, M., Tomczak, J. and Welling, M., 2018, July. Attention-based deep multiple instance learning. In International conference on machine learning (pp. 2127-2136). PMLR.
[3] Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M. and Williams, M., 2024. Towards a general-purpose foundation model for computational pathology. Nature Medicine, 30(3), pp.850-862.

Acknowledgements

JB is supported by the UKRI Engineering and Physical Sciences Research Council (EPSRC) [EP/S024336/1]. This work uses data provided by patients and collected by the NHS as part of their care and support.

Figure G.4 Poster from the MICCAI 2024 GRAIL workshop based on the work in Chapter 8.